VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

Aporte del Análisis Estadístico Implicativo a Learning Analytics

TESIS DOCTORAL

Programa de Doctorado Formación en la Sociedad del Conocimiento

Doctorando
D. RUBÉN ANTONIO PAZMIÑO MAJI

Directores
DR. D. FRANCISCO JOSÉ GARCÍA PEÑALVO
Y
DR. D. MIGUEL ÁNGEL CONDE GONZÁLEZ

Octubre, 2021

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

Aporte del Análisis Estadístico Implicativo a Learning Analytics

TESIS DOCTORAL

Programa de Doctorado Formación en la Sociedad del Conocimiento

Directores:

Dr. D. Francisco José García Peñalvo

Dr. D. Miguel Ángel Conde González

Doctorando:

D. Rubén Antonio Pazmiño Maji

Octubre, 2021

GRUPO DE INVESTIGACIÓN EN INTERACCIÓN Y ELEARNING (GRIAL)
Universidad de Salamanca. Instituto Universitario de Ciencias de la Educación
Paseo de Canalejas, 169, 37008 Salamanca (España)
Tel. (+34) 923 29 45 00 Ext. 3433 – Fax (+34) 29 45 14
grial@usal.es - httt://grial.usal.es

Dr. D. Francisco José García Peñalvo, Catedrático del Departamento de Informática y Automática de la Universidad de Salamanca y Dr. D. Miguel Ángel Conde González, Profesor Titular Departamento de Ingenierías Mecánica, Informática y Aeroespacial de la Universidad de León, en calidad de directores del trabajo de tesis doctoral titulado "*Aporte del Análisis Estadístico Implicativo a Learning Analytics*" y realizado por D. Rubén Antonio Pazmiño Maji.

HACEN CONSTAR

Que dicho trabajo tiene suficientes méritos teóricos contrastados adecuadamente mediante las validaciones oportunas, publicaciones relacionadas y aportaciones novedosas. Por todo ello consideran que procede su defensa pública.

En Salamanca, octubre 2021.

Francisco José García Peñalvo
Universidad de Salamanca

Miguel Ángel Conde González
Universidad de León

AGRADECIMIENTO

En la realización de esta tesis hubo muchas personas e instituciones que me han apoyado de una u otra forma y me gustaría darles las gracias por ello, el trabajo ha sido largo pero lleno de aprendizajes, no han faltado los errores, frustraciones y correcciones, pero al final se ha completado y sobre todo he disfrutado haciéndolo, muchas gracias a todos.

A mis directores, Dr. D. Francisco J. García Peñalvo y Dr. D. Miguel Ángel Conde González, por su paciencia, tiempo, ayuda y dedicación.

Tampoco quiero olvidar a la Escuela Superior Politécnica de Chimborazo, al Grupo de investigación CIencia DE Datos (CIDED), a la Universidad de Salamanca y al GRupo de Investigación en InterAcción y eLearning (GRIAL), por la confianza y apoyo brindados especialmente en esta época de pandemia. También deseo agradecer al Profesor Regis Gras por permitirme descubrir el Análisis Estadístico Implicativo y al amigo Raphaël Couturier, por programar en el seno de la ESPOCH el software Rchic y de esta forma promover investigaciones como las que se presentan en este trabajo.

A mi familia propia y política, a mis Padres, Hermanos y en especial a Betty, gracias por tolerarme, apoyarme y animarme. A mi esposa Carmita, por comprender, apreciar y valorar el trabajo realizado. A mis hijos Cristian, Gabriela, Vanesa y mi nieta Victoria a quienes dedico este trabajo, para que en el futuro sea ejemplo de una aspiración cumplida.

A Dios por guiarme y sostenerme en este arduo trabajo.

Rubén Pazmiño Maji
Riobamba-Ecuador

RESUMEN

En los últimos años, la Analítica de Aprendizaje (LA, del inglés *Learning Analytics*) es una línea de investigación que va creciendo considerablemente. Partiendo de la definición dada en la primera Conferencia de Analíticas de Aprendizaje y Conocimiento (LAK'11): "Las Analíticas de Aprendizaje son la medición, recopilación, análisis y comunicación de datos sobre los estudiantes y sus contextos, a efectos de comprender y optimizar el aprendizaje y los entornos en que se producen"; podemos observar que LA converge al aprendizaje y a la educación.

El Análisis Estadístico Implicativo (ASI, del francés *Analyse Statistique Implicative*) se originó hace más de 40 años en la didáctica de la matemática y se aplica actualmente en la educación y otras áreas. El ASI, permite encontrar reglas de asociación entre variables y grupos de variables basándose en la definición de cuasi-implicación: si consideramos dos subconjuntos aleatorios y disjuntos $X, Y \in E$, elegidos al azar y de igual cardinalidad de A y B respectivamente; se dice que la cuasi-implicación entre a y b $(a \to b)$ es admisible al nivel de confianza $1 - \alpha$**,** si y solo si $Pr\left[Card\left(X \cap \overline{Y}\right) \leq Card\left(A \cap \overline{B}\right)\right] \leq \alpha$.

LA y ASI convergen las dos al campo educativo, entonces con el propósito de profundizar en el aporte del ASI a LA, se planteó el siguiente problema de investigación ¿Existen elementos comunes entre el ASI y LA, se puede determinar el aporte del ASI a LA?

Utilizando varias revisiones sistemáticas por aproximadamente 11 años y la teoría de conjuntos, se demostró que el ASI y LA al menos son comunes en el campo educativo y en tres métodos de análisis: minería de relaciones, descubrimiento de estructura y estadísticas, según las clasificaciones de Baker e Inventado y Papamitsiou y Economides. Se profundizó en la comparación de la complejidad algorítmica entre las técnicas de análisis comunes entre LA y ASI, debido a que LA frecuentemente necesita el análisis de grandes cantidades y nuevos tipos de datos surgidos de fuentes diversas tales como tuits, páginas web, redes sociales, emails, foros, chats, etc. Con este propósito, se utilizó un diseño pre-experimental del tipo un solo grupo aleatorio de la forma $RGXO_1$. Se encontraron importantes resultados estadísticos sobre tiempo de ejecución y espacio de memoria entre cinco técnicas de análisis clúster y cuatro técnicas de reglas de asociación del ASI y LA.

Los aportes encontrados se los determinaron: desde la definición de LA dada en LAK 2011, desde las técnicas del ASI y desde la comparación de la complejidad algorítmica entre técnicas comunes entre el ASI y LA. Además, se describen detalladamente las opciones adicionales de las técnicas ASI factibles de aportar a LA.

Se ha promovido también la colaboración entre el Análisis Estadístico Implicativo y las Analíticas de Aprendizaje, proponiendo futuras investigaciones de beneficio común.

Palabras clave: analíticas de aprendizaje, análisis estadístico implicativo, técnicas clúster, técnicas de reglas de asociación, complejidad algorítmica, Rchic.

ABSTRACT

In recent years, Learning Analytics (LA) is a line of research that is growing considerably. Taking the definition given at the first Learning Analytics and Knowledge Conference (LAK'11): "Learning Analytics is the measurement, collection, analysis and communication of data about learners and their contexts, for the purpose of understanding and optimizing learning and the environments in which it occurs", we can see that LA converges to learning and education.

Statistical implicative analysis (ASI, from the French Analyse Statistique Implicative) originated more than 40 years ago in the didactics of mathematics and is currently applied in education and other areas. The ASI, allows to find association rules between variables and groups of variables based on the quasi implication definition: if we consider two random and disjoint subsets $X, Y \in E$, chosen at random and of equal cardinality of A and B respectively; is said to be admissible at confidence level $1-\alpha$, if and only if $Pr\left[Card\left(X \cap \overline{Y}\right) \leq Card\left(A \cap \overline{B}\right)\right] \leq \alpha$.

LA and ASI both converge in the educational field, so with the purpose of deepening in the contribution of ASI to LA, the following research problem was posed: Are there common elements between ASI and LA, can the contribution of ASI to LA be determined?

Using several systematic reviews for approximately 11 years and set theory, it was shown that ASI and LA at least are common in the educational field and in three methods of analysis: relationship mining, structure discovery and statistics, according to Baker and Inventado and Papamitsiou's and Economides classifications. The comparison of the algorithmic complexity between the common analysis techniques between LA and ASI was deepened, since LA frequently requires the analysis of large amounts and new types of data from diverse sources such as tweets, web pages, social networks, emails, forums, chats, etc. For this purpose, a pre-experimental design of the type of a single randomized group of the form $RGXO_1$ was used. Significant statistical results on execution time and memory space were found between five cluster analysis techniques and four ASI and LA association rule techniques.

The contributions found were determined: from the definition of LA given in LAK 2011, from the ASI techniques and from the comparison of the algorithmic complexity between common techniques between ASI and LA. Furthermore, additional options of ASI techniques feasible to contribute to LA are described in detail.

Collaboration between Statistical Implicative Analysis and Learning Analytics has also been promoted, proposing future research of common benefit.

Keywords: Learning Analytics, Statistical implicative analysis, cluster techniques, association rules techniques, algorithmic complexity, Rchic.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

C a p í t u l o 1ro | INTRODUCCIÓN

Se plantea el problema, hipótesis, preguntas de investigación, objetivos, metodología y se los asocia con los capítulos desarrollados.

1 Capítulo.- Introducción

En este capítulo se presenta el problema, las hipótesis, las preguntas de investigación, los objetivos y la metodología utilizadas en el desarrollo de la tesis. Además, se presenta los pasos de la investigación y su relación con cada uno de los capítulos desarrollados. Finalmente, se indica el marco del investigador académico y contextualización.

1.1 Problema

Consideramos como un problema de investigación, la contradicción (oposición, contrariedad o antagonismo) entre una situación actual del objeto y una situación deseable (Espinoza Freire, 2018; Sarguera y Rebustillo, 2017), siendo el objeto de investigación el aporte (contribución, participación, ayuda (*Sinónimos de aporte*, 2020) (ASALE y RAE, 2021)) del Análisis Estadístico Implicativo (ASI, contracción del idioma francés *Analyse Statistique Implicative*) a *Learning Analytics* (LA), el problema de investigación se refleja en la siguiente pregunta: ¿Existen elementos comunes entre el ASI y LA, se puede determinar el aporte del ASI a LA?. La situación actual del objeto es la ignorancia de los aportes del ASI a LA y la situación deseable sería el conocimiento de los aportes del ASI a LA.

Learning Analytics (LA), es una línea de investigación emergente y en constante crecimiento como lo muestra un estudio reciente realizado en los últimos 10 años (Pazmiño-Maji Rubén et al., 2021), se define como (*LAK 2011 : 1st International Conference Learning Analytics and Knowledge*, 2011): "*Learning Analytics* is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs, according to the 1st International Conference on *Learning Analytics* and Knowledge".

Que tiene por traducción: Las Analíticas de aprendizaje son la medición, recopilación, análisis y comunicación de datos sobre los estudiantes y sus contextos, a efectos de comprender y optimizar el aprendizaje y los entornos en que se producen.

De la definición, se deduce que LA es un proceso («Analítica de aprendizaje», 2021) para comprender y optimizar el aprendizaje y su contexto. De aquí se desprende que uno de

los objetivos de LA es la comprensión y optimización del aprendizaje, que es un objetivo educativo.

El Análisis Estadístico Implicativo (ASI) es una teoría estadística que nació hace más de 40 años, originado por un problema en la didáctica de la matemática (Gras, 2005). Su autor fue el francés Regis Gras que basado en los conceptos de similaridad de Israel Lerman, construyó esta nueva teoría que se sigue aplicando mayormente en la educación, pero también en otros campos como economía, ambiente, salud, psicología, arte, entre otros. El ASI, permite extraer reglas de cuasi-implicación de variables y clases de variables, considerando los contraejemplos a las reglas. Las reglas de cuasi-implicación son reglas de implicación desde el punto de vista de la lógica matemática, pero con pequeño número de excepciones. Las reglas de cuasi-implicación permiten establecer relaciones asimétricas entre variables y clases de variables (Gras, 2014).

De persistir la situación actual del objeto de investigación, es decir la ignorancia de los aportes del ASI a LA, podrían perdurar y quizá incrementarse consecuencias tales como:

- No se aprovecha las experiencias y los resultados por más de 40 años de las aplicaciones ASI en LA.
- El ASI no se enriquece de los descubrimientos realizados en LA.
- Escasa producción científica en áreas comunes al ASI y a LA (Pazmiño-Maji Rubén et al., 2021).
- Poca o ninguna colaboración entre investigadores de ambas áreas.
- No se incrementan los aportes del ASI a LA.
- La educación y el aprendizaje podrían ser afectados por la falta de colaboración entre el ASI y LA.

Los métodos del ASI están automatizados en el software propietario CHIC y en Rchic (Couturier et al., 2015) que es un paquete libre programado en R. Las salidas gráficas son dendrogramas simétricos, dendrogramas asimétricos y gráficos de implicación. Las principales técnicas de análisis del ASI son el análisis de similaridad, análisis de cohesión, análisis de implicación y reducción (Couturier y Pazmiño-Maji, 2016). De continuar la situación actual del objeto de investigación, podrían mantenerse:

- Desconocimiento de la potencialidad de las técnicas de análisis del ASI.

- No aplicación de las técnicas del ASI y las opciones antes mencionadas.
- No se resuelven con técnicas ASI, ciertos problemas como por ejemplo los de cuasi-implicación en LA.
- No se puede aproximar fácilmente a relaciones causales usando técnicas ASI.
- LA no aprovecha la resistencia al ruido de las variables principales y suplementarias.

Es importante notar que, en esta nueva área de investigación como es LA, las necesidades de tiempo de ejecución y espacio de memoria de sus algoritmos son más exigentes (Gruzd y Conroy, 2020) al tratar con grandes cantidades y nuevos tipos de datos (cualitativos, textos, imágenes, videos, audios, etc.) originados del ambiente de aprendizaje del estudiante y su contexto (Muhammad et al., 2020), con fuentes diversas y registros rastreables tales como tuits, páginas web, redes sociales, bases de datos, emails, foros, chats, etc.

En esta tesis se determinan ciertos elementos comunes al ASI y LA, basándose en la definición de LA, se determinan ciertos aportes del ASI a LA y se realiza un estudio de la complejidad algorítmica en las técnicas clúster y las técnicas de reglas de asociación del ASI y LA. De no haberse realizado este último estudio, se desconocería:

- El formato de las funciones ASI aplicado a técnicas clúster y reglas de asociación.
- Las funciones de R utilizadas en técnicas clúster y reglas de asociación.
- La complejidad de los algoritmos ASI y LA en técnicas clúster.
- La complejidad de los algoritmos ASI y LA en técnicas de reglas de asociación.
- Cómo seleccionar las técnicas de análisis más adecuadas desde el punto de vista de la complejidad algorítmica.
- Cómo utilizar técnicas que acepten múltiples tipos de datos de ingreso.
- El aporte a la solución del problema de obstrucción (imposibilidad o lentitud al ejecutar ciertos algoritmos).
- El aporte a la solución del problema de adecuación de las técnicas de análisis al contexto educativo.

Por estas razones es importante el manejo, adecuación y conocimiento de otras técnicas de análisis que pueden complementar con las técnicas de LA, he aquí la importancia de determinar los aportes de las técnicas del ASI (Couturier et al., 2015).

El Análisis Estadístico Implicativo (ASI) tiene opciones adicionales de análisis (Couturier y Almouloud, 2009) tales como nodos significativos, tipicalidad, contribución, diferentes niveles de computación dependiendo del número de datos, herramientas visuales fáciles de interpretar, variables suplementarias, acepta datos numéricos y categóricos ordinales y ambientes de experimentación que ayudan en la toma de decisiones. Estas nuevas opciones podrían servir para atenuar las siguientes dificultades que se pueden presentar en LA:

Análisis inadecuados: En el artículo de la teoría a la acción (Wiley et al., 2020) se identifica que en el campo de la educación se reúnen una diversa y gran cantidad de datos de fuentes cambiantes, pero la forma en que se analizan no es adecuada, en particular en la educación superior (Siemens y Long, 2011). La obtención, recolección y análisis de datos en educación tiene desafíos significativos en varias etapas (Axelsen et al., 2020).

Acceso limitado a herramientas de análisis global: Los factores analizados por Ferguson destacan que también hay un problema de acceso a las herramientas de análisis, es decir, ¿cómo hacer para que las herramientas de análisis en LA sean accesibles para un grupo diferente de usuarios y con necesidades también diferentes? Al encontrarnos en la sociedad del conocimiento global, el hombre está en búsqueda constante de nuevos métodos, que simplifiquen las actividades en cualquier área o campo, el problema de acceso global, es decir ¿cómo hacer para que las herramientas de análisis en LA sean accesibles asincrónicamente y desde lugares remotos?, es así que los problemas de LA no solo son tecnológicos, pues su constante desarrollo exige a LA tener acceso, adecuar o incluir nuevas herramientas de análisis en la resolución de problemas educativos (Ferguson, 2014).

Selección inadecuada de técnicas (y opciones) de análisis: Se nota la existencia de un problema de selección de técnicas de análisis y opciones apropiadas utilizadas en LA de acuerdo con el tipo de datos con los que se cuente en dependencia a las necesidades, dificultades y problemas educativos a tratar.

Lentitud de cálculo: Dentro de la etapa de procesamiento de datos, las técnicas de análisis son necesarias en LA y si las empleadas no fueran eficientes (por ejemplo en

cuanto a complejidad algorítmica) se crea el problema de lentitud de cálculo o que ciertos problemas sean computacionalmente irresolubles (Laxmi et al., 2020).

Acceso a técnicas de análisis con ingresos de tipo amplio: Uno de los desafíos del LA, es utilizar herramientas de análisis aplicables a un amplio tipo de datos generados por el contexto educativo a estudiar (Ferguson, 2014). El Análisis Estadístico Implicativo (ASI, contracción del francés *Analyse Statistique Implicative*), permite determinar reglas de cuasi-implicación entre grupos de variables, tiene aproximadamente 40 años de experiencia en el ámbito educativo y tiene tres herramientas de análisis poderosas, con sus correspondientes opciones (nodos significativos, tipicalidad, contribución, etc.) como: los árboles de similaridad, grafos implicativos y árboles de cohesión, además permiten aplicarse automáticamente a diferentes tipos de datos (Zamora, Gregori, & Orús, 2009), binaria (0-1), modal (1 a 5), frecuencial (0 a 7), intervalo (rangos numéricos) y de esta forma aportar a la solución del problema de selección de técnicas apropiadas de análisis.

1.2 Hipótesis

La hipótesis del trabajo de tesis es: "El Análisis Estadístico Implicativo aporta[1] a las Analíticas de Aprendizaje", asociadas a la hipótesis de trabajo se tienen cuatro hipótesis estadísticas que son afirmaciones sobre la comparación entre la complejidad algorítmica temporal y espacial de las técnicas clúster y de reglas de asociación.

1.3 Preguntas de investigación

1. ¿El árbol de similaridad es una técnica dentro del Análisis Estadístico Implicativo que se puede utilizar en *Learning Analytics*?
2. ¿Cuáles son las ventajas del árbol de similaridad frente a otras técnicas de análisis similares utilizadas en *Learning Analytics*?
3. ¿El grafo implicativo es una técnica dentro del Análisis Estadístico Implicativo que se puede utilizar en *Learning Analytics*?

[1] Para comprender de mejor forma la palabra aporte, consideremos sus sinónimos según la Real Academia Española: contribución, participación, ayuda (*Real Academia Española*, 2021)

4. ¿Cuáles son las ventajas del grafo implicativo frente a otras técnicas de análisis similares utilizadas en *Learning Analytics*?
5. ¿El árbol cohesivo es una técnica dentro del Análisis Estadístico Implicativo que se puede utilizar en *Learning Analytics*?
6. ¿Cuáles son las ventajas del árbol cohesivo frente a técnicas de análisis similares utilizadas en *Learning Analytics*?

1.4 Objetivo general

Caracterizar el aporte de las técnicas del Análisis Estadístico Implicativo en *Learning Analytics*.

En los últimos 40 años se ha utilizado y se sigue utilizando el Análisis Estadístico Implicativo en la solución de problemas de educación en general y educación matemática en particular, pero ¿se lo puede aplicar en ?, en donde las necesidades tiempo de ejecución y espacio de memoria son más exigentes al tratar con grandes cantidades de datos, nuevos tipos de datos (cualitativos, textos, imágenes, audios, etc.) y fuentes diversas.

1.5 Objetivos específicos

1. Describir las técnicas de análisis de datos utilizadas en el Análisis Estadístico Implicativo.
2. Seleccionar las técnicas de análisis de datos del Análisis Estadístico Implicativo aplicables en *Learning Analytics*.
3. Realizar un análisis comparativo entre las técnicas de análisis de datos utilizadas en el Análisis Estadístico Implicativo con sus similares en *Learning Analytics*.
4. Determinar las ventajas de aplicar las técnicas del Análisis Estadístico Implicativo en el marco de *Learning Analytics*.

1.6 Metodología

El método de investigación es de tipo cuantitativo debido a que utiliza un paradigma positivista, según el nivel de profundización en el objeto de estudio es inferencial ya que se utilizan pruebas de hipótesis para comparar las funciones espacio de memoria y tiempo de ejecución en los diferentes algoritmos, según la manipulación de variables es pre-experimental, según el periodo temporal es transversal (Hernández et al., 2010; Patten y Newhart, 2017).

Para la comparación de la complejidad temporal y espacial se utilizó un diseño pre-experimental aplicado al área informática (Harris et al., 2006), el cual es un método de investigación cuantitativo que permite comparar estadísticamente varios grupos de interés (Cook et al., 2002). Específicamente se aplicó un pre-experimento basado en la notación de Campbell y Stanley (Campbell y Stanley, 2015).

Además, se realizaron dos revisiones sistemáticas de literatura para determinar los aportes desde la definición de LA y los aportes desde las técnicas del Análisis Estadístico Implicativo.

1.6.1 Revisión sistemática de literatura científica

La revisión narrativa es un proceso de observación, selección y análisis para obtener características en artículos científicos, mientras que la revisión y mapeo sistemáticos de literatura científica, tienen el mismo objetivo, con una diferencia importante pero muy significativa; la sistematización de criterios de selección, inclusión, exclusión y calidad, con el fin de trabajar con información de un área, que debe llevarse a cabo bajo un control riguroso y así evitar sesgos de información (Pazmiño-Maji et al., 2016).

La revisión sistemática de literatura es una metodología que tiene como objetivo principal dar un perfil de lo que sucede con cierta área de investigación, mediante el análisis de la producción científica en la misma, pero con el valor añadido de utilizar la información adecuada para dicho análisis, criterios de calidad, inclusión y exclusión de documentos (García Holgado et al., 2020). Ésta se aplica especialmente en estudios sobre medicina, pero actualmente se amplió a otros campos como la educación (I. F. González et al., 2011), la informática (Kitchenham y Charters, 2007), Ingeniería de Software (Kitchenham et al., 2009), etc.

Las características que toda revisión sistemática debe poseer son (Vidal Ledo et al., 2015):

1. Rigurosas: en cuanto a criterios usados para definir la inclusión, exclusión y/o calidad del documento.
2. Informativas: deben tratar un problema claramente definido en tiempo y espacio, con el fin de que los resultados permitan una toma de decisiones adecuada.
3. Exhaustivas: deben usar la mayor cantidad de información pertinente al tema de análisis.
4. Explícitas: deben explicitarse teórica y prácticamente los métodos usados en la revisión.

Las revisiones sistemáticas se pueden clasificar, según la profundidad del estudio, en dos: las revisiones sistemáticas de literatura, que son estudios más estructurados con relación a los criterios de selección y evaluación de los documentos analizándolos más a fondo; y los mapeos sistemáticos de literatura, los cuales dan una visión general del panorama actual del tema en análisis (R. Pazmiño-Maji, García-Peñalvo, et al., 2019).

En esta tesis se utiliza la revisión sistemática de literatura para determinar los aportes del Análisis Estadístico Implicativo a *Learning Analytics* desde la definición y desde las técnicas comunes. Se realizaron dos revisiones sistemáticas en diferente tiempo (2011-jun2016 y jul2016-jun20121, en total 11 años), pero con el mismo propósito de determinar los aportes desde la definición y desde las técnicas del ASI comunes a LA. Cada una de las revisiones tiene un total de 5 preguntas de investigación, de las cuales las 3 primeras ayudan a determinar los aportes desde la definición y las 2 últimas preguntas de investigación ayudan a determinar los aportes desde las técnicas del ASI comunes a LA. Con el objetivo de enfatizar en los aportes, en el Capítulo 5 (Aportes del ASI a LA) se consideran las dos revisiones sistemáticas y se analizan las primeras 3 preguntas, mientras que en el capítulo 6 (Aportes con Técnicas ASI a LA) se consideran las dos revisiones sistemáticas y se analizan las 2 últimas preguntas.

1.6.2 Diseño pre-experimental

(Connaway, 2015) indica que los métodos más usados hasta 1975 fueron las encuestas o experimentos en bibliotecas, las metodologías históricas y el diseño de sistemas de información. Estudios recientes han mostrado una variedad de métodos y técnicas nuevas utilizadas en Ciencias de la Información tales como: síntesis de evidencias, estudios experimentales y estudios observacionales (Koufogiannakis et al., 2004). Hider realizó un estudio sobre las publicaciones JCR del año 2005 e indica que aumentaban notablemente los estudios experimentales, más lentamente los estudios cualitativos y se observa una disminución en la metodología histórica (Hider y Pymm, 2008).

(Connaway y Radford, 2016) indica que existen tres tipos de pre-experimentos:

- El estudio del caso de un solo grupo que se puede diagramar simplemente como GXO1. Tiene un solo grupo (si es aleatorio lo representamos por RG), no hay base para la comparación de sujetos que han recibido y no han recibido el tratamiento experimental. Sin asignación aleatoria ni pruebas previas, este diseño es susceptible a numerosas influencias alternativas o amenazas a su validez interna. Está amenazada particularmente por la historia, la maduración, los sesgos de selección y otros factores. En cuanto a la validez externa, este diseño es incapaz de controlar la interacción entre los sesgos de selección y la variable experimental. Sin embargo, por razones prácticas, el estudio del caso de un solo grupo se utiliza con bastante frecuencia.
- El diseño de un grupo con pre-test y post-test (O1GXO2) y
- Comparación (G1XO1) con un grupo estático (G2), que añade un bloque de la forma (G2_O2), donde la línea indica que no se aplica el tratamiento.

En este trabajo de investigación se utilizó un pre-experimento del caso un solo grupo, para comparar las técnicas de análisis del ASI y *Learning Analytics*. Se compara el espacio de memoria y el tiempo de ejecución entre las técnicas similares del ASI y LA. El diseño pre-experimental es necesario para medir con objetividad la diferencia existente entre las medidas de complejidad de los dos grupos de técnicas.

El pre-experimento es del tipo un solo grupo aleatorio RGXO1, donde RG representa el grupo aleatorio del pre-experimento, X representa el tratamiento que en este caso son las técnicas por analizar (se consideraron también el software, hardware e inputs para que no

afecten la validez interna) y O1 son las mediciones referentes a la complejidad algorítmica. Se trabajó con un nivel significancia de α=0,05 (la normalidad se la buscó con niveles de 0,1 y 0,01). Las variables dependientes fueron la variable memoria (para el caso de la complejidad espacial) y la variable tiempo (para el caso de la complejidad temporal), ambas de tipo numérico.

Por el paradigma de investigación es de tipo cuantitativo, por el tipo de diseño utilizado es pre-experimental, por el tiempo de estudio es transversal, el colectivo de estudio lo conforman las 100 000 bases de datos aleatorias formadas por un máximo de 1000 observaciones y 100 variables (ver Apéndice C.- Tamaño de la población (colectivo de estudio)), por la amplitud es un estudio de muestreo conformado por 383 bases de datos aleatorias binarias. La población es la información sobre la muestra de estudio, tal como: nombre del archivo, número de filas que conforman la base de datos, número de columnas que conforman la base de datos, el total de datos, tiempo de ejecución, cantidad de memoria y sistema operativo.

1.7 Pasos de la Investigación

En esta investigación se realizó una adecuación a las etapas de aplicación del método comparativo propuestas por (de León y de la Garza, 2014): donde la Etapa 1 es la configuración de una estructura teórica, que se presenta en los capítulos 2, 3 y 4 (Pasos 1 y 2). La Etapa 2 que está conformada por los criterios para la selección de la muestra, que contiene el Capítulo 1 y el Apéndice C (Paso 1). En la Etapa 3, el análisis comparativo de los casos está formado por los capítulos del 5 al 9 (Pasos 3 y 4). La relación entre las etapas del Método comparativo de Carlos Gómez Díaz de León y Elda Ayde de León de la Garza, los pasos propuestos por el tesista y los capítulos desarrollados se muestran a continuación.

Etapa 1: Configuración de una estructura teórica de *Learning Analytics* y Análisis Estadístico Implicativo

- Paso 1: CONCEPTUALIZACIÓN
    i. Capítulo 2: *LEARNING ANALYTICS* (LA).
    ii. Capítulo 3: ANÁLISIS ESTADÍSTICO IMPLICATIVO (ASI).
- Paso 2: COMPATIBILIDAD

i. Capítulo 4: APROXIMACIÓN A LOS ELEMENTOS COMUNES DE (LA) Y (ASI).

Etapa 2: Criterios para la selección de la muestra de la investigación

- Paso 1: CONCEPTUALIZACIÓN
    i. Capítulo 1: INTRODUCCIÓN (y Apéndice C)

Etapa 3: Análisis de los casos de complejidad algorítmica y aportes del ASI a LA

- Paso 3: COMPARACIÓN
    i. Capítulo 7: COMPLEJIDAD ALGORÍTMICA ENTRE TÉCNICAS CLÚSTER DE LA Y ASI
    ii. Capítulo 8: COMPLEJIDAD ALGORÍTMICA ENTRE REGLAS DE ASOCIACIÓN DE LA Y ASI
- Paso 4: APORTES
    i. Capítulo 5: APORTES DEL ASI A LA
    ii. Capítulo 6: APORTE CON TÉCNICAS ASI A LA
    iii. Capítulo 9: APORTES FACTIBLES DESDE LAS OPCIONES ADICIONALES DEL ASI

A continuación, se relacionan etapas, pasos, objetivos, preguntas de investigación y capítulos.

1.7.1 Relación entre etapas, pasos y capítulos

La Figura 1.1, muestra las etapas, pasos y capítulos respectivos en los cuales se los desarrolla, así por ejemplo la etapa 1 utiliza los Pasos 1 y 2 y se desarrollan en los capítulos 2, 3 y 4, la etapa 2 utiliza el Paso 1 y se desarrolla en el capítulo 1 y así sucesivamente.

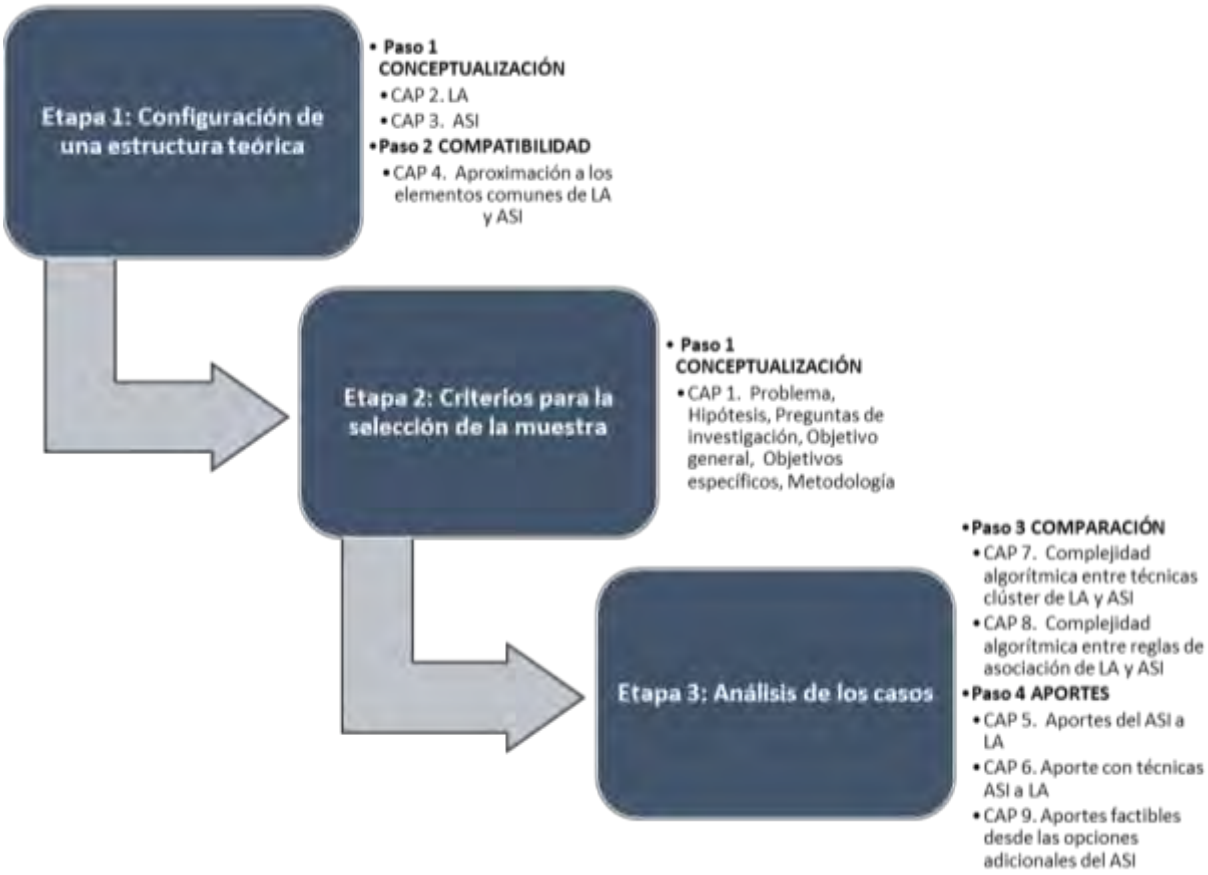


Figura 1.1.- Relación entre etapas, pasos y capítulos

1.7.2 Relación entre objetivos, pasos y capítulos

La Figura 1.2, indica los objetivos específicos planteados y los capítulos en donde se puede observar el cumplimiento de dichos objetivos. Así, por ejemplo, se tiene que el primer objetivo: descripción de las herramientas de análisis de datos del Análisis Estadístico Implicativo se encuentra en el Capítulo 3. Además, que el segundo objetivo: Seleccionar las técnicas de análisis de datos del Análisis Estadístico Implicativo aplicables en *Learning Analytics*, se encuentra en el Capítulo 6: Similaridad entre herramientas de

análisis, el tercer objetivo se cumple en los capítulos 7 y 8, y por último el cuarto objetivo se cumple en los Capítulos 5, 6 y 9.

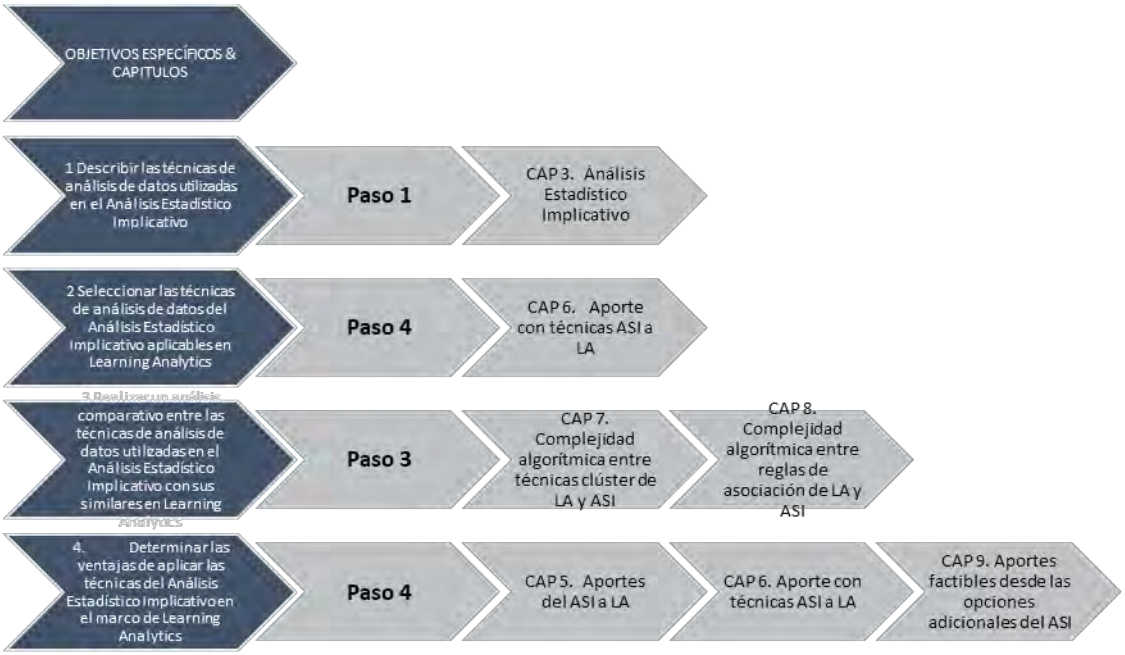


Figura 1.2.- Relación entre objetivos, pasos y capítulos

1.7.3 Relación entre pasos, preguntas de investigación y capítulos

La Figura 1.3, indica los pasos del desarrollo de la investigación y las respuestas a las preguntas de investigación planteadas.

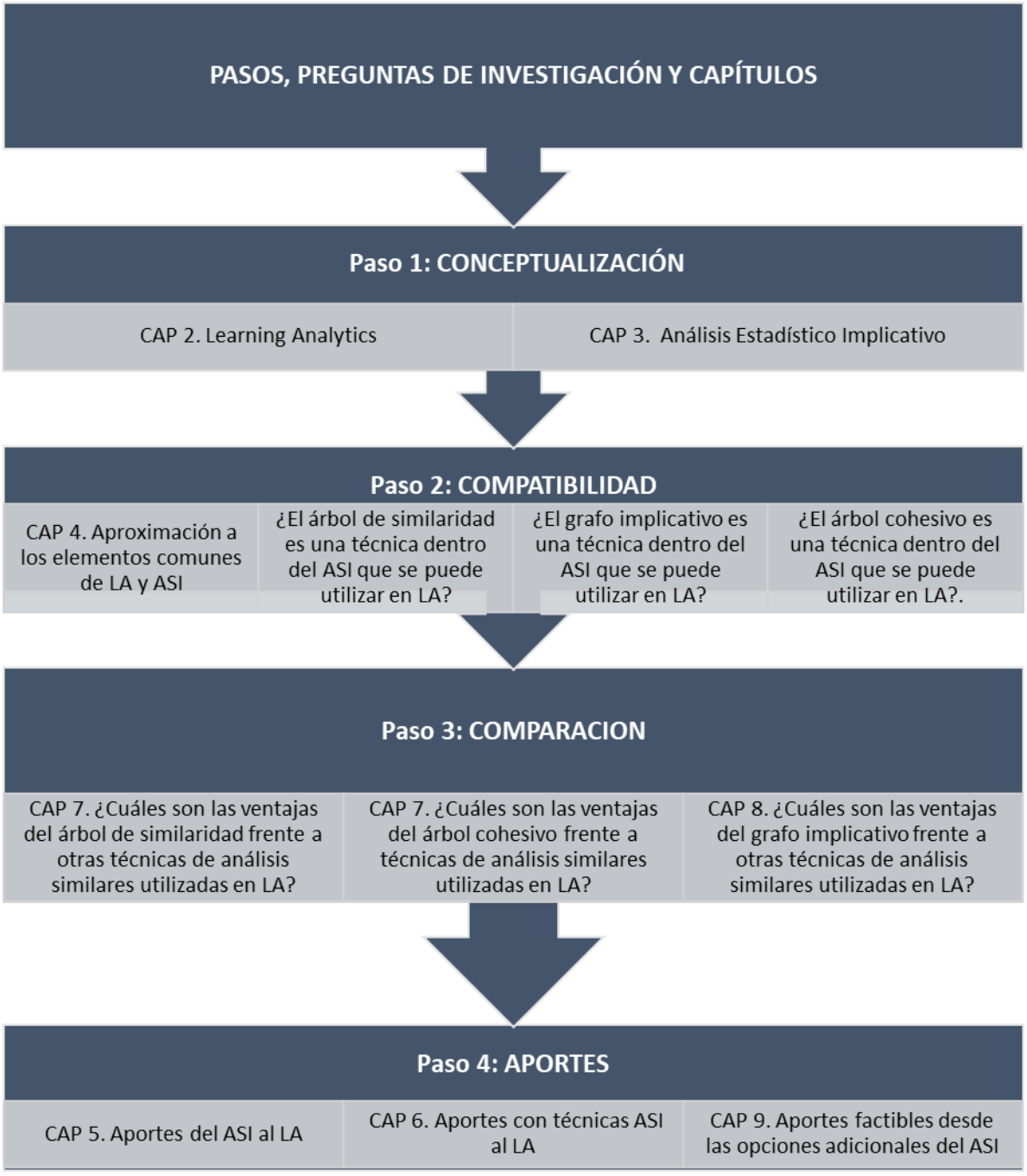


Figura 1.3.- Relación entre pasos, preguntas de investigación y capítulos

Una vez establecida la correspondencia entre etapas, pasos y capítulos, así como entre objetivos, pasos y capítulos, se esboza a continuación una descripción de los capítulos:

El **capítulo 1**, presenta el problema, hipótesis, las preguntas de investigación, los objetivos, la metodología y una asociación de los pasos seguidos con los elementos anteriores y los capítulos desarrollados.

El **capítulo 2**, describe *Learning Analytics (*LA), sobre todo contextualizado localmente con las últimas publicaciones de los autores de esta tesis.

El **capítulo 3**, describe el Análisis Estadístico Implicativo (ASI) desde el punto de vista de sus orígenes epistemológicos-educativos, la minería de datos y las técnicas de análisis que utiliza.

El **capítulo 4**, aproxima a algunos elementos comunes entre ASI y LA desde el origen y el análisis de las publicaciones sobre ASI y la definición de LA.

El **capítulo 5**, muestra detalladamente mediante revisiones sistemáticas de literatura los aportes del Análisis Estadístico Implicativo (ASI) a las Analíticas de Aprendizaje (LA).

El **capítulo 6**, vislumbra las técnicas de análisis con las que el ASI puede aportar a LA, concluyendo que son las técnicas clúster, técnicas de reglas de asociación y medidas descriptivas.

El **capítulo 7**, realiza el análisis estadístico comparativo entre las técnicas clúster de LA y ASI.

El **capítulo 8**, realiza el análisis estadístico comparativo entre las técnicas de reglas de asociación de LA y ASI.

El **capítulo 9**, describe las opciones adicionales del ASI y ejemplifica la aplicación de ASI, mediante un caso de estudio a los indicadores educativos universitarios.

El **capítulo 10**, presenta las conclusiones globales, es importante indicar que cada capítulo al final tiene detalladas sus conclusiones específicas.

1.8 Marco Investigador Académico y Contextualización

La tesis doctoral se realiza dentro del GRupo de Investigación en InterAcción y *eLearning* (GRIAL) de la Universidad de Salamanca, que es un grupo de investigación reconocido por la universidad y ahora Unidad de Investigación Consolidada UIC081 de la Junta de Castilla y León. Según (García Peñalvo et al., 2019; García-Peñalvo, 2019) tiene como actividades:

- Realización de proyectos de I+D+i tanto de forma individual como en consorcio con otras instituciones y/o grupos.
- Formación bajo demanda.
- Servicios bajo demanda de otras instituciones y empresas.

Además, las principales líneas de investigación son:

- Analítica visual.
- Calidad y evaluación en educación.
- Ciencias de la información.
- Ecosistemas tecnológicos.
- Educación médica.
- Gestión estratégica de conocimiento y tecnología.
- Humanidades Digitales.
- Ingeniería web y arquitecturas software.
- Metodologías eLearning.
- Responsabilidad social e inclusión.
- Sistemas de aprendizaje interactivos.
- Tecnologías del aprendizaje.
- TIC e Innovación educativa.

La tesis doctoral se enmarca en las líneas de investigación: TIC e Innovación educativa ya que permite contar con nuevas técnicas de análisis de LA, para innovar la educación utilizando las TIC y en la línea Gestión estratégica de conocimiento y tecnología, porque las reglas de asociación del ASI, permiten descubrir el conocimiento relacionando hechos por medio de reglas de cuasi-implicación generadas por la teoría ASI.

La presente investigación es un estudio teórico que permite el acercamiento de una teoría probabilística (ASI) y *Learning Analytics* (LA), utilizando sus respectivas técnicas de análisis comunes. El contexto es teórico formado por los artículos científicos utilizados en las revisiones sistemáticas (Capítulos 5 y 6) y las bases de datos binarias generadas aleatoriamente para su análisis (Capítulos 7 y 8).

C a p í t u l o 2do | LEARNING ANALYTICS (LA)

Se presenta el estado actual de LA en el Ecuador, mostrando antes el desarrollo de LA en países vecinos tales como Perú y Colombia, y en general en Latinoamérica.

2 Capítulo.- Learning Analytics (LA)

El ámbito de influencia de las investigaciones del tesista en ASI y LA es el Ecuador y su entorno, por eso se justifica este estudio. Los contenidos que se muestran a continuación se han extraído, adecuado y ampliado de las publicaciones de (R. Pazmiño-Maji et al., 2021).

En la introducción se hace énfasis en el estado actual de LA y además se describe el modelo de (Chatti et al., 2013).

2.1 Introducción

En este trabajo se considera la definición de *Learning Analytics* dada en la primera Conferencia Internacional sobre *Learning Analytics* y conocimiento (*LAK 2011: 1st International Conference Learning Analytics and Knowledge*, 2011) y adoptada por la Sociedad para la Investigación en LA – SOLAR (Lang et al., 2017):

> "*Learning Analytics* es la medición, recopilación, análisis y reportes de los datos sobre los alumnos y su contexto, con el fin de comprender y optimizar el aprendizaje y los entornos en los que se produce"

Se puede deducir que LA utiliza grandes cantidades y nuevos tipos de datos (cualitativos, textos, imágenes, videos, audios, etc.), además LA tiene necesidad de aumentar sus capacidades de tiempo de ejecución y espacio de memoria en sus técnicas de análisis.

2.1.1 LA en Perú

Hay dos fuentes disponibles con respecto a Perú. El primero es "*Learning Analytics* Perú: Plataforma de desarrollo para la Analítica del Aprendizaje en el Perú" (Navarrete, 2019). Esta fuente muestra casos de éxito en el uso de LA en La Educación Superior en 2018 en Perú. Por ejemplo, el 51,2% de las personas están conectadas a Internet y el 90% de los peruanos acceden a Internet con un smartphone. Se muestra como desafíos la ética y la privacidad, la investigación académica y la nueva brecha digital. Además, muestra como líneas estratégicas la promoción, el fortalecimiento de capacidades, el apoyo a la investigación y el desarrollo de soluciones. La segunda fuente es "Experiencias exitosas en el uso de *Learning Analytics* en educación superior", este video muestra historias de éxito en el uso de LA en Educación Superior en 2019 en Perú.

2.1.2 LA en Colombia

Una fuente para Colombia es: "Análisis de Aprendizaje en Colombia: Una revisión a la literatura y análisis del esfuerzo de investigación local" (Wilches y Grisales-Palacio, 2017). Este documento presenta una revisión del trabajo realizado en Colombia con respecto a LA, que identifica las instituciones e investigadores más relevantes durante los últimos años. Además, presenta un debate sobre algunas cuestiones relacionadas con los niveles de adopción en Colombia y los beneficios que podrían obtenerse si fuera posible aumentar los niveles de adopción en el sistema educativo colombiano. Concluye que existe una mayor asociación con los autores españoles, curiosamente países como Australia, Estados Unidos y el Reino Unido, que son líderes en LA aún no han establecido mayores lazos de cooperación con investigadores colombianos.

2.1.3 LA en Latinoamérica

Las dos fuentes a utilizar con respecto a América Latina, son (1) "Una Revisión Inicial de Análisis de Aprendizaje en América Latina" (Dos Santos et al., 2017) y (2) "El análisis de aprendizaje en América Latina presenta una oportunidad que no debe perderse" (Ochoa, 2019), ambos escritos por el ecuatoriano Xavier Ochoa. El primer artículo científico fue escrito en 2017 por tres autores brasileños: Henrique Lemos dos Santos, en, Joao Batista Carvalho Nunes y un ecuatoriano, Xavier Ochoa. Su objetivo es identificar las iniciativas de LA en América Latina utilizando como metodología la Cartografía Sistemática de 30 documentos de autores latinoamericanos en LA; también se utilizó una encuesta abierta para analizar datos de 28 grupos de investigación en el área de LA, el estudio se llevó a cabo de 2011 a 2016. Como parte de las conclusiones, se menciona que hay una tendencia creciente en el número de artículos científicos. De los artículos estudiados, el 20% de ellos contaban con autores no latinoamericanos de España, Canadá, Alemania, Reino Unido y Suiza, mostrando colaboraciones entre América Latina y otros países a nivel global. Artículos científicos y grupos de investigación tienen como fuentes de datos Sistemas de Gestión de Aprendizaje (Moodle, Canvas, etc.) en su mayoría, así como encuestas y sistemas institucionales. Otras fuentes de datos solo se indicaron en las encuestas como aprendizaje inmersivo e interacción cara a cara. Otro aspecto que se menciona es que la mayoría de los artículos tienen más de un fondo teórico, solo algunos de ellos muestran aplicaciones, herramientas o productos, mostrando una brecha sin explotar. En el futuro, se cruzará la información de Cartografía Sistemática y encuestas y

se creará un mapa visual e interactivo de iniciativas de análisis de aprendizaje en América Latina.

El segundo artículo de Xavier Ochoa ilustra las diferencias entre clases sociales, que la calidad de la educación es muy irregular en América Latina en todos los niveles educativos de primaria a superior (Ochoa, 2019), lo que conduce a un uso ineficiente del capital humano para un desarrollo adecuado debido a una fuerza de trabajo mal preparada y la fuga de cerebros de la élite académica. Uruguay mantiene una gran base de datos centralizada en la Administración Nacional de Educación Pública (ANEP). Además de producir estadísticas descriptivas, trabajan en un sistema de alerta temprana que predice qué estudiantes están cerca de abandonar la escuela. La Escuela Politécnica del Ejército del Ecuador (ESPOL) participó en un proyecto multimodal para proporcionar retroalimentación automática a los estudiantes sobre sus habilidades orales. Para explotar las diferentes posibilidades que LA podría ofrecer en América Latina, se ha creado un grupo internacional de investigadores llamado *Learning Analytics* Community for Latin America (F. Gutiérrez et al., 2018). También es importante tener en cuenta los siguientes dos artículos científicos sobre los congresos LALA celebrados en 2017 y 2018 respectivamente: "Red Latinoamericana de Análisis de Aprendizaje - LALA" (Sprock et al., 2017), este trabajo presenta un plan para la creación de la Red Latinoamericana de Análisis de Aprendizaje. La Red presenta una metodología de trabajo, un plan de difusión e indicadores de seguimiento, que servirán para evaluar la evaluación de la Red. Los miembros de la Red presentarán proyectos en sus universidades, con el fin de obtener fuentes de financiación para la movilidad y organización de eventos. La red está formada actualmente por 7 instituciones asociadas y 68 instituciones miembros. El documento titulado "El Proyecto LALA, Creación de Capacidad para Usar Análisis de Aprendizaje para Mejorar la Educación Superior en América Latina" (Sprock et al., 2017) resume el Proyecto LALA para desarrollar capacidad local en instituciones latinoamericanas de educación superior para diseñar, implementar y utilizar herramientas de LA para apoyar los procesos de toma de decisiones. El proyecto está formado por diferentes HEI (*Higher Education Institutions*) en América Latina y Europa, que combinan conocimiento y experiencia, incluyendo la Pontificia Universidad Católica de Chile (PUC), Universidad Austral de Chile (UACh), dos universidades en Ecuador: Universidad de Cuenca (U. Cuenca) y Escuela Superior Politécnica del Litoral (ESPOL) y tres universidades

europeas: Universidad Carlos III de Madrid (UC3M), Universidad de Edimburgo (U. Edin) y KU Lovaina (KUL).

2.1.4 LA en Ecuador

Ecuador está ubicado en América del Sur, limita al norte con Colombia, al sur y al este con Perú y al oeste con el Océano Pacífico. La extensión es de 283561 kilómetros cuadrados y tiene una población aproximada de 17 millones de habitantes (Censos, 2020).

Al finalizar su escuela secundaria, los estudiantes ecuatorianos realizan un examen de ingreso a la Educación Superior que les da acceso gratuito a 30 universidades públicas o 138 institutos públicos de educación superior, 168 en total. También hay 37 universidades privadas y 140 institutos privados de educación superior, 177 en total (*IES, listado provisional | CES - Consejo de Educación Superior | Ecuador*, 2020). En Ecuador hay un total de 345 Instituciones de Educación Superior que brindan una excelente oportunidad para la aplicación de *Learning Analytics*.

Los esfuerzos realizados por el estado ecuatoriano, universidades y educadores necesitan ayuda para mejorar los procesos de aprendizaje (Díaz, 2006), esta es la razón por la que se propone definir una línea de base (información de vanguardia que proporcione puntos de partida para la investigación y publicación) de LA con el fin de promover la investigación en esta área emergente. El *b-learning* (Gonzalez et al., 2013; Graham, 2006) y el *e-learning* (Crisol-Moya et al., 2020; García Peñalvo y Seoane Pardo, 2015; Gros y García-Peñalvo, 2016) se están volviendo populares en las universidades de Ecuador como lo indica el artículo “Entornos de aprendizaje en la escuela superior politécnica de chimborazo, transformación mediante moodle y google analytics" (Pazmiño-Maji et al., 2019), que puede ser la etapa inicial hacia el uso de LA en las instituciones de Educación Superior Ecuatorianas.

La educación superior en Ecuador se está volviendo más complicada porque la prueba de admisión no solo determina la carrera a estudiar, sino también el lugar donde se va a estudiar. Muchos estudiantes tienen que estudiar fuera de sus ciudades de origen, por lo que el aprendizaje a distancia se está utilizando cada vez más (Torres, 2003). Uno de los problemas en el desarrollo de la investigación en LA en Ecuador para jóvenes investigadores es la falta de información sobre el estado del arte.

Hace algunos años, no existía una revisión sistemática de LA en el Ecuador, y los estudios más cercanos son dos de mapeo sistemático "*Learning Analytics in Ecuador: An Initial Analysis based in a Mapping Review*" y "Un análisis inicial basado en el mapeo sistemático de los trabajos de graduación" y (R. Pazmiño-Maji, López-Ortega, et al., 2019; R. Pazmiño-Maji, Naranjo-Ordoñez, et al., 2019). La primera Revisión Sistemática responde a 4 preguntas de investigación (y 16 subpreguntas) de 68 artículos científicos, la principal pregunta de investigación es "¿Cuál es el estado del arte de la investigación sobre LA en Ecuador?" se muestran 17 características de los 68 artículos científicos, tales como: primer trabajo, número total, autores, universidades, tipo de fuentes, subárea, patrocinadores, tipos de fuente de publicación, relación entre autores y el número de artículos científicos, tendencias de la obra, etc. El segundo Mapeo Sistemático responde a 5 preguntas de investigación aplicado a tres trabajos de tesis. Los dos documentos anteriores utilizan estadísticas descriptivas y correlacionales.

Hay dos fuentes principales para Ecuador en "Análisis de Aprendizaje en Ecuador: Análisis basado en una Revisión de Mapeo" (R. Pazmiño-Maji, Naranjo-Ordoñez, et al., 2019), donde se hace una aproximación al Análisis de Aprendizaje en Ecuador, respondiendo a la pregunta principal: ¿Cuál es el estado actual de LA en Ecuador? Para ello, se realiza un mapeo sistemático en los años 2014 hasta junio de 2019 utilizando las bases de datos bibliográficas RRAAE (RRAAE Home, 2019), Scopus (Scopus - Welcome to Scopus, 2017), WOS (REUTERS, 2017) and IEEE (*IEEE Xplore Digital Library*, 2019). Se determina que la historia de LA en Ecuador comienza en 2013 con el artículo de Xavier Ochoa y una publicación sobre recursos multimodales celebrada en el congreso ICMI. El año 2018 fue el más alto en producción de investigación en Ecuador, cuya principal fuente de financiamiento fue la Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación (Senescyt). La producción científica de pregrado en LA es solo una. La producción científica de posgrado en LA no es muy significativa. No hay una red nacional que administre y aumente la investigación en LA en Ecuador. Por último, el documento "*Learning Analytics* en el Ecuador: Un análisis inicial basado en el mapeo sistemático de los trabajos de graduación" (R. Pazmiño-Maji, Conde González, et al., 2019) analiza el trabajo de pregrado y postgrado en LA en Ecuador. Se realizó un mapeo sistemático en los años 2014 a 2018 utilizando las bases de datos bibliográficas RRAAE y Senescyt (*Repositorio Digital Senescyt: Página de inicio*, 2019). Se encontraron trece

tesis, tres de las cuales fueron sobre LA a nivel de postgrado, dos maestrías y una de doctorado. Hay solo una tesis de grado sobre LA. Una fuente alternativa es "*Learning Analytics* in Continuing Training in Higher Education. Caso práctico: Universidad Nacional de Loja" (Chamba-Eras et al., 2018). El objetivo de este artículo científico es utilizar LA en la formación continua con el fin de identificar el comportamiento de los participantes en cursos virtuales. LA se utilizó para establecer la interacción en la gestión del aprendizaje de Moodle. Se concluye que el Análisis Descriptivo permitía identificar las características de los participantes y con ello entender el comportamiento en los escenarios de entrenamiento virtual. Como línea futura, se utilizaría LA predictiva con el propósito de prevenir la deserción de los cursos en línea.

2.1.5 Modelo de referencia de Chatti para LA

En el artículo "Un modelo de referencia para el análisis de aprendizaje" (Chatti et al., 2013), se plantea cuatro preguntas: ¿Qué? Tipo de datos que se recopilan, administran y utilizan, ¿Quién? Entidad a la que se dirige el análisis, ¿Por qué? Razón del análisis y ¿Cómo? Cómo llevar a cabo el análisis, las mismas que hicieron parte en la formulación de la pregunta de investigación RQ05 ¿Cuáles son las características similares del modelo de referencia Chatti?

2.2 Metodología

La metodología principal utilizada en el artículo científico fue la revisión sistemática de literatura (Neiva et al., 2016), el proceso utilizado se ve en la Figura 2.1. En la planificación de la revisión sistemática se definieron los objetivos y el protocolo (Marangunić y Granić, 2015).

El protocolo muestra el método utilizado en la revisión sistemática para minimizar el sesgo de los investigadores y demostrar que la metodología puede ser reproducida (Okoli y Schabram, 2010).

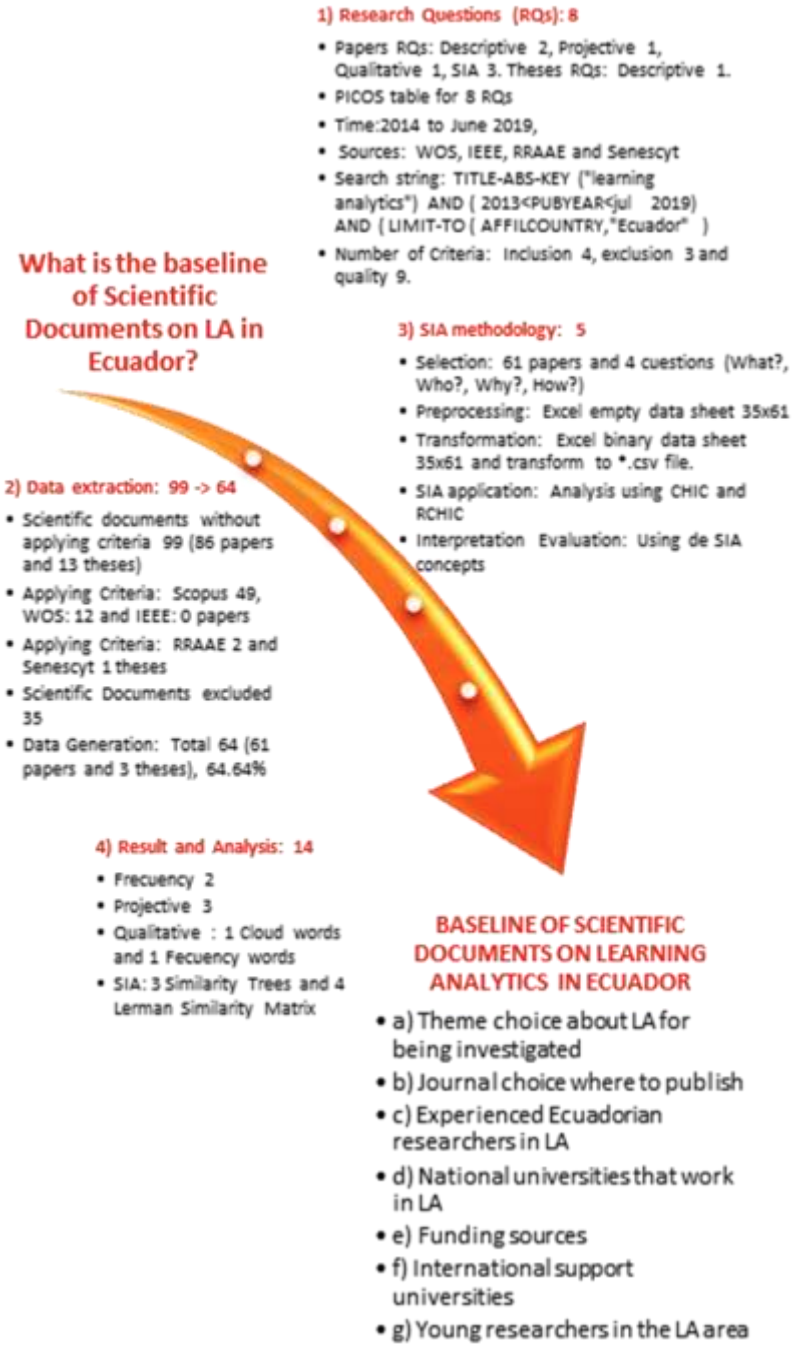


Figura 2.1.- Proceso de revisión sistemática de literatura (R. Pazmiño-Maji et al., 2021)

La generación de gráficos implicativos se basó en la metodología propuesta en el artículo "Análisis Implicativo Estadístico: Su posición en KDD y Minería de Datos" (R. Pazmiño-Maji et al., 2017c) (Figura 2.2). A continuación, resumimos el protocolo utilizado.

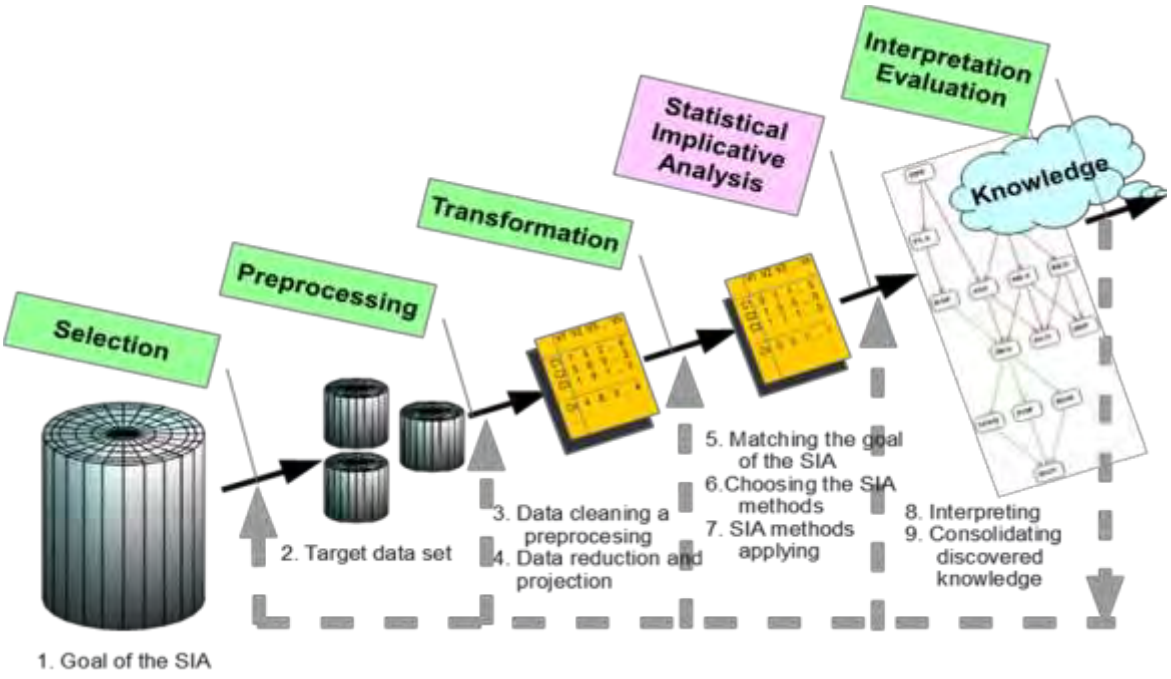


Figura 2.2.- Proceso del Análisis Estadístico Implicativo (R. Pazmiño-Maji et al., 2017c)

2.2.1 Preguntas de Investigación

Algunas preguntas en esta investigación fueron tomadas de (R. Pazmiño-Maji, Naranjo-Ordoñez, et al., 2019) y (R. Pazmiño-Maji, Conde González, et al., 2019), los resultados se han profundizado, ampliado o especificado. La investigación sistemática tenía como objetivo responder a las siguientes ocho preguntas sobre los documentos de LA en Ecuador:

RQ01: ¿Cuál es el estado del arte de los documentos de LA en Ecuador?
RQ02: ¿Cuál es la relación entre los autores y el número de trabajos por universidad?
RQ03: ¿Cuáles son las tendencias de los artículos sobre LA en Ecuador?
RQ04: ¿Cuál es la frecuencia de las palabras en los documentos sobre LA?
RQ05: ¿Cuáles son las características similares del modelo de referencia Chatti?
RQ06: ¿Cuáles son los artículos similares?

RQ07: ¿Se pueden clasificar los artículos ecuatorianos sobre LA y determinar el representante de la clase?

RQ08: ¿Cuáles son las características de la investigación de pregrado y postgrado en LA en el Ecuador?

2.2.2 Método PICOS

La Tabla 2.1, muestra la aplicación del método PICOS a las preguntas de investigación.

Tabla 2.1.- Método PICOS (R. Pazmiño-Maji et al., 2021)

RQ	P	I C	O	S
01	Papers	Does not apply	Text	Frequency
02			Text	Frequency
03			Trend line	Projective
04			Qualitative	Cloud words
05			Similarity Matrix	ASI
06			Similarity Tree	ASI
07			Reduction	ASI
08	Theses		Table	Frequency

2.2.3 Tiempo

La investigación duró 11 semestres, de 2014 a 2019.

2.2.4 Fuentes de consulta

La búsqueda se llevó a cabo en las siguientes bases de datos bibliográficas: Scopus (*Scopus - Document search*, 2020), WOS (*Web of Science - Web of Science Group*, 2010), IEEE (*IEEE - The world's largest technical professional organization dedicated to advancing technology for the benefit of humanity.*, 2020), RRAAE (*RRAAE Home*, 2019) (RRAAE es el nodo nacional que forma parte de la Red Federal de Repositorios Institucionales de Artículos Científicos de América Latina) y Senescyt (*Repositorio Digital Senescyt: Página de inicio*, 2019) (Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación de Ecuador) .

2.2.5 Cadena de búsqueda

Un grupo de estudios primarios se define en (Zhang y Ali Babar, 2010), la cadena de búsqueda final se utilizó de la siguiente manera: TITLE-ABS-KEY ("*Learning Analytics*") AND (LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018) OR LIMIT-TO (PUBYEAR, 2017) OR LIMIT-TO (PUBYEAR, 2016) OR LIMIT-TO (PUBYEAR, 2015) OR LIMIT-TO (PUBYEAR, 2014)) AND (LIMIT-TO (AFFILCOUNTRY,"Ecuador")).

Las tesis se buscaron en RRAAE y Senescyt, que se basan en el repositorio digital DSpace (*DSpace: An Open Source Dynamic Digital Repository*, 2018). DSpace no tiene cadenas de búsqueda textuales, funciona con ventanas, pero la cadena de búsqueda equivalente sería: TITLE-ABS-KEY ("*Learning Analytics*") AND ((PUBYEAR, 2014) OR (PUBYEAR, 2015) OR (PUBYEAR, 2016) OR (PUBYEAR, 2017) OR (PUBYEAR, 2018) OR (PUBYEAR, 2019)).

2.2.6 Criterios de inclusión y exclusión

Para responder a las preguntas de investigación planteadas, se definieron los criterios de inclusión y exclusión. Los criterios de inclusión y exclusión fueron:

IC1: Los documentos contienen en su título, resumen o palabras clave la frase completa "análisis de aprendizaje";

IC2: Al menos uno de los autores está afiliado a una institución ecuatoriana de educación superior;

IC3: Los artículos están escritos en cualquier idioma**.**

IC4: Los documentos pueden ser de cualquier tipo y se obtienen únicamente de las fuentes indicadas.

EC1: Es imposible acceder al documento digital.

EC2: El contenido del documento no se refiere a LA, aunque se haya encontrado en la búsqueda.

EC3: Aquí no hay conflictos de afiliación de los autores en los diferentes motores de búsqueda utilizados.

2.2.7 Criterio de calidad

Según (Kitchenham et al., 2010) se deben hacer listas de verificación de calidad, estas listas de comprobación admiten el proceso de selección. De esta manera, hemos producido la siguiente lista de comprobación de calidad:

- ¿Las búsquedas son claras y replicables?
- ¿Se han alcanzado los objetivos de investigación?
- ¿Están claramente descritos los datos utilizados y su selección justificada?
- ¿Están claramente descritos los documentos procesados?
- ¿Se describen claramente los documentos procesados en ASI?

- ¿Se describen claramente los documentos transformados en ASI?
- ¿Los artículos científicos seleccionados tratan con LA?
- ¿Se facilita la reproducibilidad de la investigación realizada?
- ¿Es evidente todo el proceso?

2.2.8 Generación de datos

Noventa y nueve artículos científicos y tesis sobre LA en Ecuador (86 artículos y 13 tesis) fueron generados utilizando fuentes y búsquedas. Por último, sesenta y cuatro documentos científicos fueron seleccionados (61 artículos y 3 tesis). Después de realizar las búsquedas, se exportaron y gestionaron con EndNote versión X9, Citavi versión 6 y Zotero versión 5,082.

2.2.9 Metodología del Análisis Estadístico Implicativo

La Metodología de Análisis Implicativo Estadístico (R. A. Pazmiño-Maji et al., 2017) se utilizó para el árbol de similitud y la reducción, como se describe a continuación.

Selección: El objetivo de las herramientas de análisis implicativo estadístico es encontrar relaciones entre 61 trabajos y el modelo de referencia Chatti para LA (Chatti et al., 2013).

Preprocesamiento: Desde los gestores bibliográficos los trabajos se exportaron a Microsoft Excel versión 2019, seleccionando la información de nombre, autores, palabras clave, resumen y el enlace para acceder a los documentos digitales. Una base de datos estadística vacía fue construida, que consta de un total de 61 filas correspondientes a los documentos y 35 columnas correspondientes a las opciones codificadas en el modelo de referencia Chatti para LA.

Transformación: La base de datos vacía se llenó de ceros y unos con la información de los documentos obtenidos; se transformó en una base de datos dicotómica en formato de texto separado por comas (*.csv).

Análisis implicativo estadístico: La base de datos estadística en formato *.csv was se analizó utilizando el software CHIC para Windows versión 6.0 y Rchic versión 0,25 (funcionó en la versión R 3.6.2 (*The Comprehensive R Archive Network*, 2015) y el RStudio 1.2.5033 (*RStudio | Open Source & Professional Software for Data Science Teams*, 2020)).

Evaluación de la interpretación: Los resultados obtenidos en el Análisis Estadístico Implicativo aplicado a la Revisión Sistemática fueron analizados y evaluados de acuerdo con el contexto de la aplicación.

2.3 Resultados

Las respuestas a las ocho preguntas de investigación formuladas se responden en detalle a continuación.

2.3.1 ¿Cuál es el estado del arte de los documentos de LA en Ecuador?

El primer documento de LA en Ecuador fue "Expertise Estimation Based on Simple Multimodal Features" (Ochoa et al., 2013) por Ochoa, X., K. Chiluiza, G. Méndez, G. Luzardo, B. Guamán, y J. Castells. Fue publicado en 2013 en el contexto del Segundo Taller Internacional de Análisis de Aprendizaje Multimodal. En total, se encontraron sesenta y cuatro (61 artículos científicos y 3 tesis) sobre LA en Ecuador. El total de artículos científicos ecuatorianos sobre LA por base de datos bibliográfica es: Scopus con cuarenta y nueve artículos, WOS doce, RRAAE dos tesis y Senescyt una tesis. Los artículos científicos ecuatorianos de LA son del 41,4% en la Escuela Superior Politécnica del Litoral (ESPOL) (*ESPOL*, 2019), seguido por el 19% en la Universidad Técnica Particular de Loja (UTPL) (*UTPL | Decide ser más*, 2012), 8,6% por la Universidad de Cuenca (*Inicio | Universidad de Cuenca*, 2019) y el 69% en otras Universidades. El principal autor ecuatoriano es Ochoa, X. (con veinte artículos científicos), el segundo autor ecuatoriano es Chiluisa, k. (con diez artículos científicos). La mayoría de los documentos de LA son documentos de la Conferencia, con un total de 55 (80,8%) artículos científicos, sub-área principal de Ciencias de la Computación con un total de 46 (67,6%) artículos científicos, mientras que el principal patrocinador de financiación es la (Senescyt – Secretaría de Educación Superior, Ciencia, Tecnología e Innovación – Ser Bachiller, Becas, Investigación, Innovación Ecuador, 2020) y la principal fuente de publicación es (CEUR-WS.org - CEUR Workshop Proceedings, 2020), con un total de 14 artículos científicos.

2.3.2 ¿Cuál es la relación entre los autores y el número de trabajos por universidad?

Si uno de los miembros del grupo de investigación cambia su producción científica, también lo hará la producción científica del grupo y de la Universidad. En la Escuela Superior Politécnica del Litoral (ESPOL), Xavier Ochoa se trasladó de la ESPOL a la Universidad de Nueva York – Steinhardt (*NYU Steinhardt*, 2020). La ausencia de Xavier Ochoa disminuirá significativamente los artículos científicos en *Learning Analytics* en Ecuador.

2.3.3 ¿Cuáles son las tendencias de los artículos sobre LA en Ecuador?

La Figura 2.3, muestra la tendencia a crecer en los próximos años del número total de documentos en LA en Ecuador. La línea de tendencia muestra un coeficiente de correlación relativamente bajo de 0,6962 (raíz del coeficiente de variación).

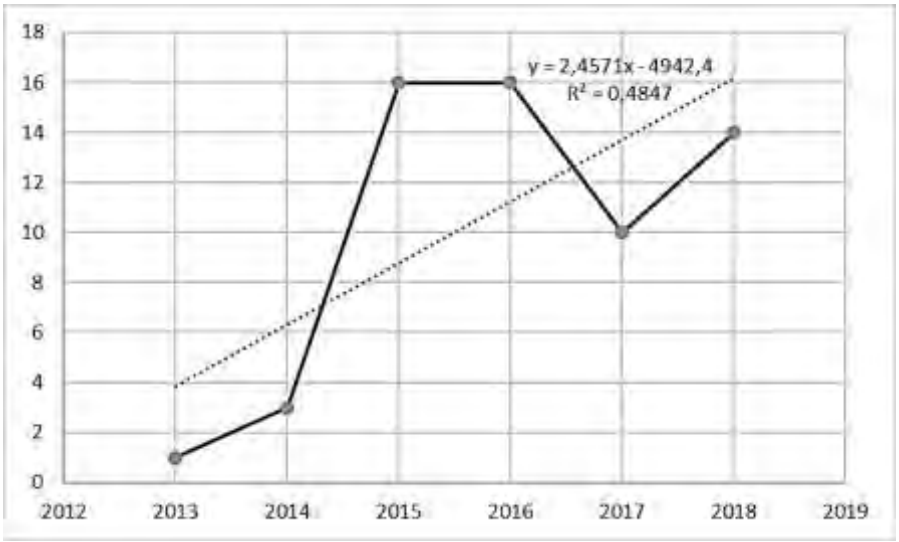


Figura 2.3.- Tendencia del número de documentos LA (R. Pazmiño-Maji et al., 2021)

A través de la línea de tendencia (bastante apropiada) se puede determinar que en el 2019 aproximadamente habrá 18 nuevos artículos científicos sobre LA en Ecuador. La Figura 2.4, hace una comparación en 2019, entre el valor proyectado por el modelo de regresión lineal 18 (punto azul) y el valor verdadero 21 (punto naranja). La condición de crecimiento se cumple sin duda, pero hay un error de sobre estimación de solo tres artículos científicos. En el año 2019 están en Scopus 14 y en WOS 7 artículos científicos

ecuatorianos sobre LA. Scopus mantiene el crecimiento de 13 artículos en 2018, 14 en 2019 y proyectados 21 en 2020. Además, se proyectó una disminución en WOS, pero hay un gran crecimiento general de 7. Esto demuestra la afirmación en (R. A. Pazmiño-Maji, Conde González, et al., 2019) de que el modelo lineal era inadecuado. Estos resultados favorecen el aumento de la producción científica en LA en Ecuador.

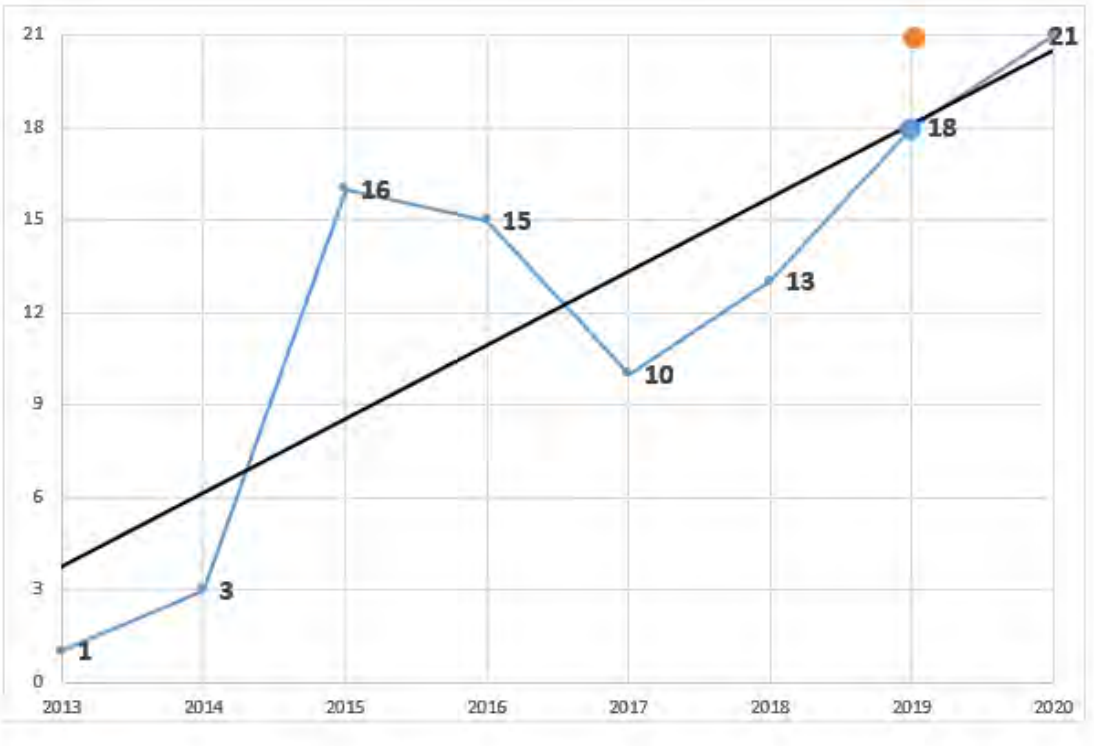


Figura 2.4.- Tendencia y valor real para el año 2019 y valor de tendencia para el año 2020 (R. Pazmiño-Maji et al., 2021)

Para predecir con mayor precisión el número de documentos en los años 2020 y 2021, se suaviza el polígono y consideran todos los datos reales (Figura 2.5). Se proyectaron un total de 22 y 25 documentos para 2020 y 2021 respectivamente.

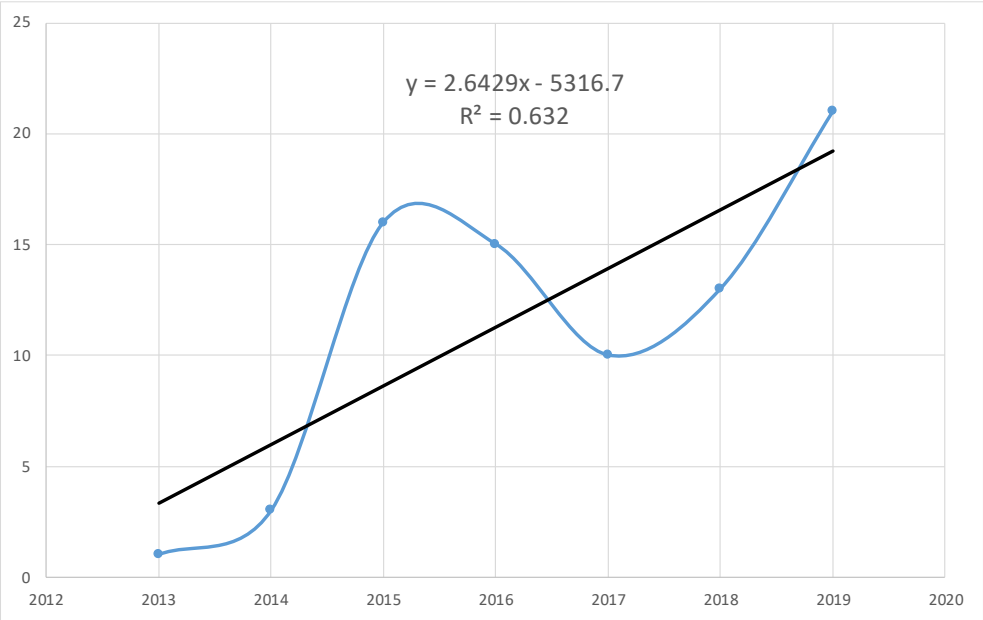


Figura 2.5.- Curva suavizada utilizando los datos históricos reales de siete años (R. Pazmiño-Maji et al., 2021)

2.3.4 ¿Cuál es la frecuencia de las palabras en los documentos sobre LA?

La palabra más fuerte es "aprendizaje" con un porcentaje del 1,3 % de frecuencia, como se nota en la palabra nube que se muestra en la Figura 2.6.

A continuación, la palabra "estudiantes" con un 0,94%, seguida de la palabra "datos" con 0,9%; la palabra "Analytics" con un 0,56%. Palabras anteriores pueden guiar al investigador a conocer el contenido de los artículos estudiados y éstos facilitan la elección de un tema de interés para ser investigado.

La Figura 2.6, muestra nube de palabras extraída de los 61 artículos científicos y procesada utilizando NVivo 12 (*NVivo qualitative data analysis software | QSR International*, 2016), ilustrando que la palabra más frecuente es "aprender" (learning).

Figura 2.6.- Nube de palabras de documentos de LA (R. Pazmiño-Maji et al., 2021)

La Tabla 2.2, muestra las 7 palabras más frecuentes en la nube de palabras de la Figura 2.6.

Tabla 2.2.- Palabras más frecuentes sobre los artículos científicos sobre LA en Ecuador (R. Pazmiño-Maji et al., 2021)

Palabras más frecuentes	Frecuencia Absoluta	Frecuencia Relativa
learning	1485	1,30%
students	1081	0,94%
data	1028	0,90%
using	919	0,80%
educational	716	0,62%
analytics	642	0,56%
courses	597	0,52%

Las palabras más frecuentes están directamente relacionadas con el objetivo de LA, en otras palabras, permite a los estudiantes mejorar su aprendizaje utilizando análisis para probar datos sobre su proceso de aprendizaje en su propio contexto.

2.3.5 ¿Cuáles son las características similares del modelo de referencia Chatti?

Se ha creado una matriz de comparación gráfica basada en la similitud de Israel (Lerman, Chantrel, et al., 1981). Los valores de similaridad se han calculado en el software Rchic y se han trazado mediante el paquete corrplot (Wei y Simko, 2017) en el lenguaje de programación estadística R. El objetivo de la matriz es determinar dentro de cada una de las preguntas propuestas por Chatty (¿Qué? ¿Quién? ¿Por qué? y Cómo?) los artículos similares, teniendo en cuenta los 61 documentos analizados.

En la Figura 2.7.**a)** (Wath (Qué)?: Tipo de datos que se recopilan, gestionan y utilizan), el valor más alto en la similitud de Lerman es 0,9. Este valor se logra para (whatVL, whatWe) y (whatMu, whatWe), esta dimensión se relacionó con los datos recogidos durante la investigación. Las fuentes de datos pueden ser entornos de aprendizaje virtual (VLE), encuestas, redes sociales como Facebook, bibliotecas electrónicas como SciELO o cualquier otro repositorio. Por lo tanto, hay un gran número de documentos que eligen (o no) la misma razón para el tipo de datos (VLE, dispositivos portátiles) o (Multimedia, dispositivos portátiles).

La Figura 2.7.**b)** (Who (Quién)?: Entidad a la que se dirige el análisis) muestra el valor más alto en la similitud de Lerman que es 0,9. Este valor se logra para (whoTe, whoSt) y (whoTe, whoCo). Esta dimensión se refiere a las partes interesadas y significa que hay un gran número de trabajos que eligen (o no) la misma razón para el análisis (Profesores, Estudiantes) o (Profesores, Coordinadores).

La Figura 2.7.**c)** (Why (Por qué) ?: Motivo del análisis) muestra el valor medio en la similitud de Lerman que es 0,5. Este valor se logra para (whyAd, whyTu). Esta dimensión tiene que ver con los objetivos del análisis y significa que hay un número medio de documentos que eligen (o no) la misma razón para el análisis (Adaptación y tutoría).

La Figura 2.7**.d)**. (How (Cómo)?: Cómo llevar a cabo el análisis) muestra que el valor más alto en la similitud de Lerman es 1. Este valor se logra para (howSo, howIn). Está relacionado con las técnicas empleadas durante la investigación para lograr los objetivos

establecidos por la dimensión why y significa que hay un número muy alto de documentos que eligen (o no) la misma técnica de análisis (Análisis de redes sociales, Visualización de información). En el software CHIC utilizamos la opción del Árbol de similitudes, para generar los siguientes árboles.

2.3.6 ¿Cuáles son los artículos similares?

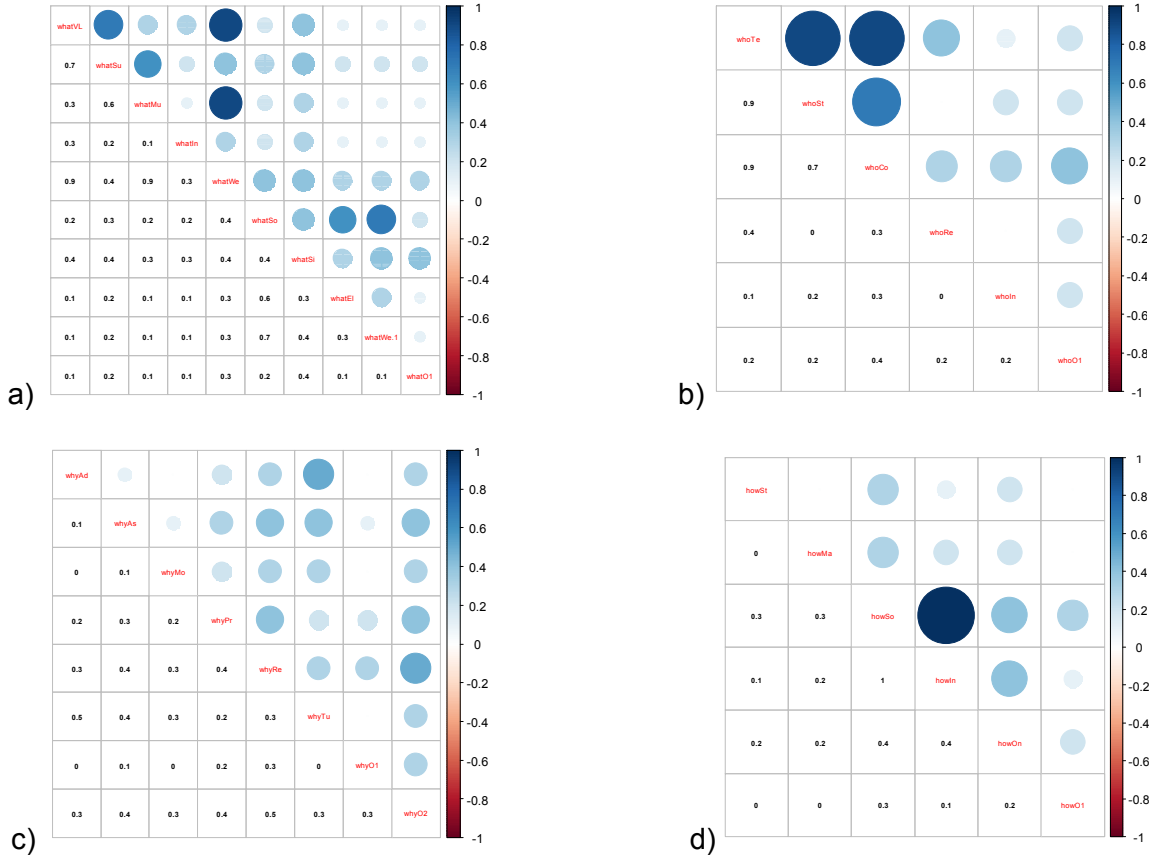


Figura 2.7.- Matriz de similaridad de Lerman, sobre: ¿Qué?, ¿Quién?, ¿Por qué? y ¿Cómo? (R. Pazmiño-Maji et al., 2021)

En la Figura 2.8**,** los datos se analizaron usando el ASI que se menciona en los tres documentos, con una diferencia significativa entre ellos: los artículos científicos número P26 (F. Gutiérrez et al., 2018) y P27 (Kizilcec et al., 2017) recopilaron datos de instituciones, mientras que los datos recogidos de P36 (Ochoa et al., 2016) se hizo en multimedia, estos tres documentos son para beneficio de los estudiantes. También se proporcionaron tutoriales usando la información de los documentos P26 y P27, mientras

que el paper P36 se utilizó para evaluar y dar retroalimentación a los estudiantes; además, los tres documentos todavía se consideran en la fase de experimentación.

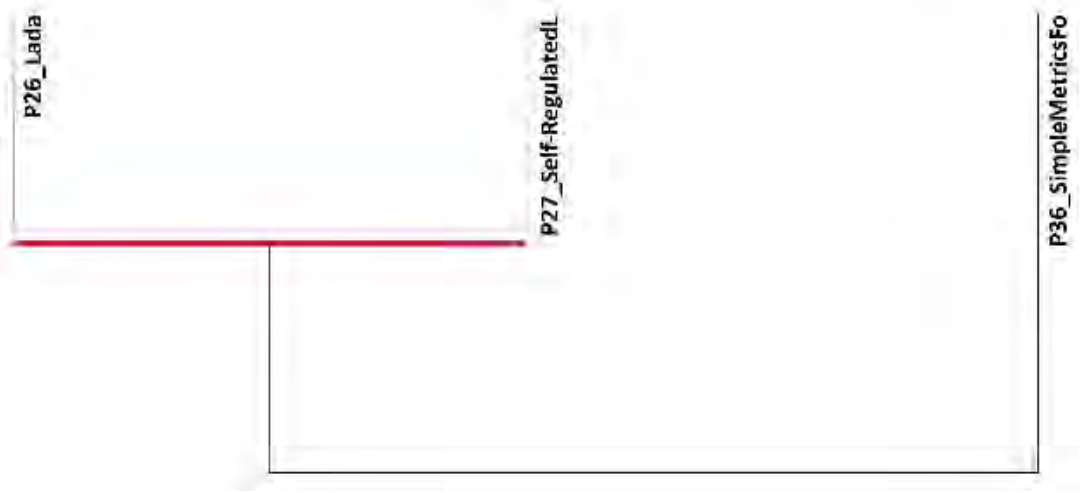


Figura 2.8.- Árbol de similaridad entre ((P26 P27) P36) (R. Pazmiño-Maji et al., 2021)

En la Figura 2.9, los documentos P20 (Fernandez y Lujan-Mora, 2016) y P33 (Naranjo Serrano y Pazmiño Maji, 2018a) utilizan una forma diferente de analizar la información, mientras los documentos P33, P29 (Luzardo et al., 2014) y P22 (García-Tinisaray et al., 2018) analizan la misma información, pero los documentos P20 y P33 son los mismos en términos de cómo recopilaron los datos porque utilizan información de programación.

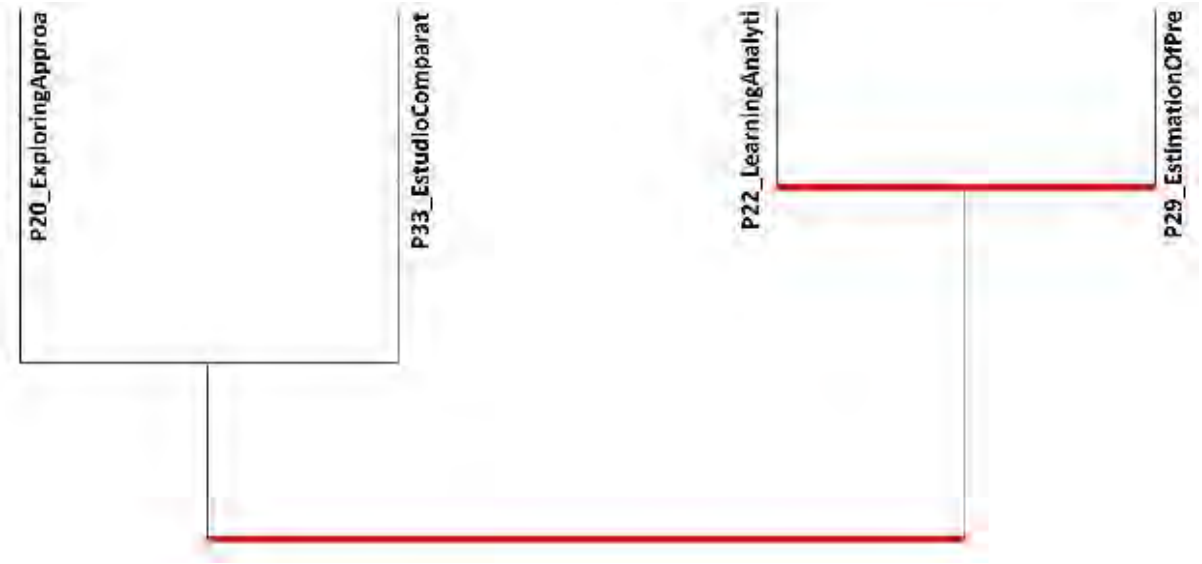


Figura 2.9.- Árbol de similaridad entre ((P20 P33) (P22 P29)) (R. Pazmiño-Maji et al., 2021)

Los documentos P29 y P22 no hacen eso. Los cuatro son útiles para los estudiantes que quieren aprobar el rendimiento académico y tratar de proporcionar la información a los estudiantes utilizando nuevos métodos; además, cuatro artículos son trabajos analíticos. Las líneas en rojo son nodos significativos que son nodos correspondientes a una clasificación compatible lo mejor posible con los valores y la calidad de la agrupación obtenida.

La Figura 2.10, muestra dos grupos de similaridad bien definidos: el grupo de la izquierda formado por 5 documentos (P02 (Aguilar, Cordero, et al., 2017), P05 (Aguilar et al., 2018), P03 (Aguilar et al., 2018), P04 (Aguilar et al., 2016), P41 (Piedra et al., 2015)) y el grupo de la derecha formado por 4 documentos (P13 (Díaz Nafría et al., 2015), P58 (Pérez-Álvarez et al., 2018), P21 (Fiallos y Ochoa, 2019), P28 (Kloos et al., 2016)), Tabla 2.3.

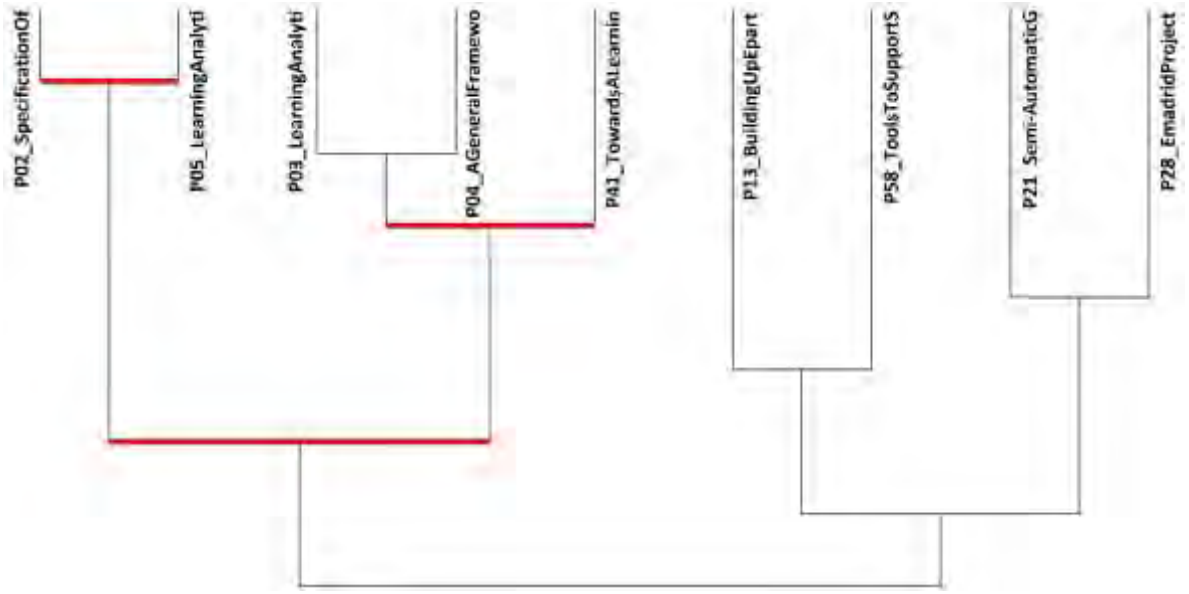


Figura 2.10.- Árbol de similaridad completo entre (4 y 5) (R. Pazmiño-Maji et al., 2021)

2.3.7 ¿Se pueden clasificar los artículos ecuatorianos sobre LA y determinar el representante de la clase?

Es posible automatizar la clasificación de los documentos y la determinación de los representantes; ASI utiliza la opción de reducción que permite generar clases representativas basadas en el concepto de índice de implicación. En este caso se utilizó la teoría clásica y el modelo binomial.

La Tabla 2.3, muestra los 61 documentos analizados en el SR. Los documentos están etiquetados con la letra P, un número y los quince primeros caracteres del título. Los trabajos se dividen en 6 clases, los documentos de cada uno de ellos son equivalentes y el representante de la clase se muestra en negrita.

Tabla 2.3.- Clasificación de los 61 documentos (documentos representativos indicados en negrita para cada clase) (R. Pazmiño-Maji et al., 2021)

Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
P36_SimpleMetricsFo	**P11_OntologyForMode**	**P17_MultimodalColla**	**P05_LearningAnalyti**	**P58_ToolsToSupportS**	**P12_LearningAnalyti**
P01_FrontiersInEduc	P06_Moocs	P08_TheUseOfToolsOf	P02_SpecificationOf	P13_BuildingUpEpart	P48_2017FourthInter
P18_TowardsCollabor	P07_TheUseOfLaToEva	P09_ApplicationOfMa	P03_LearningAnalyti	P21_Semi-AutomaticG	P50_TechnologiesAnd
P19_TowardsADistrib	P10_ProposalOfAppli	P14_MultimodalSelfi	P04_AGeneralFramewo	P28_EmadridProject	
P22_LearningAnalyti	P15_AnInitialReview	P20_ExploringApproa	P41_TowardsALearnin		
P23_ImprovingTheUse	P16_Visla	P25_LearningAnalyti			
P24_CompetencesAsSe	P30_2ndCrossmmla	P39_HowToMapLearnin			
P26_Lada	P33_EstudioComparat	P54_MethodologicPro			
P27_Self-RegulatedL	P37_MultimodalLearn	P57_EVEAAdaptiveStr			
P29_EstimationOfPre	P38_MLA?14-ThirdMul				
P31_TechniquesForDa	P42_Editorial				
P32_AHybridInfrastr	P46_ApproximationOf				
P34_TheRAPSystem	P47_CurrentAndFutur				
P35_VisualizingUnce	P52_2015MultimodalL				
P40_DesignOfAToolTo	P53_ProceedingsOf20				
P43_ModelToPredictA	P55_2015XliLatinAme				
P44_BusinessIntelli	P56_Edulearn16:8thI				
P45_FirstStepsTowar	P59_NewAdvancesInIn				
P49_ABusinessIntell					
P51_ExploringTheImp					
P60_CloudComputingI					
P61_Latin-AmericanN					

El concepto de equivalencia depende del índice de implicación, de las variables utilizadas y del llenado adecuado de la base de datos. Por ejemplo, en las cuatro columnas (Clase 4): los documentos **P05**, P02, P03, P04 y P41 son equivalentes para la implicación y el representante de la clase es **P05**. Los cinco documentos (P05 (Aguilar, Valdiviezo-Díaz, et al., 2017), P02 (Aguilar, Cordero, et al., 2017), P03 (Aguilar et al., 2018), P04 (Aguilar et al., 2016), P41 (Piedra et al., 2015)) pertenecen a la misma clase porque la mayoría de ellos tienen estas características imperativas: se emplean técnicas para el análisis de los datos recogidos (aprendizaje automático), tipo de datos utilizados (multimedia), los beneficiarios (estudiantes), trabajo analítico, etc. **P05** es el representante de la clase, ya que este documento cumple con todas las características anteriores (multimedia), los beneficiarios (estudiantes), el trabajo analítico realizado, etc.

2.3.8 ¿Cuáles son las características de la investigación de pregrado y postgrado en LA en el Ecuador?

Hay una tesis de grado de ingeniería informática llevada a cabo en LA por un becado ecuatoriano Jordi Casanovas Muñoz (Casanovas Muñoz, 2016), la cual propone una ontología que es una estandarización de los conceptos dentro del dominio de *e-learning* y la relación entre ellos. La ontología se validó utilizando datos de una escuela que utiliza AGORA como Sistema de Gestión del Aprendizaje LMS, del inglés Learning Management Systems (Conde et al., 2014; García-Peñalvo y Alier Forment, 2014).

Posteriormente, se ha implementado la asignación de los conceptos entre ontología y LMS. En esta tesis LA está relacionada con los datos tomados de los estudiantes y su interacción digital con el LMS (Casanovas Muñoz, 2016).

Hay dos tesis de postgrado, preparadas en 2018 por Naranjo Serrano Mauricio Medardo (Naranjo Serrano y Pazmiño Maji, 2018a) y en 2016 por Casanovas Muñoz Jordi (Tinisaray y Karina, 2016). Los detalles se proporcionan en la Tabla 2.4.

Tabla 2.4.- Características de la investigación de postgrado en LA en Ecuador (R. Pazmiño-Maji et al., 2021)

Características	**Tesis de posgrado**	
Título	Estudio comparativo del análisis estadístico implicativo y el *Learning Analytics* en relación al uso de las técnicas de exploración de datos educativos (Naranjo Serrano y Pazmiño Maji, 2018a)	Construcción de un modelo para determinar el rendimiento académico de los estudiantes basado en *Learning Analytics* (análisis del Aprendizaje), mediante el uso de técnicas multivariantes (Tinisaray y Karina, 2016)
Autores	Naranjo Serrano, Mauricio Medardo	Daysi Karina García Tinisaray
Co Autores	Pazmiño Maji, Rubén Antonio	José Luis Pino Mejías, Juan Manuel Muñoz Pichardo
Año	2018	2016
Palabras clave	*Learning Analytics*, análisis estadístico implicativo, métodos clúster, minería de reglas de asociación	No keywords
Tipo	Máster	Doctoral
Institución de Educación Superior	Pontificia Universidad Católica del Ecuador- Sede Ambato	Universidad de Sevilla/ Universidad Técnica Particular de Loja

En el análisis final (Naranjo Serrano y Pazmiño Maji, 2018a) en términos de técnicas clúster, el método ASI era el mejor en tiempo de ejecución, pero eran equivalentes en memoria. Se concluye que el método ASI requiere más tiempo y los otros son cada vez más similares entre sí (Apriori - eclat - weclat) en términos de tiempo y memoria sobre las reglas de asociación. En cuanto a la hora, eclat es el más rápido, seguido de apriori, luego

weclat, mientras que simlrty es el último. Hay pocos datos para aplicar una técnica de tendencia, pero se puede percibir en el caso de másteres que tienden a aumentar y en el caso de doctorado no hay suficientes datos. En esta tesis, LA proporciona las técnicas de análisis de clústeres y reglas de asociación más utilizadas y luego las compara con las de ASI (Naranjo Serrano y Pazmiño Maji, 2018a).

La segunda es una tesis doctoral, con el objetivo de construir un modelo para determinar el rendimiento académico de los estudiantes basado en LA, utilizando técnicas estadísticas multivariantes (análisis multinivel y análisis logístico bivariado). Las variables identificaron la influencia que ejercen sobre el rendimiento académico. Las estimaciones permiten a una institución educativa mejorar la focalización de las intervenciones y apoyar los servicios a los estudiantes en riesgo de problemas académicos. El género no es estadísticamente significativo y la región de origen no tiene ningún efecto en el rendimiento académico. En el caso del estudio de rendimiento académico óptimo, la edad indica que los jóvenes estudiantes tienen menos ventaja. La participación en actividades en línea muestra que los estudiantes que participan poco en actividades en línea tienen menos ventaja de lograr un rendimiento académico óptimo, en comparación con los estudiantes que participan moderada o activamente (Tinisaray y Karina, 2016). En esta tesis LA está relacionado con los datos de contexto de los estudiantes (género, rendimiento académico, edad, participación en actividades en línea, etc.).

2.4 Discusión

El paper " Learning analytics in Ecuador: An initial analysis based in a mapping review" (R. Pazmiño-Maji, Naranjo-Ordoñez, et al., 2019) indica que hay 68 artículos científicos realizados, aunque para ser precisos, esos 66 artículos provienen de Scopus, WOS, IEEE y 2 tesis se encontraron en RRAAE. El artículo "*Learning Analytics* en el Ecuador: Un análisis inicial basado en el mapeo sistemático de los trabajos de graduación" (R. Pazmiño-Maji, López-Ortega, et al., 2019), presenta un estudio de la licenciatura y posgrado en Ecuador, considera el repositorio digital DSpace del Senescyt, en lugar de una base de datos bibliográfica. Se encontraron dos tesis, una de maestría y otra de doctorado, los resultados de esta investigación profundizan los resultados de los dos trabajos anteriores y también construyen la línea de base de la investigación de LA en Ecuador. Esta línea base se construyó utilizando nuevas herramientas teóricas ASI como

el concepto de similaridad. Los resultados de similaridad al aplicar el software Rchic a los 61 trabajos permitieron la conformación de 6 clases desarticuladas. En cada clase, se determinó el documento más representativo (desde el punto de vista del modelo de referencia Chatti para LA). Se obtuvieron un total de 6 documentos representativos (según el modelo de referencia de Chatti para LA) que al leerlos dan una idea del contenido de los 61 documentos iniciales. Para tener una idea general del contenido de los 61 documentos, se utilizó una nube de palabras. Esta herramienta de análisis gráfico permite observar la frecuencia de las palabras ordenadas por tamaño. Las palabras más grandes nos darán una primera idea de lo que contienen los papeles. Los dos primeros documentos utilizaron herramientas descriptivas y proyectivas, mientras que este documento utiliza un concepto ASI de similaridad. Las respuestas descriptivas a la revisión sistemática dan información exacta para construir la línea de base. Las respuestas ASI (y nube de palabras) a la revisión sistemática permiten construir objetos (documentos, clases o palabras), que representan grandes conjuntos de documentos. Las preguntas de investigación de la revisión sistemática, que utilizó la nube de palabras, permitieron especificar las características generales de la investigación en LA en Ecuador. La similaridad se aplica de dos maneras: por primera vez se utilizan matrices de similaridad, ofreciendo una respuesta cuantitativa y gráfica a todas las relaciones de similaridad de las opciones de las cuatro preguntas propuestas por el modelo de referencia Chatti para LA. La otra forma de utilizar la similaridad es a través de árboles de similaridad o dendrogramas, que es una forma clásica de salida del software CHIC y Rchic. Las matrices de similaridad proporcionan explícitamente todos los valores de similaridad binaria y los visualizan mediante círculos proporcionales. Los dendrogramas muestran la similaridad jerárquica entre las clases de variables, pero el gráfico no observa explícitamente las cuantificaciones. Las dos formas de calcular la similaridad son combinadas y complementarias. La herramienta de similaridad también se puede aplicar en la investigación cualitativa porque permite el procesamiento de texto (Silva y de Almeida, 2017). Dendrogramas permiten la creación de los llamados clúster desde el punto de vista de las estadísticas multivariantes.

2.5 Conclusiones

Los trabajos se pueden agrupar en la investigación cualitativa utilizando la técnica de racimo para que se genere un dendrograma base en la similaridad de palabras entre los papeles. El investigador elige los criterios de similaridad entre los trabajos rellenando previamente la matriz de datos en el ASI (en este caso características similares del modelo de referencia Chatti).

Esto se detectó en el análisis de similaridad cuando se aplicó ASI, donde demuestra que es una herramienta muy útil y poderosa en el momento de determinar objetos similares, en particular diferentes artículos científicos. La mayoría de los métodos utilizan una comparación de texto para determinar duplicados; estos pequeños cambios pueden dar lugar a un método fallido. ASI ayuda a determinar los documentos duplicados en el momento de realizar una revisión sistemática para que el investigador pueda elegir cualquier criterio de similaridad (en este caso el modelo de referencia Chatti).

La tesis de Mauricio Naranjo es bastante interesante, pero los resultados se ven afectados porque la entrada de datos no es adecuada. Se sugiere repetir esas pruebas teniendo en cuenta que el método simlrty lo logra desde un archivo externo en la etapa de lectura; y los métodos eclat, weclat y apriori no lo necesitan.

A partir de los resultados obtenidos se construyó en Ecuador una línea de base (información de vanguardia que proporciona puntos de partida para la investigación y publicación) de la investigación en AL. Los pasos seguidos se presentan a continuación:

a) La elección del tema sobre LA para ser investigado se puede hacer a través de

(1) Abordar el contenido global de los artículos mediante el análisis de la frecuencia de palabras a través de la nube de palabras o la tabla*:* Sección 2.3.4

(2) Palabras clave: Tabla 2.2

(3) Documentos agrupados por similaridad: Sección 2.3.6

(4) Documentos más relevantes por clase: Sección 2.3.7

b) Elección de la revista dónde publicar:

(1) Sección 2.3.1

c) Investigadores ecuatorianos experimentados en LA:

(1) Sección 2.3.1 y Tabla 2.4 (cuarta fila)

d) Universidades nacionales que trabajan en LA:

(1) Sección 2.3.1 y Tabla 2.4 (última fila)

e) Fuentes de Financiamiento:

(1) Sección 2.3.1

f) Universidades internacionales de apoyo:

(1) Tabla 2.4 y artículo científico (Dos Santos et al., 2017)

g) Jóvenes investigadores en el área de LA:

(1) Sección 2.3.1 y Tabla 2.4 (tercera fila)

h) Lecturas iniciales sugeridas (para criterios de agrupación de acuerdo con la Sección 2.3.1)

(1) Sección 2.3.6, Sección 2.3.7 y Tabla 2.3

i) Para una idea general de los trabajos analizados

(1) Sección 2.3.4

Como resultado de esta investigación, se puede afirmar que las investigaciones de alto impacto en *Learning Analytics* en Ecuador son todavía pocas, pero suficientes para establecer una línea de base, que permitirá a los investigadores interesados tener un punto de partida para aumentar la investigación de alto impacto en el análisis de aprendizaje.

Un trabajo futuro implica mantener actualizada la revisión sistemática, con el fin de medir el avance de la investigación en *Learning Analytics* en Ecuador. Se abre una posible línea de investigación, es decir, la aplicación del proceso de Análisis Estadístico Implicativo en Revisiones Sistemáticas.

Capítulo 3ro | EL ANÁLISIS ESTADÍSTICO IMPLICATIVO (ASI)

Se muestras las principales características del Análisis Estadístico Implicativo con el objetivo de observar su origen, paradigmas y técnicas en el análisis de datos.

3 Capítulo.- El Análisis Estadístico Implicativo (ASI)

El conocimiento se construye con hechos y sus relaciones (Gras y Kuntz, 2009), es decir los hechos son importantes y aportan al conocimiento, pero el buscar relaciones entre ellos ayuda a que el conocimiento no se lo vea en forma aislada, el ASI en forma sencilla y con un fundamento teórico fuerte permite establecer relaciones asimétricas de cuasi-implicación que incrementan el conocimiento basado en hechos ya conocidos. El ASI está formado por un conjunto de técnicas de análisis que trabajan con diversidad de variables, que en forma general tiene técnicas tales como la similaridad, implicación, cohesión y reducción, que se fortalecen con opciones adicionales como los nodos significativos, la tipicalidad y la contribución y utiliza la entropía para grandes conjuntos de datos.

3.1 Introducción

El Análisis Estadístico Implicativo (del francés ASI) es una técnica nacida hace aproximadamente 40 años en Francia, para abordar un problema de categorización de los niveles cognitivos en las pruebas estandarizadas en matemáticas, resuelta por Regis Gras (Régnier et al., 2019) basado en la teoría de la similaridad de Israel Lerman (Lerman, Gras, et al., 1981). Las cuasi-reglas $a \Rightarrow b$ se determinan cuando la medida de calidad confirmatoria se logra a partir de la inverosimilitud de la ocurrencia en los datos, el número de casos que la invalidan (número de excepciones, los contraejemplos), es decir, para los cuales se verifica a sin que se verifique b. Formalizando, como Israel Lerman hizo por similaridad, consideremos dos subconjuntos aleatorios y disjuntos X e Y de E, elegidos al azar, independientes y de igual cardinalidad de A y B respectivamente; se dice que $a \Rightarrow b$ es admisible a nivel de confianza $\mathbf{1} - \boldsymbol{\alpha}$, si y solo si: $Pr\left[Card\left(X \cap \overline{Y}\right) \leq Card\left(A \cap \overline{B}\right)\right] \leq \alpha$. La variable aleatoria $\mathrm{Card}(\mathrm{X} \cap \bar{\mathrm{Y}})$ sigue la ley de Poisson de parámetro $\frac{\mathrm{n_a n_{\bar{b}}}}{\mathrm{n}}$. Finalmente, la definición se reformula como: La Implicación $a \Rightarrow b$ se cumple al nivel de confianza $\mathbf{1} - \boldsymbol{\alpha}$ si y solo si $\phi(a,b) \geq 1 - \alpha$, donde $\phi(a,b) = 1 - Pr\left[Q\left(a,\bar{b}\right) \leq q\left(a,\bar{b}\right)\right] = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$ (Gras et al., 2009).

Los procesos de cálculo se automatizan utilizando el software propietario CHIC (Couturier y Gras, 2005a) o su equivalente libre Rchic (Pazmiño et al., 2015) (R. Pazmiño-Maji et al., 2018). Los resultados mostrados son numéricos (medidas descriptivas, matriz de

similaridad, reducción, etc.) y gráficos como el árbol de similaridad (Naranjo et al., 2018), gráfico implicativo (R. Pazmiño-Maji et al., 2017a) y árbol de cohesión (R. Pazmiño-Maji et al., 2017b). Algunas de las opciones complementarias implementadas en CHIC son: la entropía que se utiliza para analizar una gran muestra de datos, los nodos significativos que dan soporte a los gráficos, las variables complementarias que son variables cualitativas como el género, el nivel educativo o la categoría económica; la contribución se utiliza para saber cuáles son los temas o clases de sujetos más responsables de las implicaciones calculadas y la tipicalidad indica los sujetos típicos de la población para las implicaciones calculadas (Gras et al., 2009). El ASI tiene muchas aplicaciones en el área educativa en general y en la educación matemática en particular. Debido a que el ASI tolera muy bien las excepciones, se aplica ampliamente en las Ciencias Sociales. Desde hace algún tiempo se han hecho aplicaciones tan diversas en áreas como la genética, la medicina, el arte, etc. (Barragán-Pazmiño y Pazmiño-Maji, 2018).

3.2 Origen epistemológico didáctico

El ASI en sus orígenes (Gras, 2014) resolvió la necesidad de organizar las preguntas por las respuestas dada por los estudiantes a una prueba objetiva de didáctica de las matemáticas (Spagnolo, 2005). Las preguntas tenían diferente nivel de complejidad fijado previamente, la organización de las preguntas debía respetar la complejidad preliminar de las mismas y esto conllevaría a la creación de un índice de implicación entre los ítems de las respuestas, para evaluar reglas como: "si a entonces generalmente b", donde a y b son respuestas a las preguntas de la prueba objetiva. Por ejemplo, un objetivo expresado en términos de la utilización de un proceso de cálculo (de la inversa de una matriz con el método de los determinantes, por ejemplo) se consideraría de complejidad inferior a un objetivo que requiere la elaboración de un nuevo proceso más eficiente (por ejemplo, por el número de operaciones). Una relación de tipo causal podría ser: las herramientas cognitivas de un objetivo superior serían suficientes para que el alumno cumpla un objetivo de nivel inferior, es decir si puede construir un nuevo proceso más eficiente, seguramente puede utilizar el método de los determinantes para el cálculo de la inversa.

La teoría se desarrolló con las múltiples aplicaciones y considerando analogías con la situación inicial, condujo a que la solución vaya más allá de los datos binarios, se resolvió el problema también para variables modales, frecuenciales y de intervalo (Gomes y

Régnier, 2005). Últimamente también se estableció una relación topológica entre sujetos y variables, que dio la pauta para aplicarse las técnicas del ASI no solo a las variables sino también a los casos (Régnier et al., 2020).

Durante los años 70, Regis Gras asistió nuevamente a investigar en las clases de secundaria y fue testigo de las dificultades de aprendizaje de los alumnos que en algunos casos eran recurrentes, la naturaleza del ASI, es didáctico pero también a menudo epistemológico (Bachelard, 1993).

3.3 Minería de datos

En este documento se responden a siete preguntas sobre el descubrimiento de conocimientos en base de datos y el Análisis Estadístico Implicativo, mediante el análisis inicial de 200 artículos científicos relacionados con el ASI. La quinta pregunta indica:

¿Qué tan cerca están los documentos de ASI de los pasos de la minería de datos?

Para responderla, se analizaron 35 artículos científicos utilizando un mapeo sistemático, se realizó un estudio basado en los pasos para la computación en KDD (del inglés Knowledge Discovery in Databases) propuestos por (Fayyad et al., 1996), en particular el paso que se refiere a la minería de datos.

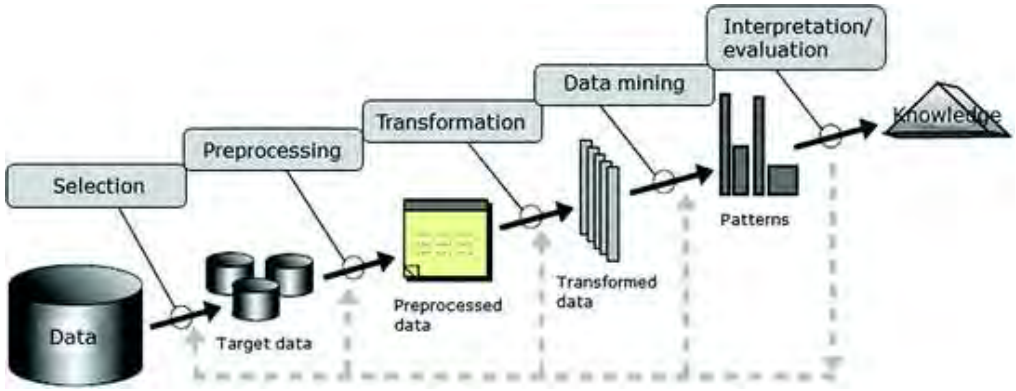


Figura 3.1.- Pasos que constituyen el proceso KDD (Fayyad et al., 1996)

La respuesta a la pregunta se realizó mediante las siguientes observaciones respecto al acercamiento a los pasos mostrados en la figura anterior:

Tabla 3.1.- Porcentaje de acercamiento del ASI al proceso de Data Mining (R. A. Pazmiño-Maji et al., 2017)

Paso		Porcentaje de acercamiento al proceso de Data Mining
Número	Nombre	
5	Elección de la función de minería de datos	94
6	Elección del algoritmo de minería de datos	54
7	Minería de datos	31
Subproceso	Minería de datos	86
	Promedio	66

El promedio de porcentaje de acercamiento al proceso de Data Mining 66% es un porcentaje medio alto. Esto se debe a que los métodos ASI pueden ser similares a los métodos de minería de datos.

Un artículo a tomar en cuenta dentro de la minería de datos es el titulado: "Minería de datos de salud, seguridad y desempeño ambiental usando Análisis Estadístico Implicativo" (Gallerati, 2008), donde inicialmente se observa que la minería de datos y el ASI, pueden enfocarse en varios aspectos de investigación como en salud, seguridad y medio ambiente. La predicción es una parte fundamental en la investigación ya que tiene la capacidad de pronosticar un fenómeno que sucederá en el futuro, a partir de información del pasado. Mediante el estudio de las variables se responden ciertas preguntas de investigación, para ello Gallerati aplica dos enfoques dentro del estudio, así como la generación de reglas de implicación. Los dos enfoques son el cualitativo que se basa en las características del entorno social y otro cuantitativo que recopila cantidades numéricas en el tiempo. Los resultados en este artículo científico fueron interesantes, novedosos y útiles, de tal forma que se pueden aplicar en casi cualquier investigación en la que se requiera de la toma de decisiones para el mejoramiento (Blanchard et al., 2005).

3.4 Grandes conjuntos de datos

La versión entrópica considera el contrapositivo $\bar{b} => \bar{a}$, que podría reforzar la afirmación de la implicación entre a y b, también podría mejorar la calidad de la discriminación de ϕ cuando la transacción se establece que aumenta: si A y B son pequeños, sus conjuntos complementarios son grandes y viceversa. Se desarrolló una extensión de la intensidad de implicación que toma en cuenta la calidad de la regla contrapositiva, la regla

estadística y la fuerza de inclusión de la regla, la intensidad de la implicación entrópica ha sido integrada en el software CHIC dedicado al análisis estadístico implicativo para procesar grandes conjuntos de datos (Gras et al., 2001).

Las comparaciones experimentales han destacado dos características interesantes cuando estas medidas no seleccionan las mismas reglas; en varias bases de datos se encontraron un subconjunto de reglas no sorprendentes con una buena confianza, demostrando la relevancia de algunas de estas reglas sobre datos de la vida real para la toma de decisiones (Gras et al., 2001).

3.5 Variables por su contenido

Por el tipo de datos, las variables que pueden utilizarse en el ASI, son de tipo binario, modal, frecuencial y de intervalo, también se encuentra en desarrollo la utilización de variables vectoriales y como caso particular variables difusas (Brousseau, 2013).

3.5.1 Variables de tipo binario

Solo toman dos valores que frecuentemente se los debe representar por 0 y 1. Estos números simbolizan dos estados opuestos, como si o no, la existencia y la ausencia, bueno o malo, la verdad y la falsedad, que cumple o no cumple, la posesión y la no posesión, etc. La suma por columnas representa el número de sujetos que poseen o satisfacen la propiedad. La suma por filas representa el número de variables satisfechas por el sujeto (Bernard y Charron, 1996).

3.5.2 Variables de tipo modal

Se asocian a valores que son números en el intervalo [0, 1] y describen grados de pertenencia o de satisfacción. Por ejemplo: "Nada satisfecho", "Poco satisfecho", "Neutral", "Muy satisfecho", "Totalmente satisfecho", utilizadas generalmente en cuestionarios de opinión Likert (Canto de Gante et al., 2020), se asumen y se transforman en valores 0, 0,25, 0,50, 0,75, 1 respectivamente.

3.5.3 Variables de tipo frecuencial

Utilizan porcentajes, que se asocian a fenómenos en la escala 0% a 100%, pero que luego se representarán en el intervalo [0, 1] (Zamora-Matamoros et al., 2015). Por

ejemplo, la variable porcentaje de estudiantes que les gusta las matemáticas podrían tomar valores de 0%, 5%, 15%, 50%, que en la escala de [0, 1] serían 0, 0,05, 0,15, 0,50.

3.5.4 Variables de tipo cuantitativo o efectivas

Son conjuntos finitos C de números reales positivos o cero que para aplicar las técnicas del ASI se deben trasladar a la escala [0, 1], donde 0 corresponderá al mínimo valor y 1 al máximo valor del conjunto C, luego para asignar cada valor de x de C al intervalo [0, 1], se realiza una traslación de x+min(C) y luego una contracción con una regla de tres asociando max(C)-min(C) correspondiente a 1, la fórmula final que lleva un x de C a un x' en [0, 1], está dada por x'= (x+min(C))/(max(C)-min(C)).

3.5.5 Variables de tipo intervalo

Son conjuntos finitos C de números reales cualquiera que para aplicar las técnicas del ASI se deben trasladar a la escala [0, 1], donde 0 corresponderá al mínimo valor y 1 al máximo valor del conjunto C, luego para asignar cada valor de x de C al intervalo [0, 1], se realiza una traslación de x+abs(min(C)) y luego una contracción con una regla de tres asociando max(C)-min(C) correspondiente a 1, la fórmula final que lleva un x de C a un x' en [0, 1], está dada por x'= (x+abs(min(C))/(max(C)-min(C)). Luego los x' se ingresan como el caso de las variables frecuenciales. Para indicar que una variable nvar es de intervalo se ingresa de la siguiente manera: nvar i

3.6 Variables por su función

Por el tipo de función que cumplen también pueden ser principales o suplementarias.

3.6.1 Variables suplementarias

Son variables de tipo binario o tipo modal (Gras et al., 2002), se indican con el nombre de la variable seguida de un espacio y la letra s: nvar s. Su función es dar más información descriptiva (Bailleul, 2000) sobre la formación de las categorías. Si por ejemplo se desea conocer si una determinada implicación está en su mayoría influenciada por el semestre al cual pertenecen los estudiantes, se deberá ingresar una nueva variable semestre s, que contendrá información tal como 0, 0,2, 0,4, 0,6, 0,8, 1 que corresponderá a los semestres primero, segundo, tercero, cuarto, quinto y sexto. Esta información se observará únicamente al utilizar las opciones de tipicalidad y contribución (Orús et al., 2005).

3.6.2 Variables principales

Todas las variables en ASI se consideran como principales. Es decir, una variable ingresada en la base de datos por defecto se considera que es principal, o sea que interviene en el cálculo de las contribuciones de las categorías.

3.7 Técnicas de análisis

A continuación, se explica las técnicas utilizadas en ASI, como son la similaridad, cuasi-implicación, cohesión y reducción (Régnier et al., 2020). En 1981, Gras, Lerman y Rostam publican el artículo que sienta las bases de su análisis implicativo para datos en formato binario (Lerman, Gras, et al., 1981). La siguiente evolución aparece con la definición del análisis de cohesiones (Bailleul y Gras, 1994), que gráficamente se representa con un árbol asimétrico a diferencia del dendrograma simétrico utilizado para representar la similaridad.

Finalmente van apareciendo sucesivamente los conceptos de nodos significativos, tipicalidad y contribución.

El ASI engloba tres procedimientos distintos e independientes (Valls, 2014):

- **Análisis de similaridades (clasificatorio):** Clúster analysis con una forma original de medir distancias, las similaridades de Lerman. Se forma un árbol jerárquico similar a un dendrograma (clúster jerárquico ascendente).
- **Análisis Implicativo:** Genera una matriz con todas las implicaciones a→b encontradas en los datos y forma un grafo implicativo, con flechas relacionando las variables con las implicaciones más fuertes.
- **Análisis de cohesiones:** Construye un árbol jerárquico parecido al árbol de similaridades, en el que se marcan de nuevo las variables con implicaciones más fuertes mediante flechas.

3.7.1 Similaridad

Efectúa el análisis de las proximidades en el sentido de la Similaridad (Lerman, Gras, et al., 1981), originando resultados numéricos y presenta el árbol jerárquico de las similaridades.

El análisis de similaridad es una técnica clúster con la que se busca encontrar agrupaciones de objetos similares dentro del conjunto de datos. Para ello se busca una forma de medir dichas similaridades entre ellos y estructurar las variables en un árbol de similaridad mediante jerarquía.

Para formar el árbol de similaridad se calculan los índices de proximidad definidos por: $s(a_i, a_j) = \Pr[Card(X_i \cap X_j) \leq K]$, donde $K = Card(A_i \cap A_j)$ son el número de copresencias observadas entre $a_i \, y \, a_j$ (coincidencias de unos) y el cálculo de la probabilidad (Pr) dependerá de la ley de probabilidad asumida. Considerando que los modelos de Poisson y Binomial se aproximan a la distribución normal, las similaridades de cada pareja (a_i, a_j) al nivel cero de la jerarquía se obtienen mediante: $\Pr\left[\frac{Card(X_i \cap X_j) - \frac{n_{a_i} * n_{a_j}}{n}}{\sqrt{\frac{n_{a_i} * n_{a_j}}{n}}} \leq K_c\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{K_c} e^{-\frac{1}{2}x^2} \, dx$, donde $K_c := \frac{K - \frac{n_{a_i} * n_{a_j}}{n}}{\sqrt{\frac{n_{a_i} * n_{a_j}}{n}}}$ indica la cantidad de copresencias. Los niveles de similaridad de los siguientes niveles (1, 2,… , n) se encuentran: para el siguiente nivel de jerarquía se combinan las clases (a_i, a_j) y se toman aquellas con mayor índice de similaridad. Para las variables aisladas, se utiliza $s((a_i, a_j), a_k) = \left[máx\left\{s(a_i, a_j); s(a_i, a_k)\right\}\right]^2$ y con clases formadas previamente $s((a_i, a_j), (a_{k1}, a_{k2}, a_{k3})) =$

$\left[máx\left\{s(a_i, a_{k1}), s(a_i, a_{k2}), s(a_i, a_{k3}), s(a_j, a_{k1}), s(a_j, a_{k2}), s(a_j, a_{k3})\right\}\right]^{2x3}$. En general, si se tiene dos clases previamente formadas llamadas por ejemplo C_1, C_2, la similaridad entre ellas se encontrará con la expresión $s(C_1, C_2) =$

$\left[\text{máx}\left\{s(a_j, a_k) : a_j \in C_1, a_k \in C_2\right\}\right]^{\text{Card}(C_1) \times \text{Card}(C_2)}$.

Los árboles de similaridad tienen la forma de la Figura 3.2, se ven 60 variables y su gráfico de similaridad respectivo.

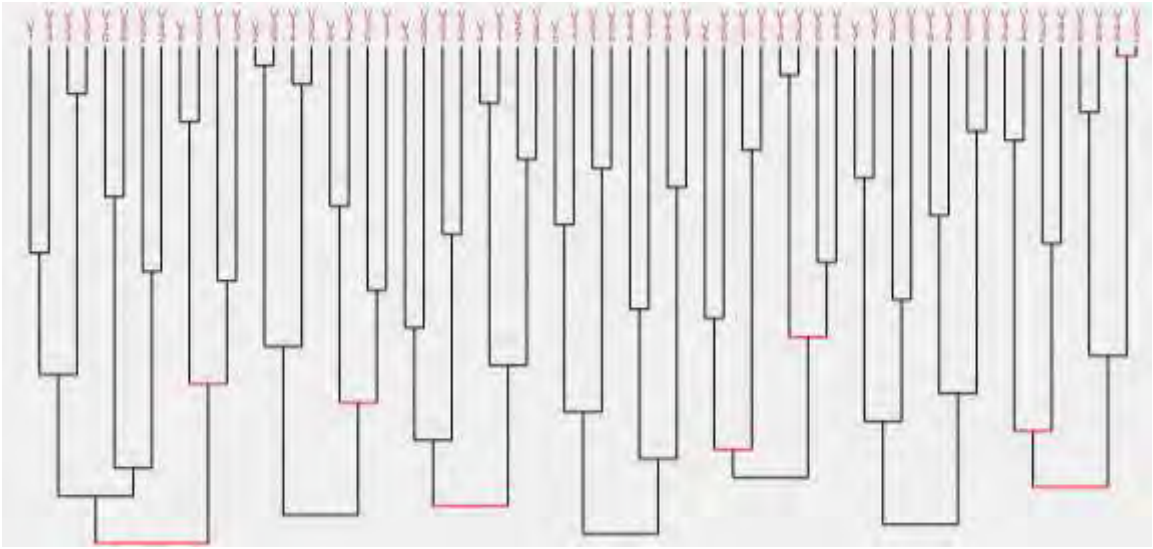


Figura 3.2.- Ejemplo de árbol de similaridad realizado en Rchic

3.7.2 Cuasi-implicación

En la cuasi-implicación al igual que en el análisis de similaridad, se parte considerando un conjunto I formado por n individuos y un conjunto A formado por p características, A = {a_1, a_2,...,a_p}. se supone también que: $A_i = \{\times \in I: a_i(\times) = 1\}, Card(I) = n, Card(A_i) = n_{a_i}$ y $Card(\overline{A_1}) = n_{\overline{a_1}}$.

En la matemática formal y en particular en la lógica matemática, la regla $a_i \Rightarrow a_j$ es verdadera si para todo x, $a_j(x)$ solo es nulo cuando $a_i(x)$ lo sea también; es decir si el conjunto A de los x por los cuales $a_i(x)=1$ está contenido en el conjunto B de los x para los cuales $a_j(x)$ =1. Sin embargo, esta inclusión estricta se observa excepcionalmente en la realidad, especialmente en las ciencias sociales y en particular en la educación, es por ello que nace el concepto de cuasi-implicación $a_i \longrightarrow a_j$ significa que “cuando a_i está presente entonces generalmente a_j está también presente”. La medida de la relación aplicativa se evalúa a partir de la inverosimilitud de la aparición, en los datos, del número de casos que la invalidan, es decir, cuantifica "el asombro" del experto ante el número inverosímilmente pequeño de contraejemplos, en comparación con los que se podrían observar en caso de una ausencia de relación (independencia).

Los nodos internos del árbol dirigido, los cuales representan la jerarquía dirigida, describen relaciones implicativas complejas, llamadas R-reglas, entre los atributos de A.

- Cuando $R \longrightarrow a_i$, a_i se interpreta como una consecuencia de R.
- La R-regla $a_i \longrightarrow R$, significa que una R-regla R puede ser deducida de la observación de a_i.
- La R-regla $R \longrightarrow R''$, significa que la propiedad R''es el corolario de una propiedad previamente definida R

Supongamos que seleccionamos aleatoriamente, de un conjunto I, dos subconjuntos U y V con $n_{a_i}\, y\, n_{a_j}$ elementos respectivamente. Sea $\left[\ X_{a_i \cap \bar{a}_j} \leq n_{a_i \cap \bar{a}_j}\right] \leq \alpha_t$. Donde $n_{a_i \cap \bar{a}_j} = Card\left(A_i \cap \bar{A}_j\right)$ es el número de contraejemplos a la regla $a_i \longrightarrow a_j$ observados en la muestra. Esto significa que el número de contraejemplos observados es pequeño. La distribución de $X_{a_i\ \bar{a}_j}$ depende del patrón de selección aleatorio asumido para seleccionar los conjuntos, pudiendo ser Binomial, Hipergeométrica o Poisson, como se detalla a continuación. Para más información ver (Bodin, 1997). Para simplificar la notación, nos referiremos a las variables (características) a_i y a_j como a y b, y a los conjuntos $A_i = \{x \in I \mid a_i(x) = 1\}\, y\, A_j = \{x \in I \mid a_j(x) = 1\}$ como A y B. Nos basamos en los modelos utilizados por (Zamora et al., 2009)

1. Modelo de la hipótesis de ausencia de relación

Forma de extraer los individuos: Se tiene n_a y n_b individuos, bajo el supuesto de que en el conjunto I existen otros individuos que poseen las características a y b.

Modelo Probabilístico: Del conjunto I se extraen de forma independiente, dos conjuntos X e Y de tamaños n_a y n_b respectivamente. $\Omega = \{(X,Y) \mid X \subset I, Y \subset I, Card(X) = n_a, Card(Y) = n_b\}$, $k_o = Card(A \cap \bar{B})$ y $K = Card(X \cap \bar{Y})$ la variable aleatoria correspondiente.

Los modelos siguientes surgen bajo el supuesto de que el conjunto I de individuos haya sido seleccionado de un conjunto mayor de individuos, que denotaremos por ζ.

2. Modelo de la hipótesis de independencia entre a y b.

Caso 1: El cardinal de ζ es un valor finito N.

Forma de extraer los individuos: Se extraen n individuos de una población ζ de tamaño N, donde N y n se consideran conocidos, mientras que n_a, n_b y $n_{a\wedge\bar{b}}$ son valores observados.

Modelo Probabilístico: Del conjunto ζ de tamaño N, se extrae una muestra I de tamaño n. En esta muestra I, se observan los individuos que poseen las características $a\ (n_a)$ y $b\ (n_b)$ y los que poseen a y no b $(n_{a\wedge\bar{b}})$.

Caso 2: El cardinal de ζ es un valor infinito.

En este caso se extraen infinitas muestras de tamaño **n** de la población ζ y en cada una de estas extracciones, la probabilidad de obtener $K = K_0$ es la misma. Para una muestra de tamaño **n** esta probabilidad se estimaría a través de $p = \frac{n_{a\wedge\bar{b}}}{n}$ la cual, bajo la hipótesis de independencia, se escribe como: $p = \frac{n_a}{n}\frac{n_{\bar{b}}}{n} = \hat{p}(a)\hat{p}(\bar{b})$.

Por lo tanto, $P(K = K_0) = C_{K_0}^n p^{K_0}(1-p)^{n-K_0}$ y K sigue la ley binomial de parámetros n y p, $K \sim B(n,p)$.

Caso 3: El cardinal de ζ es un valor infinito y el cardinal de **I** es indeterminado.

En este caso se trata de extraer de la población ζ de tamaño infinito, una muestra I, cuyo tamaño es aleatorio. Denotemos por M el tamaño de la muestra I, donde M es una magnitud aleatoria, entonces: $P(K = K_0) = \sum_{n=0}^{\infty} P(M = n) * P(K = K_0|M = n)$, para que $K = K_0$ debe cumplirse que $n \geq K_0$, por lo que la expresión anterior puede reescribirse como $P(K = K_0) = \sum_{n\geq K_0}^{\infty} P(M = n) * P(K = K_0|M = n)$.

Supongamos que la variable aleatoria M sigue una distribución Poisson de parámetro $\lambda = n$, valor observado de M. $P(K = K_0|M = n) = C_{K_0}^n p^{K_0}(1-p)^{n-K_0}$, con $p = \frac{n_a}{n}\frac{n_{\bar{b}}}{n}$.

Sustituyendo (3) en (2) y asumiendo que $M \sim P(n)$, obtenemos que: $P(K = K_0) = \sum_{n\geq K_0}^{\infty} \frac{e^{-n}n^n}{n!} * C_{K_0}^n p^{K_0}(1-p)^{K_0} = \frac{e^{-np}(np)^{K_0}}{K_0!} \sum_{n\geq K_0}^{\infty} \frac{e^{-n(1-p)}[n(1-p)]^{n-K_0}}{(n-K_0)!}$, donde el segundo factor vale 1, por lo que: $P(K = K_0) = \frac{e^{-np}(np)^{K_0}}{K_0!}$ es decir, bajo este concepto K sigue la ley

de Poisson de media np, $K \sim Poisson\,(np)$, siendo $np = \frac{n_a - n_{\bar{b}}}{n}$ bajo la hipótesis de independencia. Bajo los modelos considerados se demuestra que: $E(K) = \frac{n_a n_{\bar{b}}}{n}$

La intensidad implicativa de la regla $a_i \rightarrow a_j$ se define como: $\phi(a_i, a_j) = 1 - P\left[K < n_{a_{i\wedge}\bar{a}_j}\right]$, si $n_{a_j} \neq n$,en caso contrario, $\phi(a_i, a_j) = 0$.

La regla es retenida para un α dado si: $\phi(a_i, a_j) \geq 1 - \alpha$, lo cual es equivalente a decir que la regla es admisible, según la definición 3 dada anteriormente, con confianza $1 - \alpha$.

El índice de implicación de la regla $a_i \rightarrow a_j$ se define como: $q(a_i, a_j) = \frac{n_{a_{i\wedge}\bar{a}_j} - \frac{n_{a_i} n_{\bar{a}_j}}{n}}{\sqrt{\frac{n_{a_i} n_{\bar{a}_j}}{n}}}$. Se demuestra que: $\rho(a_i, a_j) = -q(a_i, a_j)\sqrt{\frac{n_{a_i} n_{\bar{a}_j}}{n}}$.

Los grafos de implicación tienen la forma de la Figura 3.3.

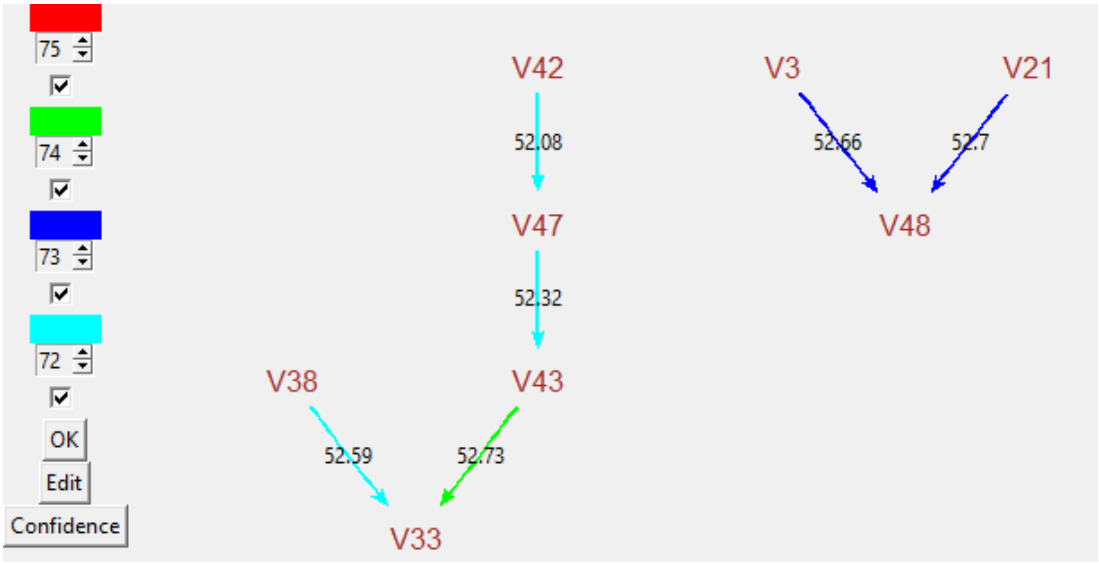


Figura 3.3.- Ejemplo de grafo de implicación en Rchic

En la Figura 3.3 se observa el gráfico de implicación correspondiente a 48 variables y 615 casos.

3.7.3 Cohesión

Se trata de construir un método de clasificación bajo la restricción de basarlo en relaciones no simétricas (del tipo de la relación de implicación) entre variables. Para resolver este problema se construyó el concepto de cohesión implicativa, como indicador del orden implicativo dentro de una clase. Esta cohesión "generalmente alimentada por la coherencia semántica o, en el caso de la didáctica, por las condiciones psicológicas, cognitivas, situacionales, etc., debe traducirse en una medida (cuantitativa)" (Gras et al., 1996).

A la medida de la calidad implicativa de las R-reglas se llama cohesión. Permite descubrir R-reglas del tipo $R' \rightarrow R''$ con una fuerte relación implicativa entre la componente de R' y las de R''. Por ejemplo, es natural formar la regla $(a_1 \rightarrow a_2) \rightarrow (a_3 \rightarrow a_4)$ si las relaciones implicativas $a_1 \rightarrow a_3, a_1 \rightarrow a_4, a_2 \rightarrow a_3$ y $a_2 \rightarrow a_4$ son lo suficientemente significativas.

Esto significa que se debe contrastar con el desorden de una experiencia aleatoria, y se plantea que la entropía es bastante conveniente para medir ese desorden.

Se considera, una regla $a_1 \rightarrow a_1$ de orden 1 y definir la variable aleatoria Y como indicadora del evento: $X_{a_i \wedge \bar{a_j}} \geq n_{a_i \wedge \bar{a_j}}$, luego: $\Pr[Y = 1] = \Pr\left[X_{a_i \wedge \bar{a_j}} \geq n_{a_i \wedge \bar{a_j}}\right] = \varphi(a_i, a_j)$, $\Pr[Y = 1] = 1 - \varphi(a_i, a_j)$.

La entropía de este experimento es $E = -p\log_2 p - (1-p)\log_2(1-p), \text{con } p = \varphi(a_i, a_j)$, que se interpreta como la cantidad media de información que reporta una fuente binaria, con probabilidad de éxito igual a la intensidad implicativa de la regla $a_i \rightarrow a_j$, o el valor medio de la incertidumbre de un observador antes de conocer la salida de una fuente binaria en la que puede o no ocurrir el evento $a_i \rightarrow a_j$.

Se define el grado de una R-regla como la cantidad de variables involucradas en la regla menos 1, por ejemplo, la R-regla $R: a_i \rightarrow a_j$ es de orden 1, la R-regla $R: R \rightarrow a_k$ es de orden 2, y así sucesivamente (Acioly-Regnier y Regnier, 2007).

Se define la cohesión de una R-regla $a_i \rightarrow a_j$ de grado 1 es: $\text{Coh}(a_i, a_j) = \begin{cases} \sqrt{1 - E^2} & \text{si } p \geq 0{,}5 \\ 0 & \text{en caso contrario} \end{cases}$.

Los árboles de cohesión tienen la forma de un dendrograma asimétrico, por ejemplo, en la Figura 3.4 se ve un dendrograma con 40 variables.

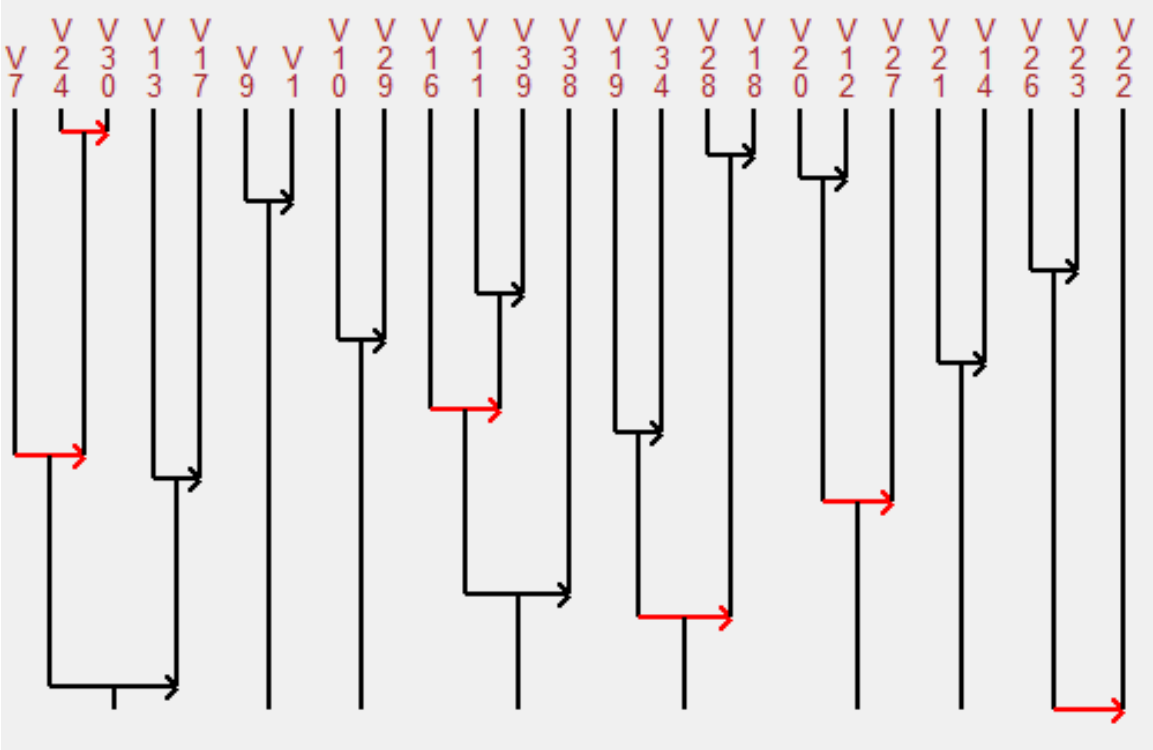


Figura 3.4.- Ejemplo de árbol de cohesión realizado en Rchic

3.7.4 Reducción

La reducción permite disminuir el número de variables en aquellas más representativas, a continuación, se describe la notación para realizar los cálculos respectivos: Sea $A \Rightarrow B$ una regla de asociación entre dos conjuntos de elementos A y B subconjuntos de un conjunto de datos I. Este valor mide la calidad de la regla en función del número de contraejemplos que se ven en la muestra. Donde I_A representa los individuos descritos por la propiedad A, I_B representa los individuos descritos por la propiedad B, I_B representa los individuos no descritos por la propiedad B y $|I_B|$ representa la cardinalidad del conjunto I_B. Por tanto, $I_{A \wedge B}$ representará a los individuos descritos por las propiedades A y B. Mientras que $I_{A \wedge \bar{B}}$ representa los individuos descritos por las propiedades A y por no B, los cuales representan los contraejemplos de la regla $A \rightarrow B$. Por otro lado, la intensidad de la implicación de una regla está definida como el número de veces que la regla no se cumple por un pequeño número de contraejemplos (Arabie et al., 2006). La intensidad de la

implicación que aquí se denominará *ImpInt,* formalmente definida por: $Imp\ Int\ (\mathrm{A} \text{ Þ } \mathrm{B}) = \frac{1}{\sqrt{2\pi}} \int_q^a e^{\frac{t^2}{2}} dt$. donde el límite inferior de integración está dado por: $q = \frac{|I_{A \wedge \bar{B}}| - n\frac{|I_A| * |I_B|}{n^2}}{\sqrt{n\frac{|I_A| * |\bar{I_B}|}{n^2}\left(1 - \frac{|I_A| * |\bar{I_B}|}{n^2}\right)}}$.

Este índice cumple funciones similares a las del ya conocido coeficiente de correlación. Otro término que se empleará es el denominado Validez y que está dado por: $Validez \quad (\mathrm{A} \text{ Þ } \mathrm{B}) = \begin{cases} 1 - E(f_1)^2, & si \quad f_1 \in [0; 0{,}5] \\ 0, & si \quad f_1 \in]0{,}5, 1] \end{cases}$ donde, f_1 es igual a: $\frac{|I_{A \wedge \bar{B}}|}{I_A}$. Y $E(f)$, representa a la función de entropía. Ésta se incorpora para obtener resultados más acordes en muestras grandes.

La Validez Global de la regla está definida en: $GloVal\ (\mathrm{A} \text{ Þ } \mathrm{B}) = [Validez\ (\mathrm{A} \text{ Þ } \mathrm{B}), Validez(-\mathrm{B} \text{Þ } -\mathrm{A})]^{\frac{1}{4}}$. La Utilidad de una regla de asociación $\mathrm{A} \rightarrow \mathrm{B}$ está definida como: $Utilidad\ (\mathrm{A} \rightarrow \mathrm{B}) = \begin{cases} 1, & si\ |I_A \cap I_B| \geq minsup \\ 0, & en\ otro\ caso \end{cases}$

Minsup, representa el mínimo número de individuos que necesita verificar la regla. Para representar la relevancia de una regla se suele emplear el término Relevancia y está dada por: $Relevancia\ (\mathrm{A} \rightarrow \mathrm{B}) = Utilidad(\mathrm{A} \rightarrow \mathrm{B}) * \sqrt{ImpInt(\mathrm{A} \rightarrow \mathrm{B}) * GloVal(\mathrm{A} \rightarrow \mathrm{B})}$ (Arabie et al., 2006).

Lógicamente dos variables A y B son equivalentes si y solo si $\mathrm{A} \rightarrow \mathrm{B}$ y $\mathrm{B} \rightarrow \mathrm{A}$, es decir la *cuasi-equivalencia* es medida por el coeficiente Quasi(A, B) definido por: $\mathrm{Quasi}(\mathrm{A}, \mathrm{B}) = \sqrt{\mathrm{Relevancia}\ (\mathrm{A} \rightarrow \mathrm{B}) \times \mathrm{Relevancia}\ (\mathrm{B} \rightarrow \mathrm{A})}$

Es oportuno aclarar que no solo se puede trabajar con cuasi-equivalencias entre dos variables sino también con cuasi-equivalencias entre clases.

Así, una clase de cuasi-equivalencia de n variables $A_1, A_2, \ldots\ldots\ldots A_n$ es medido por el coeficiente de cuasi-equivalencia de la clase definida en:

$Q\mathrm{uasiequivalenciacla\ se}\ (\mathrm{A}) = \min\{Quasi(A_i, A_j), \forall i = 1{,}2, \ldots\ldots, n-1; \forall j = 1{,}2, \ldots\ldots, n\}$

Como ejemplo se muestra la aplicación del método de reducción.

3.8 Técnicas automatizadas

Iniciamos este apartado con una breve síntesis histórica sobre el software CHIC, cuyas iniciales tienen como origen Classification Hiérarchique Implicative et Cohésitive y actualmente es administrado por el francés Raphael Couturier (Zamora et al., 2009). CHIC es un programa informático propietario desarrollado exclusivamente para el estudio y la aplicación del ASI en la plataforma windows. En sus inicios, tanto CHIC como el ASI solo trataba con variables binarias, enriqueciéndose más tarde con variables modales y frecuenciales. A lo largo de su ciclo de vida, se han ido incorporando todos los análisis que han ido surgiendo en torno al ASI, incluyendo los tres principales: análisis de similaridad, análisis cuasi-implicativo y análisis de cohesiones. Para cada uno de estos análisis proporciona también la posibilidad de realizar los cálculos tanto en la versión clásica de la implicación o en su versión entrópica, elección que influirá en gran medida en las reglas producidas.

Año 1984: El software CHIC que aún no tenía este nombre consistía en una versión primitiva elaborada por Regis Gras en Basic, implementando los cálculos de las intensidades de implicación y, sobre todo, el algoritmo de construcción de la jerarquía de similitudes de I.C. Lerman.

Año 1990: Como parte de la tesis de S. Ag Almouloud, el programa informático CHIC fue una herramienta de software confiable y lo suficientemente amigable, con el fin de tratar el análisis de similaridad de I. C. Lerman, el análisis implicativo de R. Gras y sus extensiones como: la jerarquía implicativa de clase (Larher, 1991), así como el estudio de variables numéricas y modales, el responsable de integrar todas las opciones anteriores fue S. Ag Almouloud.

Año 1992: CHIC fue programado en turbo pascal 6, consta de un programa principal y un conjunto de subprogramas, según el diagrama de flujo (Figura 3.5) correspondiente a la estructura realizada por S. Ag. Almouloud.

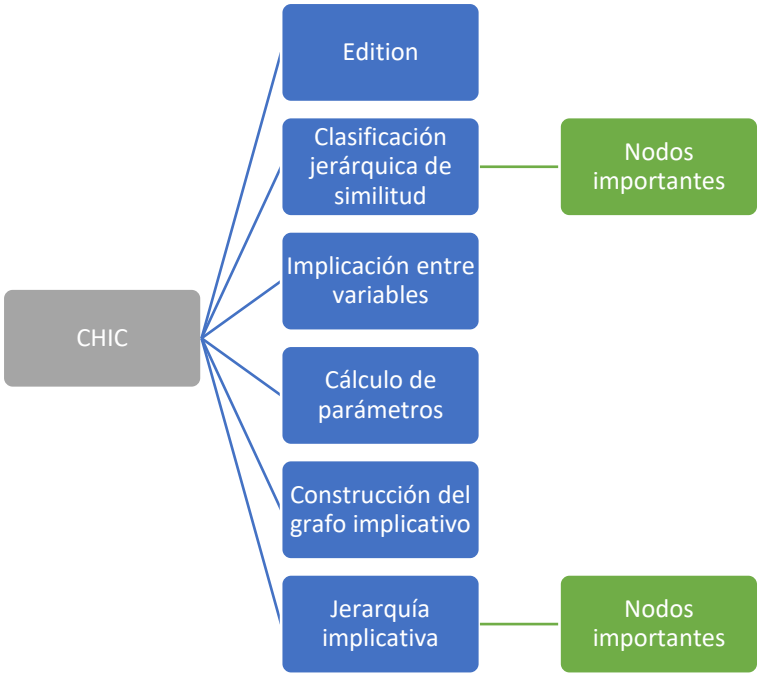


Figura 3.5.- Estructura del software CHIC de 1992 (Couturier y Gras, 2005a)

En la actualidad: El software CHIC fue totalmente programado en C++, por Raphael Couturier y es el encargado de su mantenimiento y sucesivas versiones. La última versión es la 7.00.

Figura 3.6.- Versión 2021 de CHIC

La Figura 3.6, muestra la captura de la última versión del software propietario que automatiza las principales técnicas utilizadas por el análisis estadístico Implicativo.

3.9 Conclusiones

El ASI, propuesto por Regis Gras, es un conjunto de técnicas de análisis no paramétricas multivariadas de similaridad, cuasi-implicación y cohesión (con otras opciones como por ejemplo nodos significativos, tipicalidad, contribución, etc., Capítulo 9), que las hacen atractivas de aportar en LA y pese a que su origen es la didáctica de la matemática se aplica en muchos campos como la psicología, las ciencias, el ADN, el arte, minería de datos y últimamente al tratamiento de imágenes y la investigación cualitativa mediante la organización de categorías en el tratamiento de textos. El ASI automatiza sus herramientas de análisis con el software propietario CHIC y el software libre Rchic, ambos elaborados por Raphael Couturier (Coutrier et al., 2015; Couturier y Gras, 2005b).

Capítulo 4to | APROXIMACIÓN A LOS ELEMENTOS COMUNES DE LA Y ASI

Se muestra que los orígenes de ASI están en la educación matemática y que uno de los campos de aplicación de LA es el aprendizaje, al converger ambas al ámbito educativo admiten elementos comunes.

4 Capítulo.- Aproximación a los elementos comunes de LA y ASI

En este capítulo se determinarán algunos elementos comunes entre LA y el ASI, antes de empezar a comparar las técnicas de ASI y las técnicas del LA, es importante determinar algunos elementos comunes de acuerdo a las publicaciones científicas, lo cual se desarrolla en detalle a continuación.

4.1 Introducción

El objetivo de este capítulo es hacer notar que tanto el Análisis Estadístico Implicativo (ASI) como *Learning Analytics* (LA) tienen un objeto de estudio común que es la educación en general y el aprendizaje en particular. La Sección 4.2 muestra el campo de aplicación de LA partiendo de su definición, en la Sección 4.3 se parte desde los congresos internacionales ASI y otras fuentes para determinar sus campos de aplicación, la Sección 4.4 resalta la relación entre didáctica de la matemática y el ASI y la Sección 4.5 muestra un estudio actual de la literatura científica del ASI para observar su tendencia. Por último, en la Sección 4.6 se concluye con los elementos comunes y por tanto la compatibilidad entre LA y ASI.

4.2 Campo de aplicación de LA, desde su definición

En la actualidad, la mayoría de autores de literatura sobre LA (Pazmiño-Maji Rubén et al., 2021), continúan adoptando la siguiente definición de LA, ofrecida en el 1ª Conferencia Internacional de Analítica de Aprendizaje (*LAK 2011 : 1st International Conference Learning Analytics and Knowledge*, 2011), la traducción se muestra a continuación:

La Analítica de aprendizaje es la medición, recopilación, análisis y comunicación de datos sobre los estudiantes y sus contextos, a efectos de comprender y optimizar el aprendizaje y los entornos en que se producen.

Esta definición de LA vista gráficamente se representa en la Figura 4.1.

Figura 4.1.- Definición de LA vista gráficamente

Desde el punto de vista de los procesos , se podría definir LA como (Ruipérez-Valiente et al., 2015; Chatti et al., 2012) el proceso de medición, recopilación, análisis y comunicación; aplicado a datos sobre los estudiantes, para comprender y optimizar el aprendizaje y su contexto.

Figura 4.2.- Definición de LA como un proceso

Es decir, LA es un proceso para comprender y optimizar el aprendizaje y su contexto (Figura 4.2). De aquí se desprende que uno de los objetivos de LA es la comprensión y optimización del aprendizaje, que es un objetivo de tipo educativo, es decir LA actúa en forma específica en el aprendizaje y en forma general en la educación, se concluye que uno de los campos de aplicación de las Analíticas de Aprendizaje es la educación.

4.3 Campo de aplicación del ASI, desde los congresos internacionales

El objetivo de estudio del ASI desde sus orígenes fue la didáctica de la matemática que se ha ido ampliando a la educación en general y a otros campos, esta evolución se observa en las publicaciones realizadas por ejemplo en los 10 congresos internacionales realizados desde el año 2000 hasta la actualidad (Barragán-Pazmiño y Pazmiño-Maji, 2018). A continuación, se analizan artículos científicos que permiten ver objetivamente que uno de los campos de aplicación del ASI es la educación y como conclusión existe compatibilidad con LA.

La naturaleza del ASI es epistemológico, educativo y didáctico, la distribución de la literatura en la investigación computacional, investigación teórica, aplicaciones educativas y otras aplicaciones (Pazmiño, 2014) hasta el año 2014-2015 se muestra en la Figura 4.3. Se observa que más de la mitad (58%) de los artículos científicos tienen como objetivo de estudio la educación, seguido por el 22% de investigación teórica en el campo del ASI, el 16% son otras aplicaciones que poco a poco se van extendiendo desde la educación hacia otros campos. Las fuentes de esta primera investigación fueron los artículos científicos extraídos de Google académico.

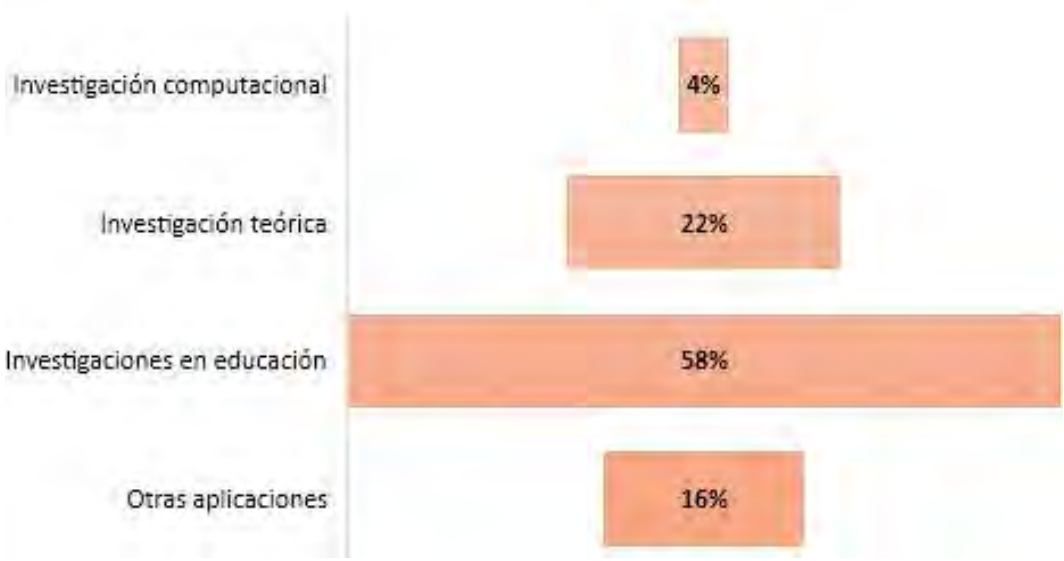


Figura 4.3.- Distribución de la literatura del ASI (R. Pazmiño-Maji, 2014b)

4.4 La didáctica de la matemática y el ASI

Se define la didáctica de la matemática como el estudio de los fenómenos ligados a la transmisión de conocimientos disciplinarios. Los didactas consideran constantemente al ASI como un método útil de análisis de datos, la razón son las reglas de cuasi-implicación que puede encontrar y que pueden interpretarse como reglas que conectan acciones, hechos y características individuales o grupales, mediante las técnicas de similaridad y sobre todo de implicación; brindando al docente la posibilidad de diagnosticar y mejorar el aprendizaje.

A continuación, se muestran los principales problemas en los que estas reglas tienen significado en didáctica de la matemática desde el artículo (Lahanier-Reuter, 2008) de Lahanier-Reuter. Para ejemplificar la regulación de los comportamientos observables de los estudiantes, se realiza un estudio de las respuestas de los estudiantes de los niveles CM1 y CM2 (9 a 10 años) a un ejercicio que se compone de dos tareas sucesivas. La primera, se pide a los estudiantes que pongan en orden los siguientes decimales y fracciones 1,2; -5,9; -7,5; -4; -9,5; -12; -5,15; -1/2; -2,5 y la segunda el colocarlos en una línea graduada. El estudio de los resultados de los estudiantes hace evidentes las diversas estrategias utilizadas para responder a las dos preguntas. Para ordenar la escritura numérica, algunos estudiantes usaron una clasificación estrategia por "tipos de escritura", al clasificar primero las fracciones, luego los números enteros que no tienen punto decimal y finalmente los números decimales. Los alumnos que adoptan esa clasificación toman en cuenta solo la longitud de los números tal como están escritos (Gras, 1991).

Aplicando ASI, resulta un grupo de asociación que permite descubrir las siguientes reglas:

- “Adoptar, definitivamente, una clasificación o escritura por tipos de escritura” implica, “aceptar una falta de acuerdo entre los dos pedidos producidos” (99%).
- “Trabajar, en definitiva, sobre la línea graduada, como una línea de escritura” implica obtener “dos órdenes coherentes, incluso si son erróneas” (95%).
- "Trabajando, definitivamente en la línea graduada, como una línea de escritura" implica obtener “dos órdenes coherentes, incluso si son erróneas” (95%).

Usando una problemática central en didáctica matemática, se mostró brevemente la eficiencia de tratar con las técnicas ASI. Las reglas de cuasi-implicación establecidas por ASI pueden prestarse fácilmente a interpretación en términos de regulación de la acción, los caminos implicativos pueden leerse en términos de redes.

Por último, la asimetría entre variables expone hipótesis explicativas de cierto fenómeno perteneciente a la enseñanza y el aprendizaje. También se plantea a través de las relaciones detalladas de ejemplos de investigación y comportamientos metodológicos particulares.

4.5 Sobre la literatura científica actual del ASI

El artículo titulado “Scientific literature on Implicative Statistical analysis: a systematic mapping of the decade that passes” (Barragán-Pazmiño y Pazmiño-Maji, 2018) se elaboró en el año 2018 con el fin de conocer la situación ASI mediante el estudio de la producción científica producida en lo que va de la década en las fuentes IEEE Explorer, Web of Science, Science Direct, Springer, SciELO, La Referencia, Dialnet, EBSCO, ACM, LatIndex, DartEurope, CSIC, PDQT, TDR, DOAJ y ACM; además se tomó en cuenta los Coloquios Internacionales de ASI 6, 7, 8 y 9 realizados en los años 2012, 2013, 2015 y 2017 respectivamente. El estudio se realizó en el período 2011-2017, mediante el análisis de 121 documentos obtenidos mediante un procedimiento sistemático contando con criterios de inclusión, exclusión y calidad. Se concluyó que el Análisis Estadístico Implicativo es mayormente aplicado en el campo de la educación (Figura 4.4) con mayor producción realizada por investigadores de Francia, y con más de la mitad de las publicaciones en francés (Figura 4.5).

Se observó, además, que los documentos en idiomas distintos al francés van creciendo en número, al igual que documentos en otros campos de aplicación aparte de la educación y con un mayor número de países en los que se está realizando nueva producción científica usando el ASI.

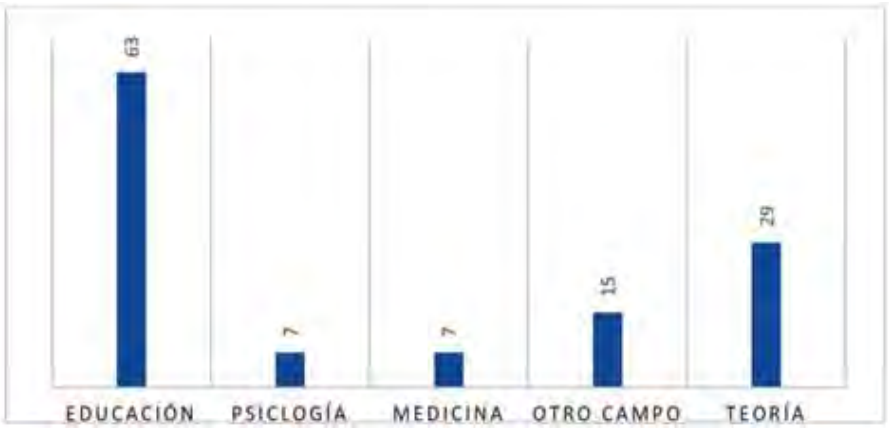


Figura 4.4.- Artículos de ASI según su campo de Aplicación (Barragán-Pazmiño y Pazmiño-Maji, 2018)

El artículo (Barragán-Pazmiño y Pazmiño-Maji, 2018), muestra además que 63 de 121 artículos científicos (52%) pertenecen al campo educativo (Figura 4.4) y aproximadamente el 24% de artículos científicos tratan del estudio de la teoría ASI, esto hace notar el constante crecimiento del desarrollo teórico de esta teoría.

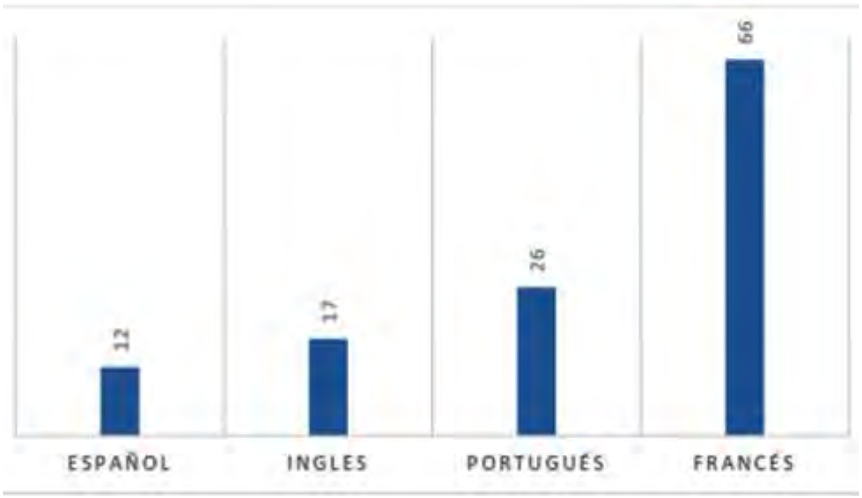


Figura 4.5.- Artículos del ASI según el idioma del documento (Barragán-Pazmiño y Pazmiño-Maji, 2018)

Como se observa en la Figura 4.6, la producción de artículos científicos sobre el ASI en su mayoría es desarrollada o aplicada por investigadores afiliados a instituciones en Francia con 62 documentos (51%), seguido de los documentos producidos en Brasil con 22 documentos (18%) y después España con nueve documentos (7%). Argelia, Chipre y Ecuador les siguen con cuatro documentos cada uno (3%), después se observan a países de Cuba, Italia, México, Nueva Zelanda y Vietnam con dos documentos cada uno (aproximadamente 2%) y finalmente los países que han producido un artículo ASI durante el periodo 2011-2017, Argentina, Colombia, Gabón, Grecia, Madagascar y Túnez que corresponde aproximadamente al 1%.

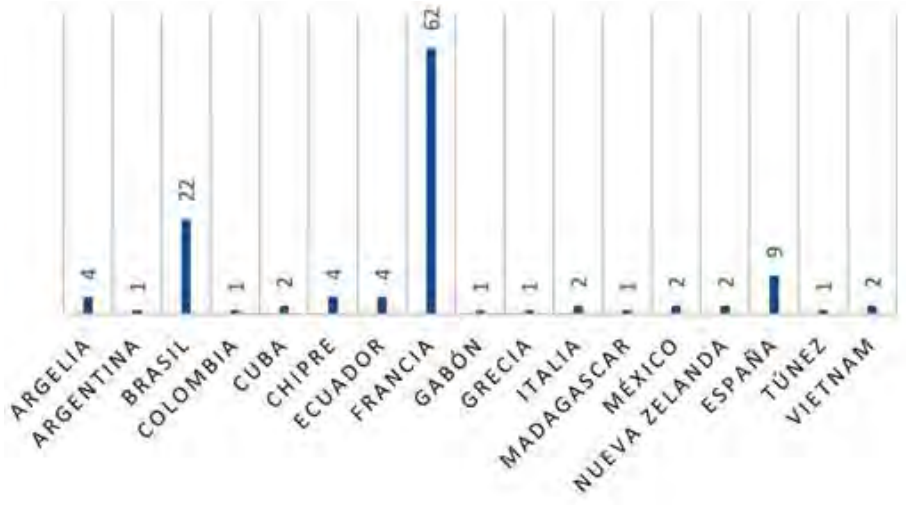


Figura 4.6.- Artículos de ASI según el país de afiliación del autor (Barragán-Pazmiño y Pazmiño-Maji, 2018)

Además del estudio en las áreas clásicas del ASI, también se encuentran algunas tendencias de aplicaciones en la minería de textos (Khaled y Couturier, 2015), aplicaciones a la investigación cualitativa, análisis de ítems (Couturier & Pazmiño, 2016) y últimamente en el análisis de imágenes (R. Pazmiño-Maji et al., 2014). También existe relación entre el ASI y teoría de probabilidades, teoría de inferencia estadística, estadística no paramétrica, teoría de la optimización, estadística multivariante y teoría de sistemas dinámicos, que se pueden encontrar en (Iurato, 2012). Es importante notar que se han estado realizando desde hace algún tiempo aplicaciones en el KDD (Knowledge discovery in databases) por parte del ASI, formalmente el aporte del ASI al KDD y a la minería de datos se lo hizo con artículo titulado “Statistical implicative analysis

approximation to KDD and data mining: A systematic and mapping review in knowledge discovery database framework" (R. Pazmiño-Maji et al., 2017c).

4.6 Conclusiones

Como se deduce de lo tratado en párrafos anteriores, tanto el Análisis Estadístico Implicativo, así como Analíticas de Aprendizaje tienen como uno de los elementos comunes a la educación, ya que comparten el mismo campo de aplicación, abriendo de esta forma la posibilidad de que tenga sentido el analizar los posibles aportes del Análisis Estadístico Implicativo a las Analíticas de Aprendizaje. En LA y ASI se observa que uno de los campos de aplicación para el ASI y LA es la educación, esto muestra una primera aproximación a sus elementos comunes. La cercanía entre ASI y LA se deduce con los estudios que indican que el 58% en 2014 y el 52% de artículos ASI en el 2018 tratan sobre educación, que permitirán aprovechar las experiencias del ASI y de esta forma diversificar y fortalecer la investigación en las Analíticas de Aprendizaje. La Sección 4.4, hizo explícita la relación y el origen educativo del ASI. El análisis de la definición de LA permite observar que LA actúa en el aprendizaje y en su contexto, así mismo el ASI tiene sus orígenes y principales aplicaciones en la educación matemática, por lo que LA y el ASI tienen en común la educación y esto los hace compatibles para continuar su estudio.

C a p í t u l o 5to | APORTES DEL ASI A LA

Se determina los primeros aportes del ASI a LA mediante dos revisiones sistemáticas realizadas en los últimos 11 años (2011-2021) considerando su definición, fuente de datos y etapas.

5 Capítulo.- Aportes del ASI a LA

A continuación, se muestra la aproximación (aportes) del ASI a LA mostrada en los resultados del artículo científico "Approximation of *Statistical Implicative Analysis* to *Learning Analytics*: a systematic review" (Pazmiño-Maji et al., 2016), donde se muestra la aproximación (aportes) del ASI y las Analíticas de Aprendizaje (LA).

El objetivo principal de este capítulo es determinar y cuantificar la aproximación (aportes) de ASI a LA, para ello se diseñó un marco de aproximación para *Learning Analytics* basado en su definición, fuente de datos y etapas.

La revisión sistemática realizada por el autor a mediados del 2016 se actualiza con otra revisión sistemática realizada en el año 2021 y se muestra con el subtítulo Aportes del ASI a LA: junio 2021.

5.1 Introducción

La cantidad de investigaciones en *Learning Analytics* están aumentando cada día, siendo necesaria la integración de nuevas herramientas, métodos y teorías. El objetivo de este capítulo es estudiar la aproximación de la teoría de Análisis Estadístico Implicativo (ASI) a *Learning Analytics*. Para ello, se ha creado un marco de aproximación basado en la definición, etapas y métodos utilizados en LA. En total, se compararon 3 criterios de enfoque y 36 subtemas. Utilizamos la revisión sistemática de literatura publicada en los 66 meses desde enero del 2011 hasta junio del 2016, en las bases de datos bibliográficas ACM, EBSCO, Google Scholar, IEEE, ProQuest, Scopus y WOS. Empezamos con 319 artículos y finalmente 24 fueron los que cumplieron con todos los criterios de calidad. Este documento contiene los temas mediante los cuales ASI contribuye a las LA, además proporciona los porcentajes por categoría de aproximación (Pazmiño-Maji et al., 2016).

Learning Analytics ha sido y sigue siendo un campo de investigación emergente, como se indica en la publicación de Horizon Report 2016 Kitchenham et al. (2010). El tiempo de adopción de *Learning Analytics* indicado en este informe fue de un año o menos, pero nos preguntamos ¿Cuántas instituciones, profesores, estudiantes y métodos de análisis están listos para esta adopción?

En el documento sobre gestión de rendimiento empresarial ("*Learning Analytics* in Enterprise Performance Management | Analytics | Business Intelligence" 2016), se clasificaron a las organizaciones en tres generaciones de madurez de *Learning Analytics* basados en su nivel de aplicación. En la generación 1, descriptivo y parcialmente diagnóstico se encontraron el 90% de las organizaciones; en la generación 2, descubrimiento y parcialmente predictivo se encontraron del 5-10% de las organizaciones y en la generación 3, parcialmente predictivos y prescriptivos no se encontró a ninguna organización. Esto quiere decir que casi el total de las organizaciones se encuentra apenas en una madurez inicial, es decir en la generación 1, descriptivo y parcialmente diagnóstico, la Figura 5.1, muestra en un gráfico de pastel lo antes indicado.

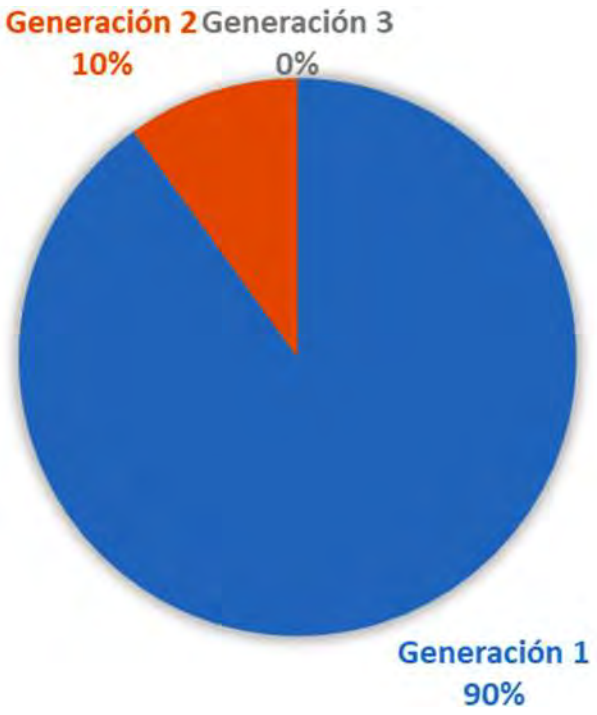


Figura 5.1.- Porcentaje de organizaciones en las tres generaciones en LA (Pazmiño-Maji et al., 2016)

(Bichsel, 2012), propone un modelo de análisis de madurez para evaluar el progreso en el uso de analíticas académicas y analíticas de aprendizaje. En los avances, han producido resultados positivos, pero la mayoría de instituciones están por debajo del 80%. La

mayoría de instituciones también obtuvieron resultados bajos en herramientas de análisis de datos, informes y experiencia (Bichsel, 2012).

Además, se debe considerar que con los métodos de análisis tanto en minería de datos como en *Learning Analytics* son actividades constantes la adaptación, optimización y renovación (Papamitsiou & Economides, 2014). Algunas de las actividades son: precisión, exactitud, sensibilidad, coherencia, medidas de ajuste, confianza, elevación y ponderaciones de similitudes.

En la Sección 5.2 se describe el marco de aproximación utilizado. La Sección 5.3 muestra las etapas de la revisión sistemática. En la Sección 5.4 se encuentran los resultados alcanzados.

5.2 Aportes del ASI a LA: Desde 2011 hasta junio 2016

A continuación, se muestran los aportes de ASI a las analíticas de aprendizaje en un estudio realizado a mediados del año 2016.

5.2.1 Marco de aproximación

Para formar nuestro marco de aproximación, se utilizó primeramente la definición de analíticas de aprendizaje proporcionada en la primera Conferencia Internacional sobre *Learning Analytics* and Knowledge (*LAK 2011: 1st International Conference Learning Analytics and Knowledge*, 2011).

5.2.1.1 Las entradas

1) Las entradas, según Bernhardt. V. (*using-data-improve-student-learning*, 2016), los datos sobre los estudiantes y sus contextos pueden ser:

- *Aprendizaje de los estudiantes* (¿Cuál es el puntaje de los estudiantes en un examen de la escuela?, ¿Hay diferencia en los resultados de los estudiantes en pruebas estandarizadas en diferentes años?).
- *Demografía* (¿Cuántos estudiantes están matriculados en la escuela por año?, ¿Cómo ha cambiado la matrícula en la escuela durante los últimos cinco años?).
- *Percepción* (¿Cuál es el grado de satisfacción de los padres, los estudiantes y personal con el ambiente de aprendizaje?, ¿Cómo han cambiado la percepción de los estudiantes del ambiente de aprendizaje en el tiempo?).

- *Procesos de la institución educativa* (¿Qué programas están en funcionamiento en la institución educativa este año?, ¿Qué programas han funcionado en la institución educativa en los últimos cinco años?).

5.2.1.2 El proceso

2) El proceso, formado por cuatro pasos: medición, recopilación, análisis y comunicación. Además, Campbell describe las analíticas académicas como un "motor de decisiones o acciones" y define cinco pasos: capturar, informar, predecir, actuar y refinar (Campbell & Oblinger, 2007), las etapas de Campbell son las que utilizamos también en este capítulo.

5.2.1.3 Las salidas

3) Las salidas, la definición indica que los efectos son comprender y optimizar el aprendizaje y los entornos. Comprender y optimizar el aprendizaje a menudo usa simple e intuitiva, dinámicas y representaciones de análisis de datos adaptable, por ejemplo, mapas, tablas, cuadros, gráficos, diagramas, etcétera.

Usando la definición, etapas y métodos de LA, construimos un marco de aproximación de ASI a las LA. La Figura 5.2, contiene los componentes (definición, etapas y métodos) y los subcomponentes.

5.2.2 Representación gráfica del marco de aproximación

En la Figura 5.2, los objetos azules representan tres partes de la definición (entrada, proceso y salida). La flecha representa el flujo de las etapas. Los métodos de análisis actúan directamente en las etapas de informe (REPORT) y predicción (PREDICT). La Figura 5.2, también ilustra la relación entre la definición, etapas y métodos de las LA, representa gráficamente el marco de aproximación utilizado.

Figura 5.2.- Representación gráfica del marco de aproximación (Pazmiño-Maji et al., 2016)

5.2.3 Marco de aproximación detallado

En la Tabla 5.1, se aprecia en forma detallada el marco de aproximación basado en la definición, etapas y la fuente de datos de *Learning Analytics*:

Tabla 5.1.- Marco de aproximación detallado (Pazmiño-Maji et al., 2016)

DEFINICIÓN (*LAK 2011 : 1st International Conference Learning Analytics and Knowledge*, 2011)	**CATEGORÍAS Y SUBCATEGORÍAS DE LA DEFINICIÓN**	**ETAPAS** (Campbell & Oblinger, 2007)	**FUENTE DE DATOS** (*using-data-improve-student-learning*, 2016)
Learning Analytics son la medición, recopilación, análisis y comunicación	EL PROCESO • medición, • recopilación, • análisis • comunicación	2 INFORMAR 3 PREDECIR	
de datos sobre los estudiantes y sus contextos	LOS INGRESOS • Datos estudiantes • Datos contexto	1 CAPTURAR	• Aprendizaje de los estudiantes • Demografía • Percepción • Procesos de la institución educativa
a efectos de comprender y optimizar el aprendizaje y los entornos en que se producen	LAS SALIDAS • Comprensión y optimización del aprendizaje y de su entorno	4 ACTUAR 5 REFINAR	

5.2.4 Etapas en la revisión sistemática de literatura

La metodología utilizada fue la revisión sistemática de literatura de investigaciones empíricas (Okoli y Schabram, 2010) sobre Análisis Estadístico Implicativo durante 66 meses, desde enero del 2011 hasta junio del 2016.

5.2.4.1 Preguntas de investigación

Esta sección, que tiene como objetivo determinar y resumir información descriptiva sobre la aproximación de ASI a LA, aborda tres preguntas de investigación (Li, Lam, & Lam, 2015):

PI1: ¿Qué partes de la definición de *Learning Analytics* se observan en los artículos sobre Análisis Estadístico Implicativo?

PI2: ¿Cuáles son las fuentes de datos de *Learning Analytics* observadas en los artículos sobre Análisis Estadístico Implicativo?

PI3: ¿Cuáles de las cinco etapas de análisis en *Learning Analytics* se observan en los artículos sobre Análisis Estadístico Implicativo?

5.2.4.2 Preguntas cortas de investigación

PI1: ¿Qué artículos del ASI están en la definición de las LA?

PI2: ¿Cuáles datos del ASI están en las fuentes de datos de las LA?

PI3: ¿Cuáles etapas del ASI están en las cinco etapas de análisis de las LA?

En las respuestas a las preguntas de investigación, además de identificar los artículos científicos se añadirán los porcentajes de pertenencia individuales, parciales y totales de cada una de las preguntas de investigación.

5.2.4.3 Bases de datos bibliográficas utilizadas

Se examinaron los artículos científicos sobre ASI, cuya fuente fueron siete bases de datos bibliográficas: Biblioteca Digital ACM, EBSCO, Google Scholar, librería digital IEEE, ProQuest, base de datos de resúmenes y citación Scopus de Elsevier y la base de datos internacional Web of Science (WOS). La revisión se limitó a estudios publicados en los últimos 66 meses, entre 2011 y el primer semestre del 2016. Las fuentes de datos y las características de la búsqueda se resumen en la Tabla 5.2.

Tabla 5.2.- Características de la base de datos y búsqueda bibliográficas (Pazmiño-Maji et al., 2016)

BASE DE DATOS BIBLIOGRÁFICAS		ACM, EBSCO, Google Scholar, IEEE, ProQuest, Scopus, WOS
BÚSQUEDA	Principales criterios de búsqueda	"*Statistical implicative analysis*" ASI
	Resultados (aplicando los criterios de exclusión y el tiempo):	24 documentos científicos sobre ASI
	Tiempo	66 meses desde enero del 2011 hasta junio del 2016
	Tópicos: Criterios de aproximación	• La definición • Las fuentes de datos • Las etapas

Durante la búsqueda, el subtema (o sinónimo) se ha guiado por la Tabla 5.3.

Tabla 5.3.- Criterios de examinación utilizados (Pazmiño-Maji et al., 2016)

PREGUNTA DE INVESTIGACIÓN	CRITERIOS DE APROXIMACIÓN	SUBCRITERIOS
1	Aproximación de ASI a las LA, definición y fuente de datos	**Entrada LA:** * Aprendizaje * Contexto del aprendizaje * Fuente de datos **Proceso LA:** *medición, *recopilación *análisis *comunicación **Salidas LA:** * comprensión * optimización
2	Datos de ingreso	* Aprendizaje del estudiante * Demografía * Percepciones * Procesos escolares
3	Aproximación de ASI a las etapas de las LA	* Captura * Informar * Predecir * Actuar * Refinar

5.2.4.4 Cadenas lógicas de búsqueda

El grupo de estudios primarios se definió basándonos en (Zhang & Ali Babar, 2010). La cadena de búsqueda final se describió como sigue: ("*Statistical implicative analysis*" OR SIA) AND **(**LIMIT-TO (PUBYEAR, 2016) OR LIMIT-TO (PUBYEAR, 2015) OR LIMIT-TO (PUBYEAR, 2014) OR LIMIT-TO (PUBYEAR, 2013) OR LIMIT-TO (PUBYEAR, 2012) OR LIMIT-TO (PUBYEAR, 2011)**)** como se indica en (Kutvonen, 2008; Tolk, Turnitsa, & Diallo, 2006), las publicaciones del primer semestre del 2016 y en el área de educación fueron filtradas después utilizando el software de gestión de referencias EndNote.

5.2.4.5 Criterios de inclusión y exclusión

Al final de la etapa de recopilación de datos, se aplicó rigurosamente tanto los criterios de inclusión, así como los criterios de exclusión (Tabla 5.4).

Tabla 5.4.- Criterios de inclusión y exclusión (Pazmiño-Maji et al., 2016)

INCLUSIÓN	EXCLUSIÓN
• Artículos publicados en bibliotecas digitales: ACM, EBSCO, Google Scholar, IEEE, ProQuest, Scopus, WOS. • Artículos publicados sobre educación, los artículos deben presentar resultados cuantitativos y utilizar el Análisis Estadístico Implicativo. • Trabajos presentados en idioma español, italiano, francés y portugués se consideran si existe la traducción al inglés.	• Capítulos de Libros, • Artículos publicados sobre Análisis Estadístico Implicativo, pero en el área de investigación computacional, en el área de teoría de la investigación, revisiones de literatura, históricos u otros similares. • Artículos que no presentan datos empíricos. • Artículos presentados en idioma distinto del inglés.

5.2.4.6 Criterios de calidad

Los criterios de calidad utilizados en la selección de la literatura seleccionada fueron:

- **Claridad en la metodología utilizada:** objetivo, datos, población de estudio, métodos de análisis, software, resultados y publicación de resultados.
- **Suficiencia de resultados:** gráficos, figuras, tablas y discusión.

Los pasos que a continuación se siguieron son:

1) Lectura y análisis de los artículos que cumplieron los criterios de calidad. Se elaboró una base de datos de artículos científicos en una hoja electrónica en Excel.
2) Registrar los subtemas que se encuentran para la definición, etapas y métodos de L.A.
3) Se utilizaron métodos estadísticos descriptivos para analizar, representar e interpretar los resultados.
4) Por último, se utilizaron métodos estadísticos descriptivos para la síntesis de la revisión.

5.2.4.7 Proceso de selección de artículos

En la Figura 5.3, se muestra el proceso de selección de los artículos finales, 7,5% de los originales que se leyeron a profundidad y a los cuales se aplicó el resto de las etapas de revisión sistemática de literatura.

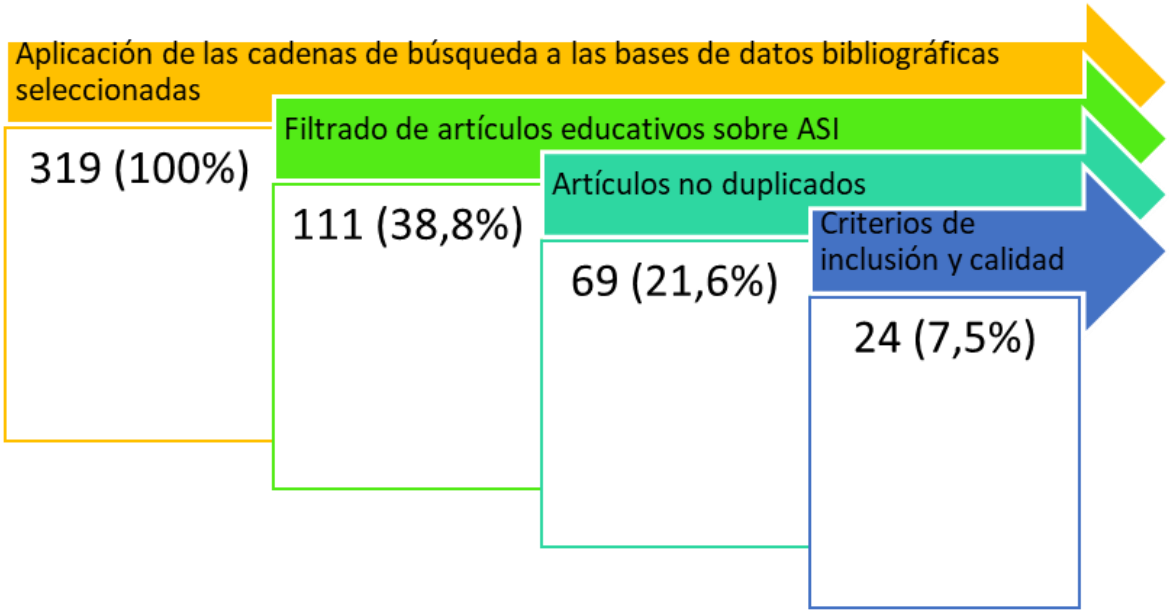


Figura 5.3.- Proceso de selección de artículos (Pazmiño-Maji et al., 2016)

Luego de aplicar las cadenas lógicas de búsqueda y el software de gestión de referencias EndNote para seleccionar los artículos sobre ASI en el primer semestre del 2016, se generaron un total de 319 artículos científicos. De los 319, 111 eran artículos científicos sobre educación, 42 de los cuales fueron registros duplicados generados en las diferentes bases de datos bibliográficas, por último 24 artículos científicos ASI cumplieron con todos los criterios de inclusión y calidad.

El gráfico de línea de tiempo sobre los artículos seleccionados de ASI en los 66 meses desde el 2011 hasta el primer semestre del 2016 se muestran en la Figura 5.4.

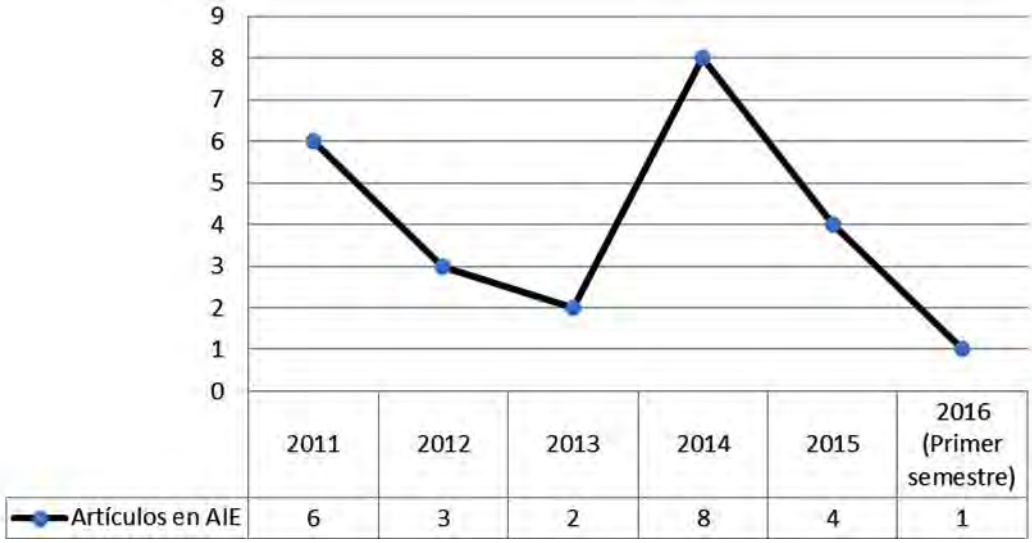


Figura 5.4.- Artículos sobre ASI por año (Pazmiño-Maji et al., 2016)

En la primera búsqueda en las bases de datos bibliográficas se utilizaron las palabras clave: "*Statistical Implicative Analysis*" o su contracción francesa "ASI", después se aplicó EndNote (herramienta de software estándar para la publicación y gestión de bibliografías, citas y referencias) para detectar registros del primer semestre del 2016 sobre educación y duplicados. Los criterios sobre fechas, área educativa y subtemas se encuentran en detalle en la Tabla 5.5 a continuación.

Tabla 5.5.- Resultados de los criterios de búsqueda e inclusión (Pazmiño-Maji et al., 2016)

BASE DE DATOS BIBLIOGRÁFICA	BÚSQUEDA PRINCIPAL: "*Statistical implicative analysis*" or ASI	DOCUMENTOS QUE CUMPLEN LOS CRITERIOS DE INCLUSIÓN	NO DUPLICADOS	ENTRE EL 2011 Y EL PRIMER SEMESTRE DEL 2016
ACM	1	0	0	0
EBSCO	12	6	4	2
Google Scholar	224	86	49	11
IEEE	1	1	1	1
ProQuest	15	8	8	6
Scopus	60	6	4	1
WOS	6	4	3	3
Total	319	111	69	24

5.2.5 Resultados de la pregunta de investigación 1 (PI1)

La Tabla 5.6, muestra las referencias a los artículos científicos sobre ASI, cantidad y frecuencia y que están enmarcados en la definición de *Learning Analytics*.

Tabla 5.6.- Artículos ASI que están enmarcados en la definición de las LA (Pazmiño-Maji et al., 2016). Ver referencias completas en Apéndice

SUBCATEGORÍAS DE DEFINICIÓN LA	ARTÍCULOS EN CATEGORÍA DE DEFINICIÓN LA	NÚMERO	FREQ (%)
Categoría ingresos:			
Estudiantes	[1], [2], [3], [9], [10], [14], [16], [17], [18], [19], [57], [28], [31], [32], [35], [36], [38], [42], [44], [47], [54], [56]	22	91,7%
Contexto de aprendizaje del estudiante	[40], [45]	2	8,3%
	Promedio categoría ingresos:		**50%**
Categoría procesos:			
Medición	[1], [2], [3], [9], [10], [14], [16], [17], [18], [19], [57], [28], [31], [32], [35], [36], [38], [40], [42], [44], [45], [47], [54], [56]	24	100%
Recolección	[[1], [2], [3], [9], [10], [14], [16], [17], [18], [19], [57], [28], [31], [32], [35], [36], [38], [40], [42], [44], [45], [47], [54], [56]	24	100%
Análisis	[[1], [2], [3], [9], [10], [14], [16], [17], [18], [19], [57], [28], [31], [32], [35], [36], [38], [40], [42], [44], [45], [47], [54], [56]	24	100%
Informes	[14], [57], [40], [42], [44]	5	20,8%
	Promedio categoría procesos:		**80,2%**
Categoría salidas:			
Comprender	[1], [2], [3], [9], [10], [16], [17], [18], [19], [57], [28], [31], [32], [35], [36], [38], [45], [47], [54], [56]	20	83,3%
Optimizar	[14], [40], [42], [44]	4	16,7%
	Promedio categorías salidas:		**50%**
	Promedio definición de Analíticas de Aprendizaje (LA)		**61,1%**

5.2.5.1 Categoría ingresos

La Figura 5.5, muestra que los datos de ingreso en casi todos los artículos leídos tratan sobre valoraciones realizadas directamente a los estudiantes (91,7%), muy poco se analiza el contexto de aprendizaje del estudiante (8,3%). El fenómeno anterior no tiene que ver con las técnicas de análisis de datos utilizadas en el ASI, más bien es una

tendencia educativa. Las LA están motivando a que se analice también el entorno del estudiante y en él se podrían aplicar las técnicas del ASI.

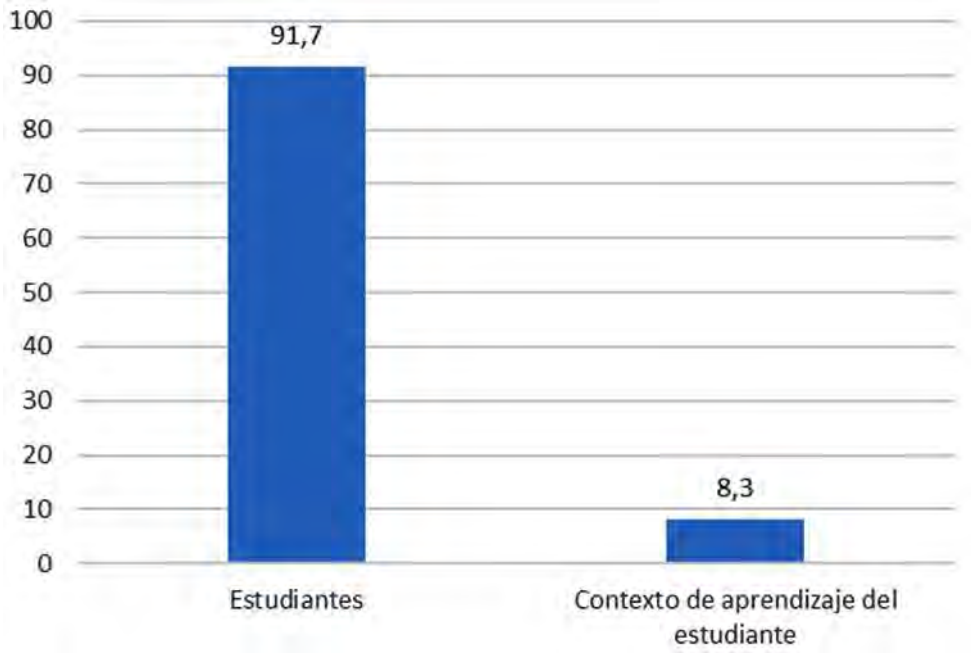


Figura 5.5.- Diagrama de Barras sobre la categoría ingreso de datos en la definición de las LA (Pazmiño-Maji et al., 2016)

Siendo ASI el conjunto de los artículos científicos que utilizan Análisis Estadístico Implicativo y LA, el conjunto formado por los elementos del conjunto anterior (unido con los artículos científicos que utilizan *Learning Analytics*) y que siguen la definición de las LA respecto a la categoría ingresos. Utilizando la teoría de conjuntos habría una relación de intersección no vacía respecto a la definición de las LA respecto a la categoría ingresos que se representa por la fórmula:

$$\underbrace{AEI \cap AA \neq \emptyset}_{Definición,ingresos\ (50\%)}$$

Utilizando la frecuencia porcentual hay un 50% de casos que validan la fórmula anterior. También gráficamente mediante los diagramas de Euler Venn se expresa la relación anterior y ayuda a comprender de mejor manera como se muestra en la Figura 5.6

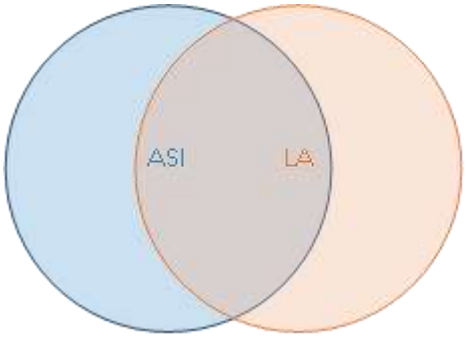


Figura 5.6.- Diagramas de Euler desde el punto de vista de los datos de ingreso (Pazmiño-Maji et al., 2016)

5.2.5.2 Categoría procesos

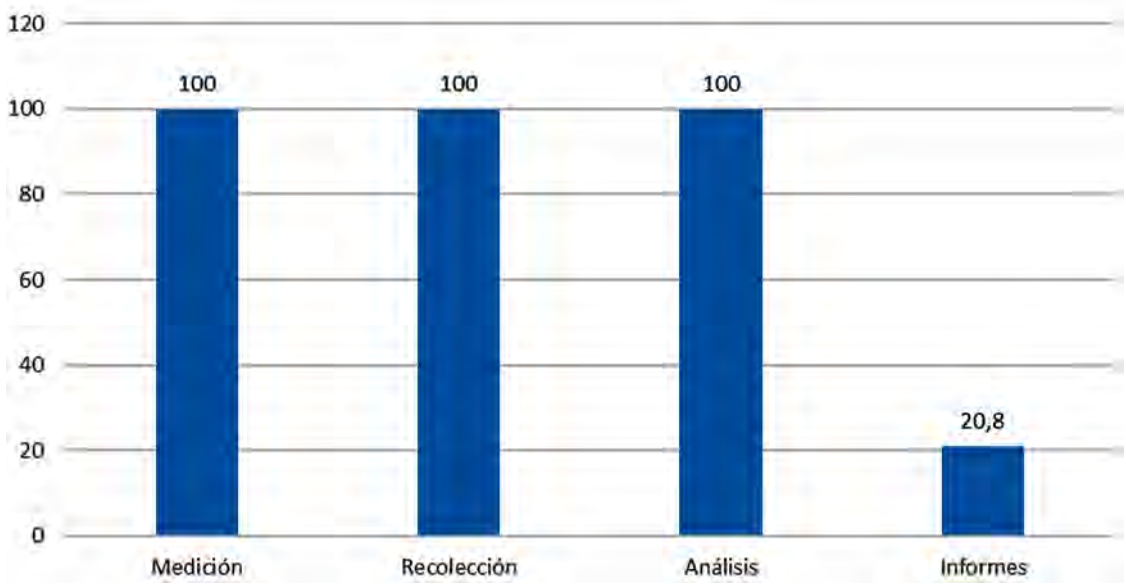


Figura 5.7.- Diagrama de Barras sobre la categoría procesos en la definición de las LA (Pazmiño-Maji et al., 2016)

La categoría más frecuente es la de *procesos* con un total de 77 artículos equivalente al 80,2% de artículos del ASI que están en la categoría procesos de la definición de las LA, las subcategorías de procesos que están en la definición de las LA más frecuentes son

medición, *recolección* y *análisis* con un 100%, mientras que *informes* tiene solamente un 20,8%, como se ve en la Figura 5.7.

Siendo ASI el conjunto de los artículos científicos que utilizan Análisis Estadístico Implicativo y LA, el conjunto formado por los elementos del conjunto anterior (unido con los artículos científicos que utilizan *Learning Analytics*) y que siguen la definición de las LA respecto a la categoría procesos y a la subcategoría medición. Utilizando la teoría de conjuntos habría una relación de intersección no vacía respecto a la definición de las LA respecto a la categoría procesos y a la subcategoría medición que se representa por la fórmula $\underbrace{AEI \cap LA \neq \emptyset}_{Definición,procesos\ (80{,}2\%)}$:

$$\underbrace{AEI \cap AA \neq \emptyset}_{Definición,procesos,medición\ (100\%)}$$

Utilizando la frecuencia porcentual, se puede decir que hay un 100% de casos que validan la fórmula anterior.

Siendo ASI el conjunto de los artículos científicos que utilizan Análisis Estadístico Implicativo y LA, el conjunto formado por los elementos del conjunto anterior (unido con los artículos científicos que utilizan *Learning Analytics*) y que siguen la definición de las LA respecto a la categoría procesos y a la subcategoría recolección. Utilizando la teoría de conjuntos habría una relación de intersección no vacía respecto a la definición de las LA respecto a la categoría procesos y a la subcategoría recolección que se representa por la fórmula $\underbrace{AEI \cap LA \neq \emptyset}_{Definición,procesos\ (80{,}2\%)}$:

$$\underbrace{AEI \cap LA \neq \emptyset}_{Definición,procesos,recolección\ (100\%)}$$

Utilizando la frecuencia porcentual, se puede decir que hay un 100% de casos que validan la fórmula anterior.

Siendo ASI el conjunto de los artículos científicos que utilizan Análisis Estadístico Implicativo y LA, el conjunto formado por los elementos del conjunto anterior (unido con los artículos científicos que utilizan *Learning Analytics*) y que siguen la definición de las LA respecto a la categoría procesos y a la subcategoría análisis. Utilizando la teoría de

conjuntos habría una relación de intersección no vacía respecto a la definición de las LA respecto a la categoría procesos y a la subcategoría análisis que se representa por la fórmula $\underbrace{AEI \cap LA \neq \emptyset}_{Definición,procesos\ (80{,}2\%)}$: $\underbrace{AEI \cap AA \neq \emptyset}_{Definición,procesos,análisis\ (100\%)}$

Utilizando la frecuencia porcentual, se puede decir que hay un 100% de casos que validan la fórmula anterior.

Siendo ASI el conjunto de los artículos científicos que utilizan Análisis Estadístico Implicativo y LA, el conjunto formado por los elementos del conjunto anterior (unido con los artículos científicos que utilizan *Learning Analytics*) y que siguen la definición de las LA respecto a la categoría procesos. Utilizando la teoría de conjuntos habría una relación de intersección no vacía respecto a la definición de las LA respecto a la categoría procesos que se representa por la fórmula $\underbrace{AEI \cap LA \neq \emptyset}_{Definición,procesos\ (80{,}2\%)}$: $\underbrace{AEI \cap LA \neq \emptyset}_{Definición,procesos\ (80{,}2\%)}$

Utilizando la frecuencia porcentual, se puede decir que hay un 80,2% de casos que validan la fórmula anterior. También gráficamente mediante los diagramas de Euler Venn se expresa la relación anterior y ayuda a comprender de mejor manera como se muestra en la Figura 5.8.

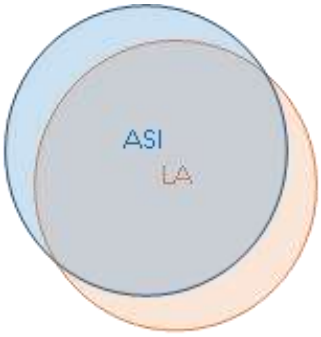


Figura 5.8.- Diagrama de Euler desde el punto de vista de los procesos de LA (Pazmiño-Maji et al., 2016)

5.2.5.3 Categoría salida

La Figura 5.9, muestra que los datos de salida en casi todos los artículos leídos tratan de comprender en el ámbito educativo (83,3%), muy poco se pretende optimizar (16,7%). El caso anterior no tiene que ver con las técnicas de análisis de datos utilizadas en el ASI, más bien responde a las necesidades educativas del aula. Las LA están motivando a que se optimice y que el paso de comprender sea un paso necesario, pero no suficiente.

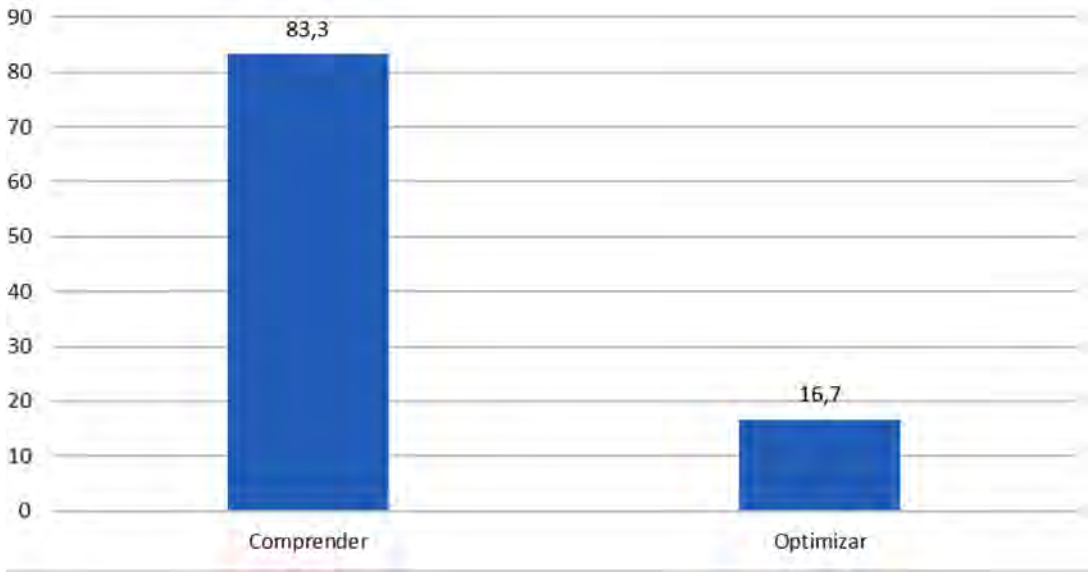


Figura 5.9.- Diagrama de Barras sobre la categoría salida de datos en la definición de las LA (Pazmiño-Maji et al., 2016)

Siendo ASI el conjunto de los artículos científicos que utilizan Análisis Estadístico Implicativo y LA, el conjunto formado por los elementos del conjunto anterior (unido con los artículos científicos que utilizan *Learning Analytics*) y que siguen la definición de las LA respecto a la categoría salidas. Utilizando la teoría de conjuntos habría una relación de intersección no vacía respecto a la definición de las LA respecto a la categoría salidas que se representa por la fórmula:

$$\underbrace{AEI \cap AA \neq \emptyset}_{Definición,salidas\ (50\%)}$$

Utilizando la frecuencia porcentual, se puede decir que hay un 50% de casos que validan la fórmula anterior. También gráficamente mediante los diagramas de Euler Venn se expresa la relación anterior y ayuda a comprender de mejor manera como se muestra en la Figura 5.10.

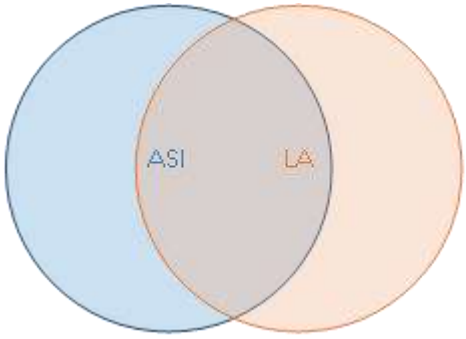


Figura 5.10.- Diagramas de Euler desde el punto de vista de los datos de salida (Pazmiño-Maji et al., 2016)

5.2.6 Resultados de la pregunta de investigación 2 (PI2)

Los subtemas más frecuentes en la aproximación del ASI a LA fuente de datos son aprendizaje (83,3%) y en el proceso escolar (12,5%), menos frecuentes son la demografía (4,2%) y percepciones (0%) (Tabla 5.7).

Tabla 5.7.- Artículos del ASI en la fuente de datos de LA (Pazmiño-Maji et al., 2016). Ver referencias completas en Apéndice

FUENTE DE DATOS	ARTÍCULOS CIENTÍFICOS	NÚMERO	FREQ (%)
Aprendizaje de los estudiantes	[1], [2], [3], [9], [10], [14], [16], [17], [18], [19], [57], [28], [31], [32], [36], [38], [42], [44], [47], [54]	20	83,3%
Demografía	[40]	1	4,2%
Percepción		0	0
Procesos de la institución educativa	[35], [42], [45]	3	12,5%

El diagrama circular de la Figura 5.11, muestra las proporciones en la categoría fuente de datos e indica que casi el total de datos se deben al aprendizaje de los estudiantes, dejando un 17% para demografía, percepción y procesos de la institución educativa.

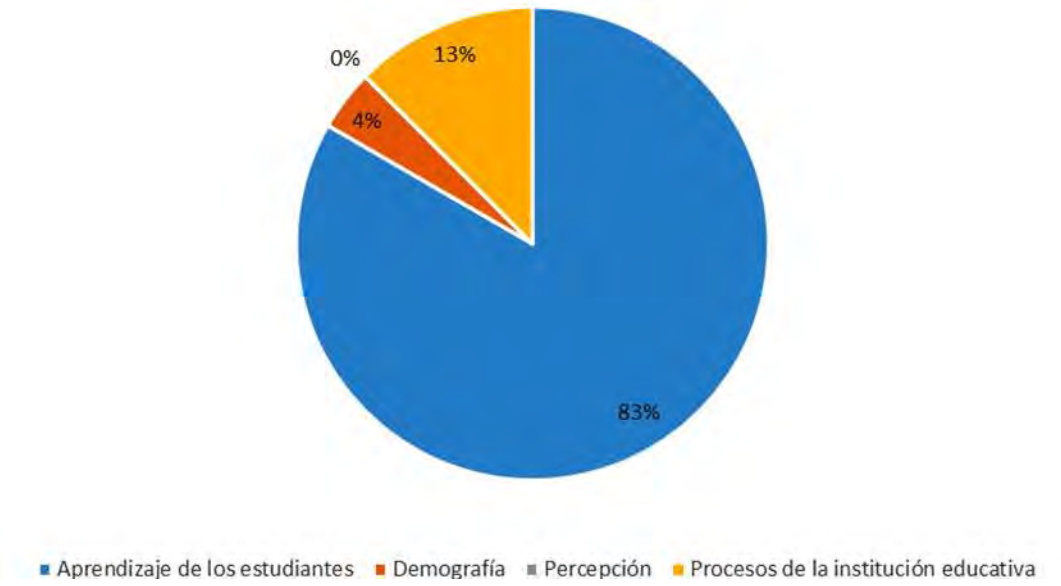


Figura 5.11.- Diagrama circular sobre la categoría fuente de datos (Pazmiño-Maji et al., 2016)

Siendo ASI el conjunto de los artículos científicos que utilizan Análisis Estadístico Implicativo y LA el conjunto formado por los elementos del conjunto anterior (unido con los artículos científicos que utilizan *Learning Analytics*) y que están en la categoría fuente de datos de las LA y subcategoría aprendizaje de los estudiantes. Utilizando la teoría de conjuntos habría una relación de intersección no vacía respecto a la definición de las LA respecto a la categoría fuente de datos y subcategoría aprendizaje de los estudiantes que se representa por la fórmula:

$$\underbrace{AEI \cap LA \neq \emptyset}_{Definición,procesos\ (80{,}2\%)} :$$

$$\underbrace{AEI \cap LA \neq \emptyset}_{Fuente\ de\ Datos,aprendizaje\ de\ los\ estudiantes\ (83.2\%)}$$

Utilizando la frecuencia porcentual, se puede decir que hay un 83,2% de casos que validan la fórmula anterior. También gráficamente mediante los diagramas de Euler Venn

se expresa la relación anterior y ayuda a comprender de mejor manera como se muestra en la Figura 5.12.

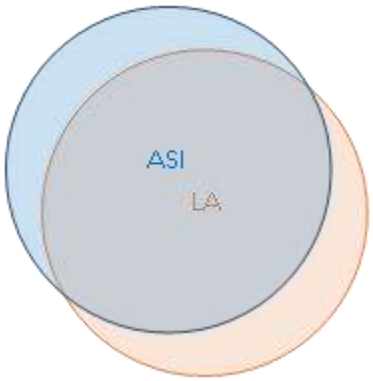


Figura 5.12.- Diagramas de Euler desde el punto de vista de la categoría fuente de datos (Pazmiño-Maji et al., 2016)

La respuesta a las preguntas de investigación uno (**PI1:** ¿Qué partes de la definición de *Learning Analytics* se observan en los artículos sobre Análisis Estadístico Implicativo?) y dos (**PI2:** ¿Cuáles son las fuentes de datos de *Learning Analytics* observadas en los artículos sobre Análisis Estadístico Implicativo?) se ven en forma conjunta y cuantificada de la siguiente manera para el caso de artículos científicos que utilizan el Análisis Estadístico Implicativo:

LA son la medición (100%), colección (100%), análisis (100%) y presentación de informes (20,8%) de datos (aprendizaje de los estudiantes (83,3%) y el proceso en la institución educativa (12,5%)) sobre aprendizaje (91,7%) y sus contextos (8,3%), a efectos de comprender (83,3%) y optimizar el aprendizaje (16,7%) y los ambientes en que se produce.

5.2.7 Resultados de la pregunta de investigación 3 (PI3)

La Tabla 5.8, muestra la referencia a artículos sobre Análisis Estadístico Implicativo, su número, frecuencia y las etapas de LA a las cuales pertenecen.

Tabla 5.8.- Artículos de ASI en las diferentes etapas de LA (Pazmiño-Maji et al., 2016) . Ver referencias completas en Apéndice

ETAPAS DE LAS LA	ARTÍCULOS EN LAS DIFERENTES ETAPAS DE LAS LA	NÚMERO	FREQ (%)
CAPTURAR	[1], [2], [3], [9], [10], [14], [16], [17], [18], [19], [57], [28], [31], [32], [35], [36], [38], [40], [42], [44], [45], [47], [54], [56]	24	100%
INFORMAR	[[1], [2], [3], [9], [10], [14], [16], [17], [18], [19], [57], [28], [31], [32], [35], [36], [38], [40], [42], [44], [45], [47], [54], [56]	24	100%
PREDECIR	[14], [18], [31], [40], [42], [47]	6	25%
ACTUAR	[40]	1	4,2%
REFINAR		0	0
	Promedio artículos en las diferentes etapas de las LA		**45,8%**

Se muestra además que las etapas más frecuentes son captura y reporte con un 100%, luego le sigue predecir con un 25% y actuar con un 4,2%, no tiene ningún caso registrado la etapa refinar.

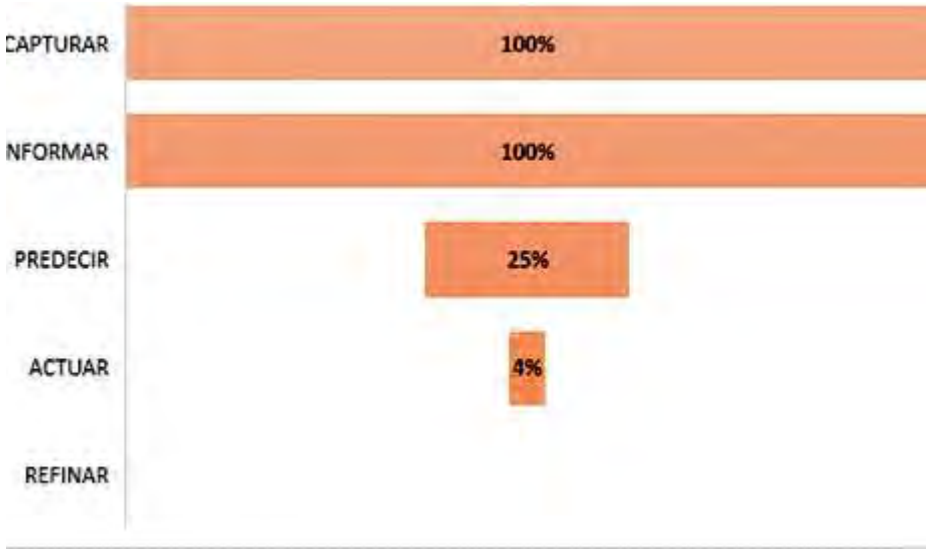


Figura 5.13.- Artículos de ASI en las diferentes etapas de LA (Pazmiño-Maji et al., 2016)

Siendo ASI el conjunto de los artículos científicos que utilizan Análisis Estadístico Implicativo y LA el conjunto formado por los elementos del conjunto anterior (unido con los artículos científicos que utilizan *Learning Analytics*) y que siguen las etapas de las LA

(Campbell & Oblinger, 2007). Utilizando la teoría de conjuntos habría una relación de intersección no vacía sobre las 5 etapas de las LA y respecto a la categoría capturar que se representa por la fórmula:

$$\underbrace{AEI \cap AA \neq \emptyset}_{Etapas,Capturar\ (100\%)}$$

Utilizando la frecuencia porcentual hay un 100% de casos que validan la fórmula anterior. También gráficamente mediante los diagramas de Euler Venn se expresa la relación anterior y ayuda a comprender de mejor manera como se muestra en la Figura 5.14.- Diagrama de Euler desde el punto de vista de las cinco etapas de LA y categoría capturar .

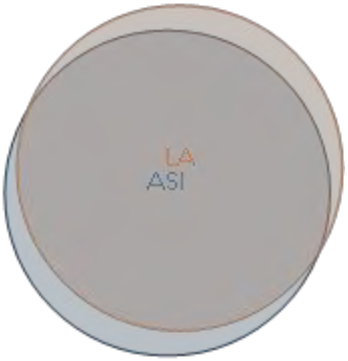


Figura 5.14.- Diagrama de Euler desde el punto de vista de las cinco etapas de LA y categoría capturar (Pazmiño-Maji et al., 2016)

Siendo ASI el conjunto de los artículos científicos que utilizan Análisis Estadístico Implicativo y LA el conjunto formado por los elementos del conjunto anterior (unido con los artículos científicos que utilizan *Learning Analytics*) y que siguen las etapas de las LA (Campbell & Oblinger, 2007). Utilizando la teoría de conjuntos habría una relación de intersección no vacía respecto a las 5 etapas de las LA respecto a la categoría Informar que se representa por la fórmula:

$$\underbrace{AEI \cap LA \neq \emptyset}_{Etapas,Informar\ (100\%)}$$

Utilizando la frecuencia porcentual, se puede decir que hay un 100% de casos que validan la fórmula anterior. También gráficamente mediante los diagramas de Euler Venn se

expresa la relación anterior y ayuda a comprender de mejor manera como se muestra en la Figura 5.15.

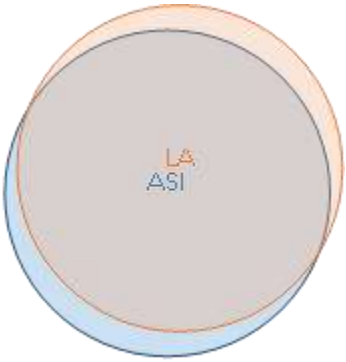


Figura 5.15.- Diagramas de Euler LA desde el punto de vista de las cinco etapas de LA y categoría informar (Pazmiño-Maji et al., 2016)

La respuesta global a la pregunta de investigación 3, ***PI3:*** *¿Cuáles de las cinco etapas de análisis en Learning Analytics se observan en los artículos sobre Análisis Estadístico Implicativo? es: captura e informe con el 100%, son menos frecuentes las etapas actuar 4,2% y refinar 0%.*

5.3 Aportes del ASI a LA: Desde 2016 hasta junio 2021

A continuación, se muestran aportes realizados mediante una nueva revisión sistemática de literatura[2]. La revisión sistemática 2021, es similar a la realizada en el año 2016 con el objetivo de actualizarla y realizar una comparación adecuada, guardando diferencias en lo que se refiere al tiempo y los resultados, sin considerar otros factores que afecten la comparación. En este capítulo se actualizan los resultados a las preguntas PI1: ¿Qué artículos científicos del ASI están en la definición de las LA?, PI2: ¿Cuáles datos del ASI están en las fuentes de datos de las LA?, y PI3: ¿Cuáles etapas del ASI están en las cinco etapas de análisis de las LA?

5.3.1 Preguntas de investigación (PI)

Se tiene como objetivo determinar y resumir información descriptiva sobre la aproximación de ASI a LA, se abordan tres preguntas de investigación (Li, Lam, & Lam, 2015):

[2] https://drive.google.com/drive/folders/1lqp52Xcx2TnJAQlULeS2YC2eTrbH7lS6?usp=sharing

PI1: ¿Qué partes de la definición de *Learning Analytics* se observan en los artículos sobre Análisis Estadístico Implicativo?

PI2: ¿Cuáles son las fuentes de datos de *Learning Analytics* observadas en los artículos sobre Análisis Estadístico Implicativo?

PI3: ¿Cuáles de las cinco etapas de análisis en *Learning Analytics* se observan en los artículos sobre Análisis Estadístico Implicativo?

5.3.2 Preguntas cortas de investigación

Se redacta las preguntas de investigación en forma corta para mejorar la comprensión de la asociación con el marco de aproximación.

PI1: ¿Qué artículos del ASI están en la definición de las LA?

PI2: ¿Cuáles datos del ASI están en las fuentes de datos de las LA?

PI3: ¿Cuáles etapas del ASI están en las cinco etapas de análisis de las LA?

En las respuestas de cada una de las preguntas de investigación se identifican los artículos científicos, se añaden los porcentajes de pertenencia individuales, parciales y totales.

5.3.3 Bases de datos bibliográficas utilizadas

Se examinaron los artículos científicos sobre ASI, cuya fuente fueron siete bases de datos bibliográficas: Biblioteca Digital ACM, EBSCO, Google Scholar, librería digital IEEE, ProQuest, base de datos de resúmenes y citación Scopus de Elsevier y la base de datos internacional Web of Science (WOS). La revisión utilizó estudios publicados en los últimos 5.5 años (66 meses), desde el 2016 hasta el primer semestre del 2021. Las fuentes de datos y las características de búsqueda se resumen en la Tabla 5.9.

Tabla 5.9.- Características de la base de datos y búsqueda bibliográficas (Pazmiño-Maji et al., 2016)

BASE DE DATOS BIBLIOGRÁFICAS		**ACM, EBSCO, Google Scholar, IEEE, ProQuest, Scopus, WOS**
BÚSQUEDA	Principales criterios de búsqueda	"*Statistical implicative analysis*" SIA
	Resultados (aplicando los criterios de exclusión y el tiempo):	37 documentos científicos sobre ASI. 21 en el área educativa y 16 en otras áreas.
	Tiempo	66 meses desde enero del 2016 hasta junio del 2021
	Tópicos: Criterios de aproximación	La definición Las fuentes de datos Las etapas

Durante la búsqueda, el subtema (o sinónimo) se ha guiado por la Tabla 5.10.

Tabla 5.10.- Criterios de examinación utilizados (Pazmiño-Maji et al., 2016)

PREGUNTA DE INVESTIGACIÓN	CRITERIOS DE APROXIMACIÓN	SUBCRITERIOS
1	Aproximación de ASI a las LA, definición y fuente de datos	Entrada LA: * Aprendizaje * Contexto del aprendizaje * Fuente de datos Proceso LA: *medición *recopilación *análisis *comunicación Salidas LA: * comprensión * optimización
2	Datos de ingreso	* Aprendizaje del estudiante * Demografía * Percepciones * Procesos escolares
3	Aproximación de ASI a las etapas de las LA	* Captura * Informar * Predecir * Actuar * Refinar

5.3.4 Criterios de inclusión, exclusión y calidad

Los criterios de inclusión y exclusión son los mismos de la Sección 5.2.4.5 y los criterios de calidad son los mismos de la Sección 5.2.4.6.

5.3.5 Cadenas lógicas de búsqueda

El grupo de estudios primarios se definió basándonos en (Zhang & Ali Babar, 2010). La cadena de búsqueda final se describió como sigue: ("*Statistical implicative analysis*" OR SIA) AND **(**LIMIT-TO (PUBYEAR, 2021) OR LIMIT-TO (PUBYEAR, 2020) OR LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018) OR LIMIT-TO (PUBYEAR, 2017) OR LIMIT-TO (PUBYEAR, 2016)**)** como se indica en (Kutvonen, 2008; Tolk, Turnitsa, & Diallo, 2006), las publicaciones del primer semestre del 2021 y en el área de educación fueron filtradas después utilizando el software de gestión de referencias Citavi versión 6.10.

5.3.6 Proceso de selección de artículos

La Figura 5.16 muestra las cadenas de búsqueda, se encontraron en total 157 artículos científicos (100%) que luego de eliminar los documentos repetidos y aplicar los criterios de

exclusión e inclusión se analizaron en profundidad 48 artículos (30,5%), se determinó si alguno de ellos no trataba de SIA y aplicando los criterios de calidad se obtuvo 37 (23,5%), finalmente se puso énfasis en los documentos de tipo educativo que fueron 21 (13,3%)

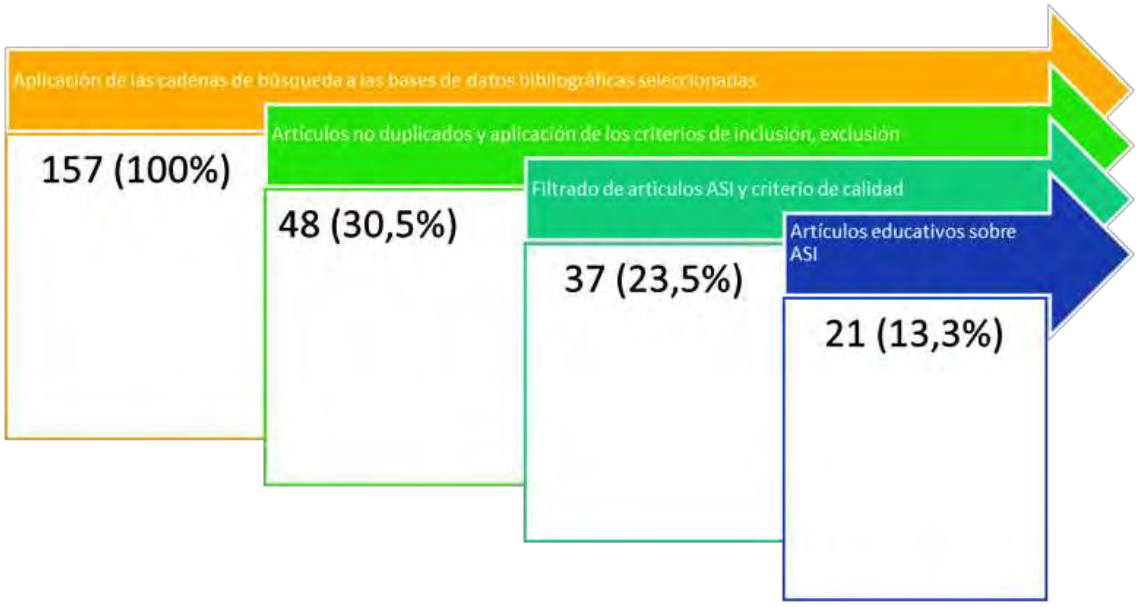


Figura 5.16.- Proceso de revisión sistemática desde 2016 hasta junio 2021 (Pazmiño-Maji et al., 2016)

En la primera búsqueda en las bases de datos bibliográficas se utilizaron las palabras claves: "*Statistical implicative analysis*", después utilizamos Citavi 6.1 (*Citavi - Reference Management and Knowledge Organization*, 2016).

Para aplicar los criterios de inclusión, exclusión y calidad, los resultados por base de datos bibliográfica se encuentran en detalle a continuación en la Tabla 5.11.

Tabla 5.11.- Resultados de los criterios de búsqueda e inclusión (Pazmiño-Maji et al., 2016)

BASES DE DATOS BIBLIOGRÁFICAS	BÚSQUEDA PRINCIPAL: "*Statistical implicative analysis*"	ARTÍCULOS NO DUPLICADOS Y APLICACIÓN DE LOS CRITERIOS DE INCLUSIÓN, EXCLUSIÓN	FILTRADO DE ARTÍCULOS ASI Y CRITERIO DE CALIDAD	ARTÍCULOS EDUCATIVOS SOBRE ASI, ENTRE EL 2016 Y EL SEMESTRE 1 DEL 2021
ACM	8	3	3	3
EBSCO	0	0	0	0
Google Scholar	70	29	21	9
IEEE	5	3	2	2
ProQuest	28	0	0	0
Scopus	33	10	8	5
WOS	13	3	3	2
Total	157	48	37	21

El gráfico de línea de tiempo sobre los artículos seleccionados de ASI en los 66 meses desde el 2016 hasta el primer semestre del 2021 se muestran en la Figura 5.17.

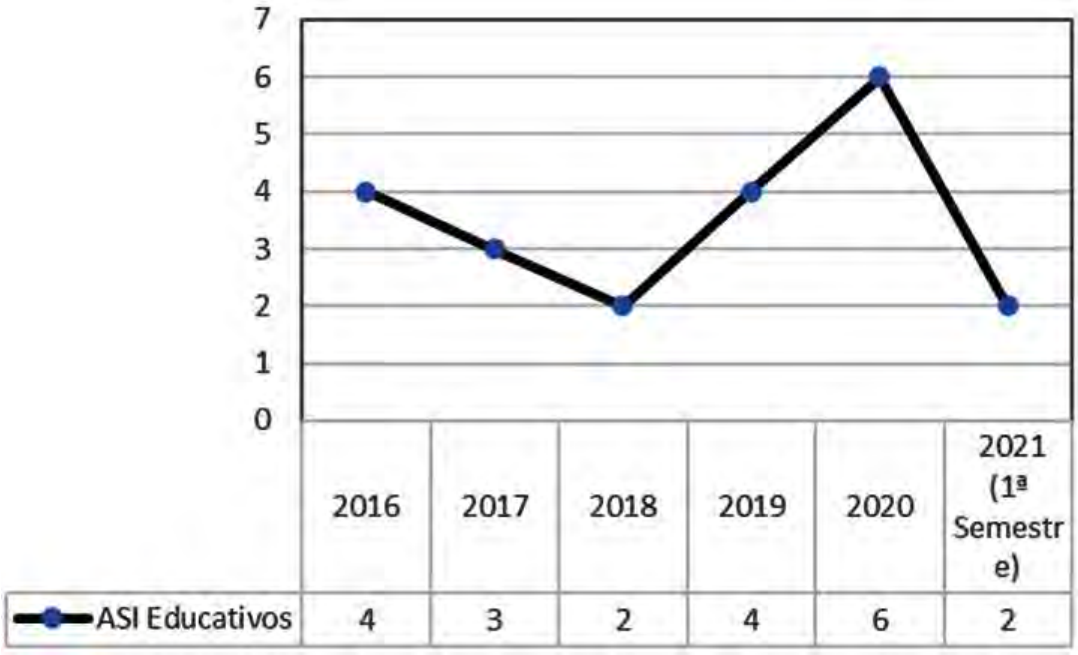


Figura 5.17.- Artículos educativos sobre ASI por año (Pazmiño-Maji et al., 2016)

5.3.7 Actualización de resultados pregunta 1 (PI1)

La Tabla 5.12, muestra las referencias a los artículos científicos sobre ASI, cantidad y frecuencia y que están enmarcados en la definición de *Learning Analytics*.

Tabla 5.12.- Artículos ASI que están enmarcados en la definición de las (Pazmiño-Maji et al., 2016). Ver referencias completas en Apéndice

SUBCATEGORÍAS DE DEFINICIÓN LA	TÍTULO DE LOS ARTÍCULOS EN CATEGORÍA DE DEFINICIÓN LA	NÚMERO	FREQ (%)
Categoría ingresos:			
Estudiantes	[1], [2], [3], [4], [6], [9], [10], [11], [12], [13], [14], [15], [16], [17], [19]	15	71,4%
Contexto de aprendizaje del estudiante	[1], [2], [3], [6], [7], [8], [9], [10], [12], [13], [15], [16], [20], [21]	14	66,6%
Promedio categoría ingresos:			69%
Categoría procesos:			
Medición	[4], [6], [7], [8], [9], [10], [11], [13], [14], [15], [17], [19], [20], [21]	14	66,7%
Recolección	[1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [13], [18], [19], [20], [21]	16	76,2%
Análisis	[1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [13], [15], [16], [17], [18], [19], [20], [21]	19	90,5%
Informes	[1], [2], [3], [4], [5], [6], [16], [19], [20], [21]	10	47,6%
Promedio categoría procesos:			70,2%
Categorías salidas:			
Comprender	[1], [2], [3], [4], [5], [7], [8], [11], [13], [15], [17], [19], [20]	13	61,9%
Optimizar	[3], [6], [7], [9]	4	19,0%
Promedio categorías salidas:			40,48%
Promedio definición de Analíticas de Aprendizaje (LA)			59,9%

5.3.7.1 Categoría ingresos

La Figura 5.18, muestra que los datos de ingreso en la mayoría de los artículos tratan sobre valoraciones realizadas directamente a los estudiantes (71,4%), más de la mitad se analiza sobre el contexto de aprendizaje del estudiante (66,6%). Las LA están motivando a que cada vez más se analice también el entorno del estudiante y en él se podrían aplicar las técnicas del ASI.

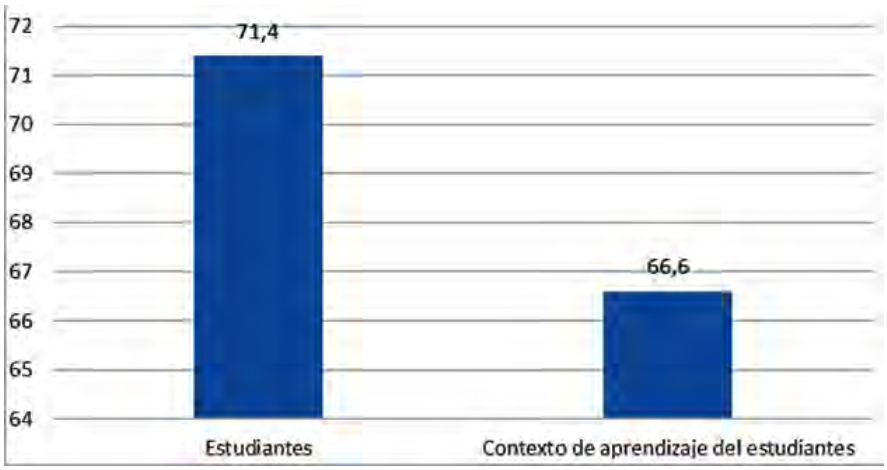


Figura 5.18.- Diagrama de Barras sobre la categoría ingreso de datos en la definición de las LA (Pazmiño-Maji et al., 2016)

5.3.7.2 Categoría procesos

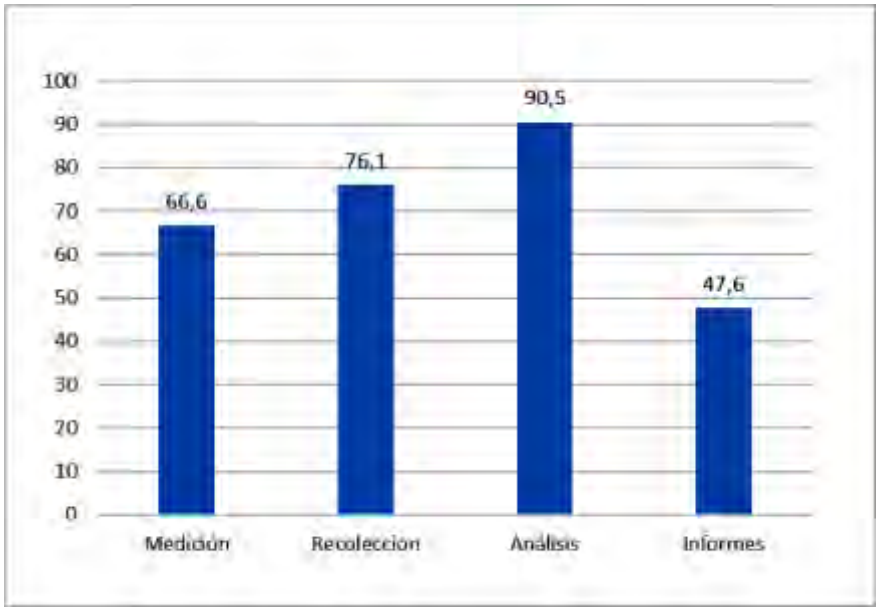


Figura 5.19.- Diagrama de Barras sobre la categoría procesos en la definición de las LA (Pazmiño-Maji et al., 2016)

Las subcategorías de proceso que hacen parte de la definición de las LA más frecuentes son análisis con un 90,5%, seguida por recolección con 76,1% y medición con un 66,6%, mientras que informes tiene solamente un 47,6%, como se puede ver en la Figura 5.19.

5.3.7.3 Categoría salida

La Figura 5.20, muestra que los datos de salida en casi todos los artículos leídos tratan de comprender en el ámbito educativo (61,9%), muy poco se pretende optimizar (19,0%). El caso anterior no tiene que ver con las técnicas de análisis de datos utilizadas en el ASI, más bien responde a las necesidades educativas del aula. Las LA están motivando a que se optimice y que el paso de comprender sea un paso necesario, pero no suficiente.

La respuesta a la pregunta de investigación uno (**PI1:** ¿Qué partes de la definición de *Learning Analytics* se observan en los artículos sobre Análisis Estadístico Implicativo?) se puede ver en forma cuantificada de la siguiente manera para el caso de artículos científicos que utilizan el Análisis Estadístico Implicativo:

Learning Analytics (LA) son la medición (66,7%), recolección (76,2%), análisis (90,5%) y presentación de informes (47,6%) de datos sobre aprendizaje (71,4%) y sus contextos (66,6%), a efectos de comprender (61,9%) y optimizar el aprendizaje (19,0%) y los ambientes en que se produce.

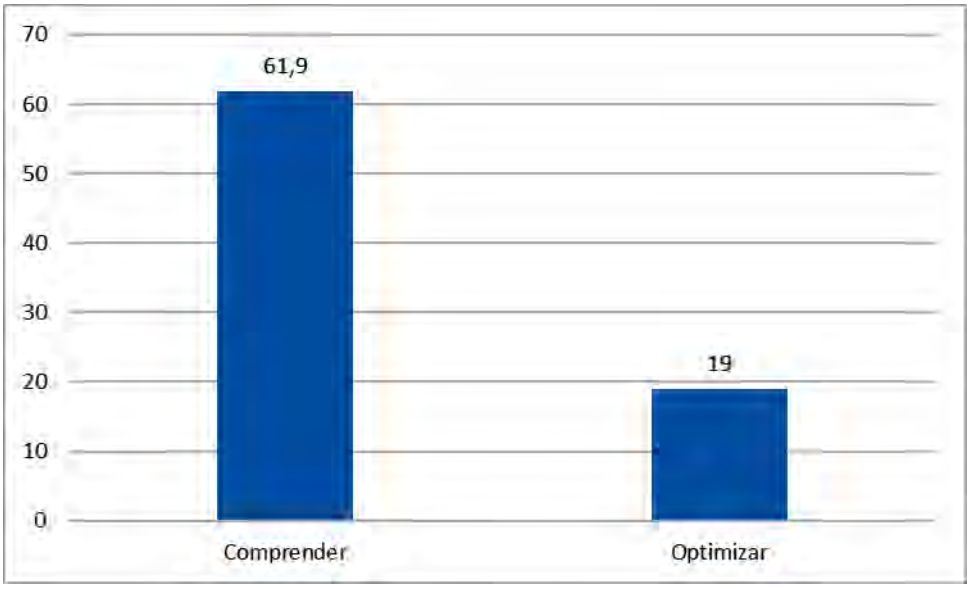


Figura 5.20.- Diagrama de Barras sobre la categoría salida de datos en la definición de las LA (Pazmiño-Maji et al., 2016)

5.3.8 Actualización de resultados pregunta 2 (PI2)

Los subtemas más frecuentes en la aproximación del ASI a LA fuente de datos son la percepción (71,4%), el aprendizaje (66,6%) y en el proceso escolar (23,8%), menos frecuentes son la demografía (19,0%) (Tabla 5.13).

Tabla 5.13.- Artículos del ASI en la fuente de datos de LA (Pazmiño-Maji et al., 2016). Ver referencias completas en Apéndice

FUENTE DE DATOS	ARTÍCULOS CIENTÍFICOS	NÚMERO	FREQ (%)
Aprendizaje de los estudiantes	[1], [3], [4], [6], [8], [9], [10], [13], [14], [16], [17], [19], [20], [21]	14	66,6%
Demografía	[3], [7], [10], [15]	4	19,0%
Percepción	[1], [2], [3], [5], [6], [7], [10], [11], [12], [14], [16], [17], [18], [20], [21]	15	71,4%
Procesos de la institución educativa	[3], [4], [9], [12], [13]	5	23,8%

La Figura 5.21, muestra las proporciones en la categoría fuente de datos e indica que casi el total de datos se deben a la percepción (71,4%), dejando un 19% para la demografía, es importante notar que la suma de los porcentajes de todas las tablas no será del 100%, debido a que un mismo artículo científico puede estar en varias categorías a la vez.

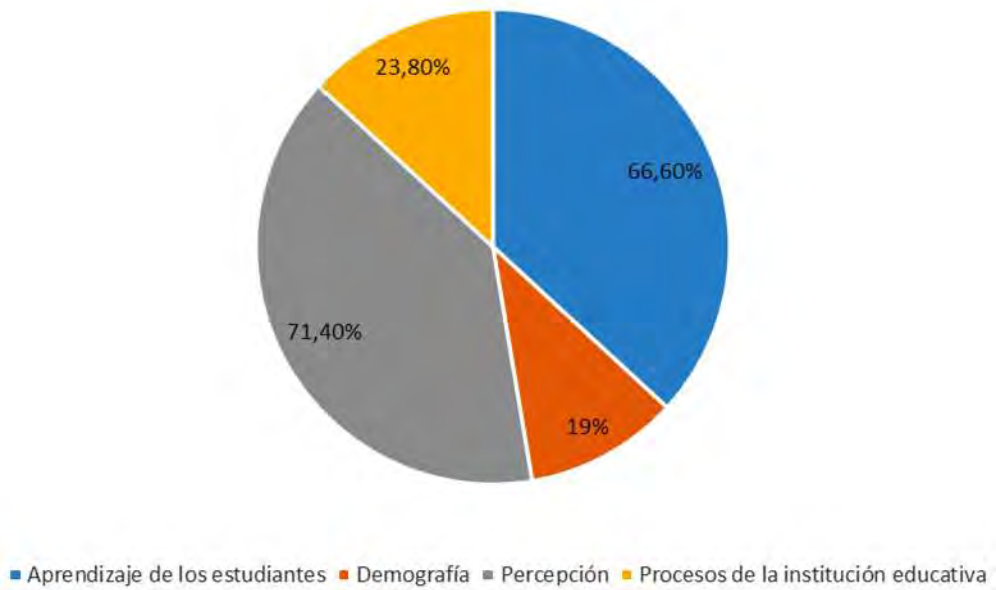


Figura 5.21.- Diagrama circular sobre la categoría fuente de datos (Pazmiño-Maji et al., 2016)

La respuesta a las preguntas de investigación uno (**PI1:** ¿Qué partes de la definición de *Learning Analytics* se observan en los artículos sobre Análisis Estadístico Implicativo?) y dos (**PI2:** ¿Cuáles son las fuentes de datos de *Learning Analytics* observadas en los artículos sobre Análisis Estadístico Implicativo?) se pueden ver en forma conjunta y cuantificada de la siguiente manera para el caso de artículos científicos que utilizan el Análisis Estadístico Implicativo:

Learning Analytics (LA) son la medición (66,7%), recolección (76,2%), análisis (90,5%) y presentación de informes (47,6%) de datos (aprendizaje de los estudiantes (66,6%), demografía (19,0%), percepción (71,4%) y el proceso en la institución educativa (23,8%)) sobre aprendizaje (71,4%) y sus contextos (66,6%), a efectos de comprender (61,9%) y optimizar el aprendizaje (19,0%) y los ambientes en que se produce.

5.3.9 Actualización de resultados pregunta 3 (PI3)

La Tabla 5.14, muestra la referencia a artículos sobre Análisis Estadístico Implicativo, su número, frecuencia y a las etapas de LA a las cuales pertenecen.

Tabla 5.14.- Artículos de ASI en las diferentes etapas de LA (Pazmiño-Maji et al., 2016) . Ver referencias completas en Apéndice

ETAPAS DE LAS LA	ARTÍCULOS EN LAS DIFERENTES ETAPAS DE LAS LA	NÚMERO	FREQ (%)
CAPTURAR	[1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [12], [13], [14], [15], [17], [18], [19], [20], [21]	19	90,4%
INFORMAR	[1], [2], [3], [4], [5], [6], [11], [13], [15], [19], [20], [21]	12	57,1%
PREDECIR	[3], [7]	2	9,5%
ACTUAR	[2], [3], [4], [5], [7], [13], [17]	7	33,3%
REFINAR	[9], [10], [11], [13]	4	19%
	Promedio artículos en las diferentes etapas de las LA		41,9%

Se muestra además que las etapas más frecuentes son captura 90,4% e informe 57,1%, la etapa predecir tiene la frecuencia más bajo con el 9,5%.

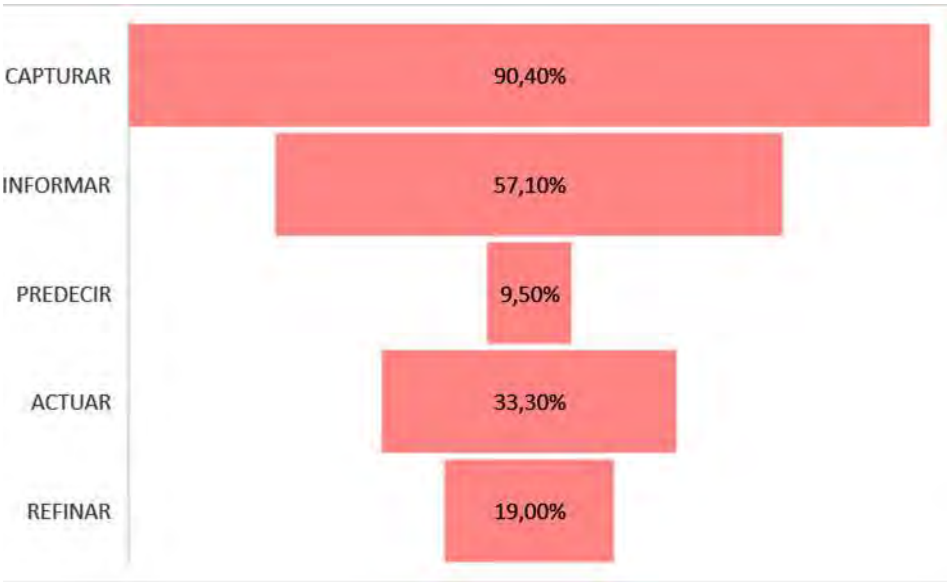


Figura 5.22.- Artículos de ASI en las diferentes etapas de LA (Pazmiño-Maji et al., 2016)

La respuesta global a la pregunta de investigación 3, **PI3:** ¿Cuáles de las cinco etapas de análisis en *Learning Analytics* se observan en los artículos sobre Análisis Estadístico Implicativo? es: captura 90,4%, informe 57,1% y actuar 33,3%, son menos frecuentes las etapas refinar 19% y predecir 9,5%.

5.4 Conclusiones

Las LA tiene algunas definiciones, se seleccionó aquella más común (Pazmiño-Maji Rubén et al., 2021) dada en la primera Conferencia Internacional sobre análisis de aprendizaje y conocimiento (*LAK 2011 : 1st International Conference Learning Analytics and Knowledge*, 2011).

A la primera pregunta PI1 sobre el aporte (aportación, contribución, tributo, cuota, participación, ayuda)[3] del ASI a la definición de LA, los resultados obtenidos en las revisiones sistemáticas realizadas del 2011 hasta jun2016 y 2016 hasta jun2021 se muestran a continuación en la Tabla 5.15.

Tabla 5.15.- Comparativo de PI1 de las revisiones sistemáticas realizadas desde el 2011 hasta el 2021 (Pazmiño-Maji et al., 2016)

SUBCATEGORÍAS DEFINICIÓN LA	2011-jun2016		2016-jun2021	
	NÚMERO_A[4]	FREQ_A	NÚMERO_D	FREQ_D
Categoría ingresos:				
Estudiantes	22	91,7%	15	71,4%
Contexto de aprendizaje del estudiante	2	8,3%	14	66,6%
Promedio ingresos		50%		69%
Categoría procesos:				
Medición	24	100%	14	66,7%
Recolección	24	100%	16	76,2%
Análisis	24	100%	19	90,5%
Informes	5	20,8%	10	47,6%
Promedio procesos		80,2%		70,2%
Categorías salidas:				
Comprender	20	83,3%	13	61,9%
Optimizar	4	16,7%	4	19,0%
Promedio salidas		50%		40,48%
Promedio Total		61,1%		59,9%

Se obtuvo una contribución global del ASI a la definición de las Analíticas de Aprendizaje del 61,1% en el 2016 y del 59,9% en el año 2021, que es un porcentaje similar manteniéndose un buen tributo del ASI a LA. Se nota que el ASI tiene como fuentes principales los aprendizajes de los estudiantes y se tiende notablemente a considerar cada vez más el contexto de éste (de 8,3% a 66,6%). El ASI ayudaba más a comprender (2011-jun2016), pero actualmente aumentan los artículos científicos que optimizan los aprendizajes (de 16,7% a 19,0%). Con los porcentajes calculados, el ASI actualmente contribuye significativamente al LA en el análisis y en los reportes del aprendizaje y su contexto por su facilidad de interpretación (árboles de similaridad, árboles de cohesión y gráfico implicativo). Es importante notar que actualmente los datos considerados no son

[3] (*Sinónimos de aporte*, 2020) (*aporte - sinónimos y antónimos - WordReference.com*, 2021)(ASALE y RAE, 2021)

[4] _A representa los resultados de la revisión sistemática de literatura entre los años 2011 hasta junio del 2016.

_D representa los resultados de la revisión sistemática de literatura entre los años 2016 hasta junio del 2021.

únicamente de los estudiantes, sino también de su contexto (69%), superando ampliamente con el 19% el estudio actual del 2021 al estudio anterior del 2011 en la categoría ingresos (medición, recolección, análisis e informes).

En la segunda pregunta PI2 sobre el aporte del ASI a las fuentes de datos de LA (using-data-improve-student-learning, 2016), como resultados se tiene que el ASI continúa contribuyendo en los datos de los aprendizajes (66,6%), pero ha aumentado su contribución en los datos complementarios al aprendizaje como son los datos de los procesos de la institución educativa con un incremento del 11,30%, la demografía (¿Cuántos estudiantes están matriculados en la escuela por año?, ¿Cómo ha cambiado la matrícula en la escuela durante los últimos cinco años?) con un incremento del 18,8% y la percepción (¿Cuál es el grado de satisfacción de los padres, los estudiantes y personal con el ambiente de aprendizaje?, ¿Cómo ha cambiado la percepción de los estudiantes del ambiente de aprendizaje en el tiempo?) con un incremento significativo de 71,40% (Tabla 5.16).

Tabla 5.16.- Comparativo de PI2 de las revisiones sistemáticas realizadas desde el 2011 hasta el 2021 (Pazmiño-Maji et al., 2016)

FUENTE DE DATOS	2011-jun2016		2016-jun2021	
	NÚMERO_A	FREQ_A	NÚMERO_D	FREQ_D
Aprendizaje de los estudiantes	20	83,3%	14	66,6%
Demografía	1	4,2%	4	19,0%
Percepción	0	0%	15	71,40%
Procesos de la institución educativa	3	12,5%	5	23,8%

En la tercera pregunta PI3, utilizamos las 5 etapas de Campbell en LA: captura, informe, predecir, actuar y refinar.

En la revisión sistemática del 2016 (Tabla 5.17) se observa que el ASI contribuye en la captura, información y predicción, pero contribuye poco en la etapa de actuación y hasta el año 2016 no se detectó la contribución en la etapa de refinamiento. En la revisión sistemática del 2021 ya se puede detectar que el ASI aporta al LA en el actuar, con un incremento desde el 4,2% hasta el 33,3% y en refinar de ningún aporte en los años 2011 hasta junio del 2016, se tiene en los años 2016 hasta junio del 2021 un incremento del 19%.

Tabla 5.17.- Comparativo de PI3 de las revisiones sistemáticas realizadas desde el 2011 hasta el 2021 (Pazmiño-Maji et al., 2016)

ETAPAS DE LAS LA	2011-jun2016		2016-jun2021	
	NÚMERO_A	FREQ_A	NÚMERO_D	FREQ_D
CAPTURAR	24	100%	19	90,4%
INFORMAR	24	100%	12	57,1%
PREDECIR	6	25%	2	9,5%
ACTUAR	1	4,2%	7	33,3%
REFINAR	0	0%	4	19%

Las conclusiones obtenidas en este capítulo muestran detalladamente los aportes del Análisis Estadístico Implicativo (ASI) a las Analíticas de Aprendizaje (LA) en la definición de LA dada en (*LAK 2011 : 1st International Conference Learning Analytics and Knowledge*, 2011), las fuentes de datos de LA (aprendizaje de los estudiantes, demografía, percepción y procesos de la institución educativa) y las etapas de LA definidas por Campbell (captura, información, predicción, actuación y refinamiento).

Capítulo 6to | APORTE CON TÉCNICAS ASI A LA

Mediante revisiones sistemáticas de literatura se determinan las técnicas de análisis con las cuales aporta el ASI a LA.

6 Capítulo.- Aporte con técnicas ASI a LA

A continuación, se muestran los resultados del artículo científico “Approximation of *Statistical implicative analysis* to *Learning Analytics*: a systematic review”, donde se muestran las técnicas de análisis similares entre el Análisis Estadístico Implicativo (ASI) y las Analíticas de Aprendizaje (LA).

6.1 Introducción

El objetivo principal de este capítulo es determinar las herramientas de análisis similares entre ASI y LA por la función cumplida, para ello se ha diseñado un marco de aproximación para *Learning Analytics* basado en las clasificaciones de herramientas de análisis de Baker e Inventado y Papamitsiou y Economides. Se utilizó un mapeo sistemático de la literatura publicada en los 66 meses desde enero del 2011 hasta junio del 2016, en las bases de datos bibliográficas ACM, EBSCO, Google Scholar, IEEE, ProQuest, Scopus y WOS. Se inició con 319 artículos y finalmente 24 fueron los que cumplieron con todos los criterios de calidad. Este documento contiene las herramientas de análisis mediante los cuales ASI puede contribuir por su similaridad a LA.

La aproximación a los métodos de análisis se basó en las clasificaciones de los métodos de análisis realizadas por Baker e Inventado (Baker & Inventado, 2014), que se basa con la similaridad a los métodos de la minería de datos educativos y una clasificación propuesta recientemente por Papamitsiou y Economides (Papamitsiou y Economides, 2014).

6.2 Aporte con técnicas ASI a LA: Desde 2011 hasta junio 2016

Se retoma el procedimiento y los resultados obtenidos en la revisión sistemática realizada por el autor de la tesis en el año 2016 (Pazmiño-Maji et al., 2016).

6.2.1 Clasificaciones de herramientas de análisis de Baker e Inventado y Papamitsiou y Economides

La clasificación de Baker e Inventado se muestra en la Tabla 6.1.

Tabla 6.1.- Clasificación de los métodos de análisis propuesto por Baker e Inventado (Baker y Inventado, 2014; Pazmiño-Maji et al., 2016)

MÉTODO	TÉCNICA
Predicción	Clasificación Regresión Estimación del Conocimiento Latente
Minería de relaciones	Minería de reglas de asociación Minería de patrones secuenciales Minería de correlaciones Minería de datos causales
Descubrimiento de estructuras	Clúster Análisis de factores Descubrimiento de estructuras de dominio
Descubrimiento con modelos	

Para detallar los métodos de análisis también se ha utilizado una clasificación reciente propuesta por una investigación de Papamitsiou y Economides (Papamitsiou & Economides, 2014) que publicaron los métodos de *Data Mining* y *Learning Analytics* (DM/LA) organizados de la siguiente manera:

- Clasificación.
- Agrupamiento.
- Regresión.
- Minería de texto.
- Minería de regla de asociación.
- Análisis de redes sociales.
- Descubrimiento con modelos.
- Visualización.
- Estadísticas.

La Tabla 6.2, muestra la forma de aproximación a los métodos de análisis basados en las clasificaciones para *Learning Analytics*, propuestos por Baker e Inventado y Papamitsiou y Economides (Baker y Inventado, 2014; Papamitsiou y Economides, 2014).

Tabla 6.2.- Métodos de análisis utilizados (Pazmiño-Maji et al., 2016)

ETAPAS (Campbell & Oblinger, 2007)	**MÉTODOS**	
	Baker e Inventado (Baker & Inventado, 2014)	Papamitsiou and Economides (Papamitsiou & Economides, 2014)
CAPTURAR		
INFORMAR PREDECIR	Predicción (Clasificación, Regresión, Estimación del Conocimiento Latente), Minería de relaciones (minería de patrones secuenciales de minería de reglas de asociación, minería de correlaciones, minería de datos causales), Descubrimiento de Estructura (agrupación, análisis de factores, descubrimiento de estructuras de dominios) Descubrimiento con modelos	Clasificación Agrupamiento Regresión Minería de texto Minería de regla de asociación Análisis de redes sociales, Descubrimiento con modelos Estadísticas Visualización de texto
ACTUAR REFINAR		

El método utilizado fue el mapeo sistemático de literatura, a continuación, se muestra el desarrollo parcial de la metodología.

6.2.2 Etapas del mapeo sistemático de literatura

La metodología utilizada fue la de mapeo sistemático de literatura adaptado de los autores (Okoli y Schabram, 2010). El mapeo se realizó en las publicaciones científicas sobre el Análisis Estadístico Implicativo (ASI) durante 66 meses, desde enero del 2011 hasta junio del 2016.

6.2.2.1 Preguntas de investigación

Esta sección, tiene como objetivo determinar los métodos de análisis de datos similares (por su aplicación) entre el ASI a LA, se elaboraron dos preguntas de investigación según lo indicado por (Li et al., 2015):

MI1: ¿Qué métodos similares a los propuestos por la clasificación de Baker e Inventado (Baker y Inventado, 2014) se observan en las investigaciones educativas con un enfoque del ASI?

MI2: ¿Qué métodos similares a los propuestos por la clasificación de Papamitsiou y Economides (Papamitsiou y Economides, 2014) se observan en las investigaciones educativas con un enfoque del ASI?

6.2.2.2 Preguntas de investigación y el método pico

La metodología utilizada fue una revisión sistemática de la literatura de investigación empírica (Okoli y Schabram, 2010) sobre Análisis Estadístico Implicativo. Este capítulo, tiene como objetivo reunir y resumir información descriptiva sobre la aproximación de ASI a las LA mediante los métodos de análisis similares (Li et al., 2015).

La primera pregunta de investigación **MI1:** ¿Qué métodos similares a los propuestos por la clasificación de Baker e Inventado se observan en las investigaciones educativas con un enfoque del ASI?, tendrá el desarrollo de la metodología PICO de la Tabla 6.3.

Tabla 6.3.- Metodología PICO aplicada a la primera pregunta de investigación (Pazmiño-Maji et al., 2016)

P	24 documentos científicos sobre ASI
I	Métodos de análisis de datos clasificados según la clasificación de Baker e Inventado classification and Papamitsiou and Economides methods classification
C	Sin comparación
O	Número de artículos de ASI que contiene métodos de análisis similares a los de la clasificación de Baker e Inventado

La segunda pregunta de investigación **MI2:** ¿Qué métodos similares a los propuestos por la clasificación de Papamitsiou y Economides se observan en las investigaciones educativas con un enfoque del ASI?, tendrá el desarrollo de la metodología PICO de la Tabla 6.4.

Tabla 6.4.- Metodología PICO aplicada a la segunda pregunta de investigación (Pazmiño-Maji et al., 2016)

P	24 documentos científicos sobre ASI
I	Métodos de análisis de datos clasificados según la clasificación de Papamitsiou y Economides (Papamitsiou y Economides, 2014)
C	Sin comparación
O	Número de artículos de ASI que contiene métodos de análisis similares a los de la clasificación de Papamitsiou y Economides (Papamitsiou y Economides, 2014)

6.2.2.3 Bases de datos bibliográficas utilizadas

Primeramente, se examinaron los artículos científicos sobre ASI de las principales bases de datos bibliográficas, sin tomar en cuenta otras fuentes como los congresos ASI (ver en http://sites.univ-lyon2.fr/asi9/?page=6&lang=en), las fuentes utilizadas fueron: Biblioteca Digital ACM, EBSCO, Google Scholar, librería digital IEEE, ProQuest, base de datos de resúmenes y citación Scopus de Elsevier y la base de datos internacional Web of Science (WOS). El mapeo sistemático se basó a estudios publicados en los 66 meses, entre el 2011 y el primer semestre del 2016. La búsqueda de los métodos de análisis fue guiada por la siguiente tabla.

Tabla 6.5.- Criterios de examinación utilizados (Pazmiño-Maji et al., 2016)

MI	APROXIMACIÓN	MÉTODO
1	Aproximación de ASI a los métodos de las LA: Clasificación de Baker e Inventado (Baker y Inventado, 2014)	Predicción (Clasificación, Regresión, Estimación del Conocimiento Latente). Minería de relaciones (minería de reglas de asociación, minería de patrones secuenciales, minería de correlaciones, minería de datos causales). Descubrimiento de estructuras (clúster, análisis de factores, descubrimiento de estructuras de dominio). Descubrimiento con modelos.
2	Aproximación de ASI a los métodos de las LA: Clasificación de Papamitsiou y Economides (Papamitsiou y Economides, 2014)	LA/EDM métodos: * Clasificación. * Agrupamiento. * Regresión. * Minería de textos. * Minería de regla de asociación. * Análisis de redes sociales. * Descubrimiento con modelos. * Visualización. * Estadísticas.

6.2.2.4 Cadenas lógicas de búsqueda

Los estudios primarios se seleccionaron basándonos en (Zhang & Ali Babar, 2010). La cadena de búsqueda final se describió como sigue: ("*Statistical implicative analysis*" OR SIA) AND **(**LIMIT-TO (PUBYEAR, 2016) OR LIMIT-TO (PUBYEAR, 2015) OR LIMIT-TO (PUBYEAR, 2014) OR LIMIT-TO (PUBYEAR, 2013) OR LIMIT-TO (PUBYEAR, 2012) OR LIMIT-TO (PUBYEAR, 2011)**)** como se indica en (Kutvonen, 2008; Tolk et al., 2006), las publicaciones del primer semestre del 2016 y en el área de educación fueron filtradas

después utilizando el software de gestión de referencias EndNote. Los métodos de análisis luego fueron determinados por la lectura exhaustiva de cada uno de los documentos, tomando en cuenta entre otros parámetros el criterio de aplicabilidad para determinar métodos similares.

6.2.2.5 Criterios de inclusión y exclusión

Al final de la etapa de recopilación de datos, se aplicó rigurosamente tanto los criterios de inclusión así como los criterios de exclusión.

6.2.2.6 Criterios de calidad

A continuación, se muestran los dos criterios de calidad que se utilizaron en la selección de la literatura:

- **Claridad de la metodología:** título, resumen, objetivo, datos, población de estudio, métodos de análisis, objetivo de los métodos de análisis, criterio de aplicabilidad de los métodos utilizados. software y resultados.
- **Idoneidad de resultados:** análisis y gráficos estadísticos, dendrogramas, gráficos causales, gráficos en estructura arborescente, resultados numéricos, tablas y discusión.

Las etapas seguidas luego de la aplicación de los criterios de calidad fueron:

1) Lectura y determinación de las técnicas de análisis de datos de los artículos.
2) Elaboración de una base de datos de artículos científicos y sus principales atributos.
3) Registro de los criterios de aplicabilidad de cada uno de los artículos analizados.
4) Utilización de estadísticas descriptivas para analizar, interpretar, registrar y sintetizar los resultados.

6.2.3 Proceso de selección de artículos sobre ASI

La Figura 6.1, muestra el proceso de mapeo sistemático para la selección de artículos. Los artículos que se leyeron y analizaron fueron el 7,5 % de los originales que fueron en un total de 319.

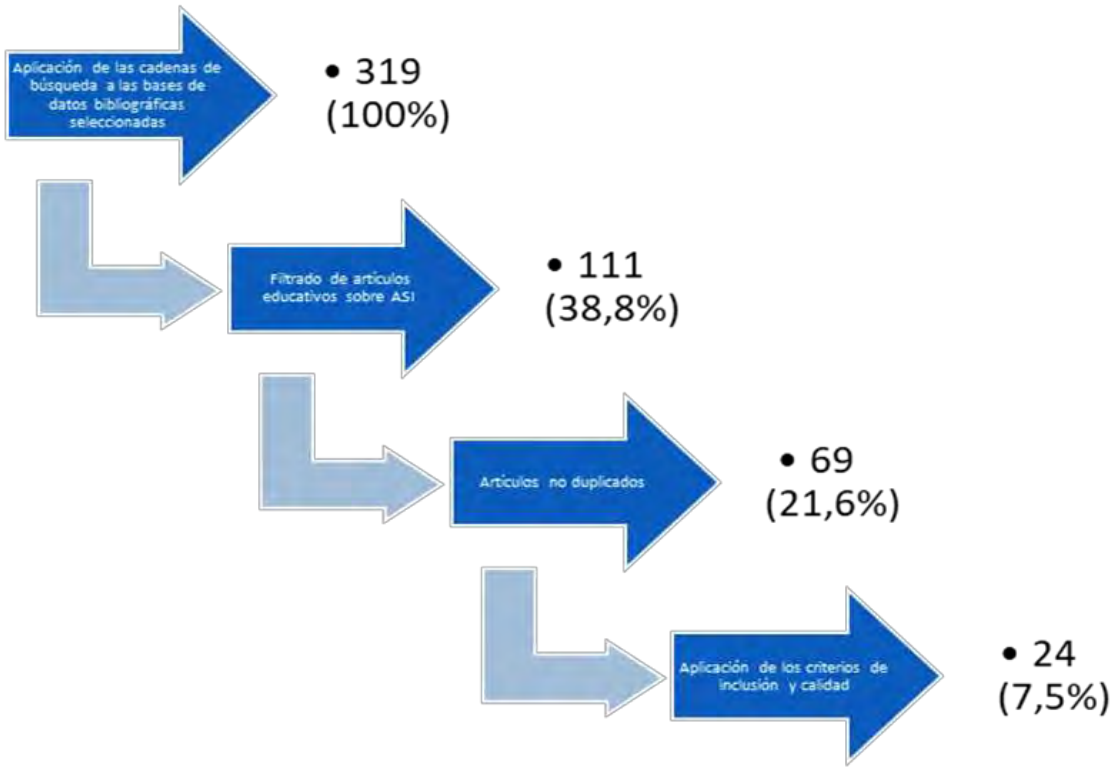


Figura 6.1.- Proceso de selección de artículos en el mapeo sistemático (Pazmiño-Maji et al., 2016)

Para la búsqueda inicial se utilizaron las palabras claves: "*Statistical implicative analysis*" o su contracción francesa “ASI”, luego se empleó EndNote como herramienta para la publicación y gestión bibliográficas y además para detectar registros del primer semestre del 2016, sobre educación y duplicados, los criterios sobre fechas, área educativa y subtemas se fijaron previamente.

6.2.4 Resultados pregunta (MI1)

La Tabla 6.6, muestra la referencia de artículos de ASI, su frecuencia absoluta y frecuencia porcentual que responde a la pregunta de investigación MI1.

Tabla 6.6.- Artículos ASI en los métodos de análisis según Baker e Inventado (Baker y Inventado, 2014; Pazmiño-Maji et al., 2016). Ver referencias completas en Apéndice

MÉTODOS DE ANÁLISIS EN LAS LA SEGÚN BAKER E INVENTADO (Baker y Inventado, 2014)		**ARTÍCULOS**	**NÚMERO**	**FREQ (%)**
Predicción	Clasificación		0	0
	Regresión		0	0
	Estimación del Conocimiento Latente		0	0
Minería de relaciones	Minería de reglas de asociación	[1], [2], [3], [9], [10], [14], [16], [17], [18], [19], [57], [28], [31], [32], [35], [36], [38], [40], [42], [44], [45], [47], [56]	23	95,8%
	Minería de patrones secuenciales		0	0
	Minería de correlaciones		0	0
	Minería de datos causales	[54]	1	4,2%
Descubrimiento de estructura	Clúster	[2], [9], [14], [17], [57], [28], [32], [40], [54]	9	37,5%
	Análisis de factores		0	0
	Descubrimiento de estructuras de dominio		0	0
Descubrimiento con modelos			0	0
Descubrimiento con modelos			0	0

Se observa que el 95,8% de los artículos utilizan métodos del ASI similares a la minería de reglas de asociación que hacen parte de la minería de relaciones, el 37,5% de los artículos utilizan métodos del ASI similares a clúster para el descubrimiento de estructura, solo un 4,2% de los artículos utilizan métodos del ASI similares a la minería de datos causales que hacen parte de la minería de relaciones.

La respuesta a la pregunta MI1 es que los artículos que utilizan métodos del ASI similares a los métodos en las LA según la clasificación de Baker e Inventado son minería de reglas de asociación (95,8%), clúster (37,5%), minería de datos causales (4,2%), los otros métodos muestran no ser utilizados (Figura 6.2).

Figura 6.2.- Artículos ASI en los métodos de análisis según Baker e Inventado (Baker y Inventado, 2014; Pazmiño-Maji et al., 2016)

6.2.5 Resultados pregunta (MI2)

La Tabla 6.7, muestra los artículos del ASI, su número y porcentaje. Es importante notar que muchos de los artículos analizados utilizaban más de una técnica de análisis del ASI y por ello pueden contabilizarse dentro de varios métodos de las analíticas del aprendizaje. Hay que observar también que cuando se seleccionó la técnica clúster, su indicador de utilidad (dados en los capítulos 2 y 3) era la búsqueda de agrupaciones y éstas se las encuentra mediante clúster no jerárquico y clúster jerárquico, pero en la selección no se hace diferencia entre los dos. Cuando se habla de estadística se refiere a ciertas medidas que se utilizan en los documentos ASI y son generados por los softwares CHIC y Rchic no con un software adicional. Las medidas descriptivas a las que se hace referencia son: análisis de frecuencia, medidas descriptivas básicas, medidas de similaridad y medidas de correlación. El análisis de frecuencia es absoluto (el conteo de las ocurrencias), porcentual y entre pares de variables, las medidas descriptivas son la media aritmética y la desviación estándar, las medidas de similaridad se concentran en los

índices de similaridad, las medidas de correlación se basan en el coeficiente de correlación de Pearson.

Tabla 6.7.- Artículos ASI en los métodos de análisis según Papamitsiou y Economides (Papamitsiou y Economides, 2014; Pazmiño-Maji et al., 2016). Ver referencias completas en Apéndice

MÉTODOS DE ANÁLISIS EN LAS LA	ARTÍCULOS EN LOS MÉTODOS DE ANÁLISIS DE LAS LA	NÚMERO	FREQ (%)
Clasificación		0	0
Clúster	[2], [9], [14], [17], [57], [28], [32], [40], [54]	9	37,5%
Regresión		0	0
Minería de texto		0	0
Minería de regla de asociación	[1], [2], [3], [9], [10], [14], [16], [17], [18], [19], [57], [28], [31], [32], [35], [36], [38], [40], [42], [44], [45], [47], [56]	23	95,8%
Análisis de redes sociales		0	0
Descubrimiento con modelos		0	0
Visualización		0	0
Estadísticas	[1], [17], [57], [31], [36]	5	20,8%

Se observa que el 95,8% de los artículos utilizan métodos del ASI similares a la minería de reglas de asociación, que el 37,5% de los artículos utilizan métodos del ASI similares a clúster, que un 20,8% de los artículos utilizan métodos del ASI similares a la estadística.

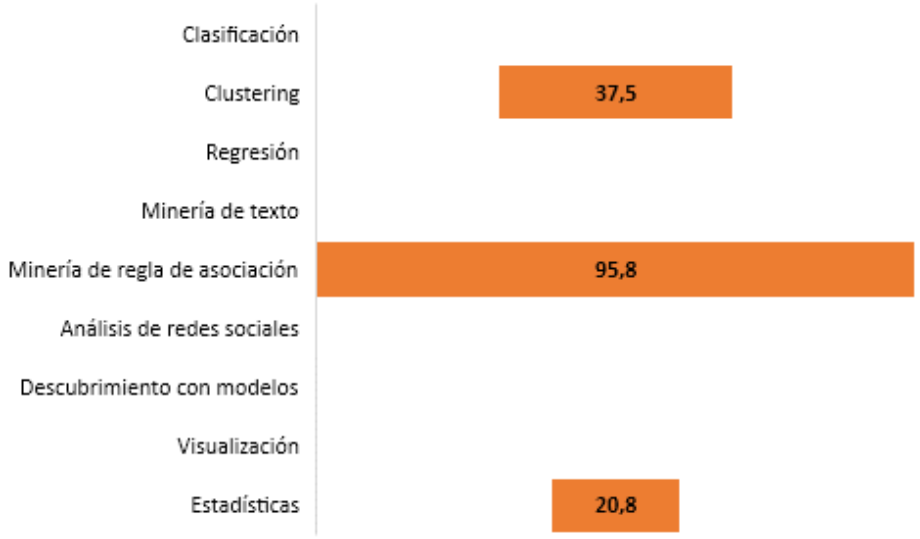


Figura 6.3.- Artículos ASI en los métodos de análisis según Papamitsiou y Economides (Papamitsiou y Economides, 2014; Pazmiño-Maji et al., 2016)

La respuesta a la pregunta MI2 (Figura 6.3) es que los artículos que utilizan métodos del ASI similares a los métodos en las LA según la clasificación de Papamitsiou y Economides son minería de reglas de asociación (95,8%), clúster (37,5%), estadísticas (20,8%) los otros métodos no muestran ser utilizados.

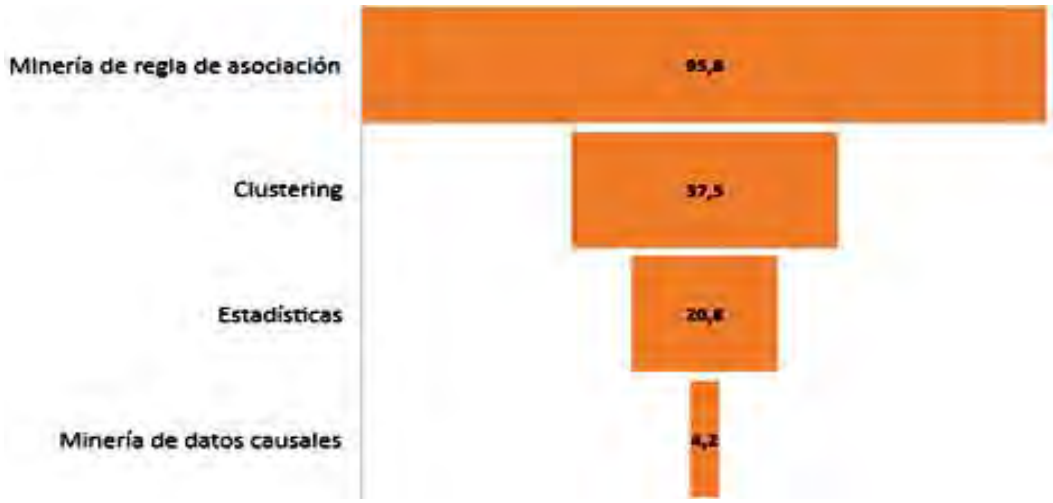


Figura 6.4.- Artículos ASI en los métodos de análisis según según la clasificación de Baker e Inventado y de Papamitsiou y Economides (Baker y Inventado, 2014; Papamitsiou y Economides, 2014; Pazmiño-Maji et al., 2016)

6.3 Aporte con técnicas ASI a LA: Desde 2016 hasta junio 2021

A continuación, se muestran aportes realizados mediante una nueva revisión sistemática de literatura respecto a las técnicas del ASI[5]. La revisión sistemática 2021, es similar a la realizada en el año 2016 con el objetivo de actualizarla y realizar una comparación adecuada, guardando diferencias en lo que se refiere al tiempo y los resultados, sin considerar otros factores que afecten la comparación. En este capítulo se actualizan los resultados a las preguntas MI1: ¿Qué métodos similares a los propuestos por la clasificación de Baker e Inventado se observan en las investigaciones educativas con un enfoque del ASI? y MI2: ¿Qué métodos similares a los propuestos por la clasificación de Papamitsiou y Economides se observan en las investigaciones educativas con un enfoque del ASI?

[5] https://drive.google.com/drive/folders/1lqp52Xcx2TnJAQlULeS2YC2eTrbH7lS6?usp=sharing

6.3.1 Preguntas de investigación (MI)

MI1: ¿Qué métodos similares a los propuestos por la clasificación de Baker e Inventado se observan en las investigaciones educativas con un enfoque del ASI?

MI2: ¿Qué métodos similares a los propuestos por la clasificación de Papamitsiou y Economides se observan en las investigaciones educativas con un enfoque del ASI?

6.3.2 Bases de datos bibliográficas utilizadas

Se examinaron los artículos científicos sobre ASI, cuya fuente fueron siete bases de datos bibliográficas: Biblioteca Digital ACM, EBSCO, Google Scholar, librería digital IEEE, ProQuest, base de datos de resúmenes y citación Scopus de Elsevier y la base de datos internacional Web of Science (WOS). La revisión se limitó a estudios publicados en los últimos meses 66, entre 2016 y el primer semestre del 2021. Las fuentes de datos y las características de la búsqueda se resumen en la Tabla 6.8.

Tabla 6.8.- Características de la base de datos y búsqueda bibliográficas (Pazmiño-Maji et al., 2016)

BASE DE DATOS BIBLIOGRÁFICAS		**ACM, EBSCO, Google Scholar, IEEE, ProQuest, Scopus, WOS**
BÚSQUEDA	Principales criterios de búsqueda	"*Statistical implicative analysis*" ASI
	Resultados (aplicando los criterios de exclusión y el tiempo)	21 documentos científicos sobre ASI
	Tiempo	66 meses desde enero del 2016 hasta junio del 2021
	Tópicos: Criterios de aproximación	La definición Las fuentes de datos Las etapas

Durante la búsqueda, el subtema (o sinónimo) se ha guiado por la Tabla 6.9.

Tabla 6.9.- Criterios de examinación utilizados (Pazmiño-Maji et al., 2016)

PREGUNTA DE INVESTIGACIÓN	CRITERIOS DE APROXIMACIÓN	SUBCRITERIOS
1	Aproximación de ASI a las LA, definición y fuente de datos	Entrada LA: * Aprendizaje * Contexto del aprendizaje * Fuente de datos Proceso LA: *medición *recopilación *análisis *comunicación Salidas LA: * comprensión * optimización
2	Datos de ingreso	* Aprendizaje del estudiante * Demografía * Percepciones * Procesos escolares
3	Aproximación de ASI a las etapas de las LA	* Captura * Informar * Predecir * Actuar * Refinar

6.3.3 Cadenas lógicas de búsqueda

El grupo de estudios primarios se definió basándonos en (Zhang & Ali Babar, 2010). La cadena de búsqueda final se describió como sigue: ("*Statistical implicative analysis*" OR SIA) AND **(**LIMIT-TO (PUBYEAR, 2021) OR LIMIT-TO (PUBYEAR, 2020) OR LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018) OR LIMIT-TO (PUBYEAR, 2017) OR LIMIT-TO (PUBYEAR, 2016)**)** como se indica en (Kutvonen, 2008; Tolk, Turnitsa, & Diallo, 2006), las publicaciones del primer semestre del 2021 y en el área de educación fueron filtradas después utilizando el software de gestión de referencias Citavi 6 (*Citavi - Reference Management and Knowledge Organization*, 2016).

6.3.4 Actualización de resultados pregunta MI1

La Tabla 6.10, muestra la referencia de artículos de ASI, su frecuencia absoluta y frecuencia porcentual que responde a la pregunta de investigación MI1 (¿Qué métodos similares a los propuestos por la clasificación de Baker e Inventado se observan en las investigaciones educativas con un enfoque del ASI?). Las referencias completas se encuentran numerada en el apéndice.

Tabla 6.10.- Artículos ASI en los métodos de análisis en LA según Baker e Inventado (Baker y Inventado, 2014) . Ver referencias completas en Apéndice

MÉTODOS DE ANÁLISIS EN LAS LA SEGÚN BAKER E INVENTADO		ARTÍCULOS	NÚMERO	FREQ (%)
Predicción	Clasificación		0	0
	Regresión		0	0
	Estimación del Conocimiento Latente		0	0
Minería de relaciones	Minería de reglas de asociación	[1], [2], [3], [5], [6], [7], [8], [9], [14], [17], [18], [19], [20], [21]	14	66,7%
	Minería de patrones secuenciales		0	0
	Minería de correlaciones		0	0
	Minería de datos causales	[5]	1	4,7%
Descubrimiento de estructura	Clúster	[8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [20], [21]	13	61,9%
	Análisis de factores		0	0
	Descubrimiento de estructuras de dominio		0	0
Descubrimiento con modelos			0	0
Descubrimiento con modelos			0	0

Se observa que el 66,7% de los artículos utilizan métodos del ASI similares a la minería de reglas de asociación que hacen parte de la minería de relaciones, el 61,9% de los artículos utilizan métodos del ASI similares a clúster para el descubrimiento de estructura, solo un 4,7% de los artículos utilizan métodos del ASI similares a la minería de datos causales que hacen parte de la minería de relaciones.

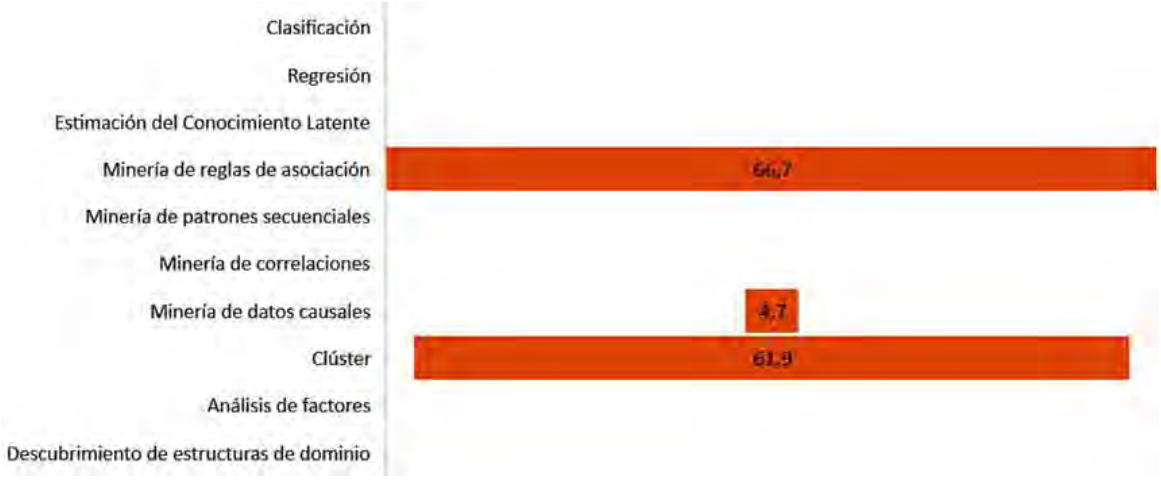


Figura 6.5.- Artículos ASI en los métodos de análisis según Baker e Inventado (Baker y Inventado, 2014)

La respuesta a la pregunta MI1 se resume en que los artículos que utilizan métodos del ASI similares a los métodos en las LA según la clasificación de Baker e Inventado son minería de reglas de asociación (66,7%), clúster (61,9%), minería de datos causales (4,7%), los otros métodos muestran no ser utilizados.

6.3.5 Actualización de resultados pregunta MI2

La Tabla 6.11, muestra los artículos del ASI, su número y porcentaje que responden a la pregunta 2 (¿Qué métodos similares a los propuestos por la clasificación de Papamitsiou y Economides se observan en las investigaciones educativas con un enfoque del ASI?). Es importante notar que muchos de los artículos analizados utilizaban más de una técnica de análisis ASI y por ello pueden contabilizarse dentro de varios métodos de las analíticas del aprendizaje. Hay que observar también que cuando se seleccionó la técnica clúster su indicador de utilidad era la búsqueda de agrupaciones y éstas se las encuentra mediante clústeres no jerárquicos y clústeres jerárquicos, pero en la selección no se hace diferencia entre los dos. Cuando se habla de estadística se refiere a ciertas medidas que se utilizan en los documentos ASI y son generados por los softwares CHIC y Rchic no con un

software adicional. Las medidas descriptivas a las que se hace referencia son: análisis de frecuencia, medidas descriptivas básicas, medidas de similaridad y medidas de correlación. El análisis de frecuencia es absoluto (el conteo de las ocurrencias), porcentual y entre pares de variables, las medidas descriptivas son la media aritmética y la desviación estándar, las medidas de similaridad se concentran en los índices de similaridad, las medidas de correlación se basan en el coeficiente de correlación de Pearson.

Tabla 6.11.- Artículos ASI en los métodos de análisis en LA según Papamitsiou (Papamitsiou y Economides, 2014). Ver referencias completas en Apéndice

MÉTODOS DE ANÁLISIS EN LAS LA	ARTÍCULOS EN LOS MÉTODOS DE ANÁLISIS DE LAS LA	NÚMERO	FREQ (%)
Clasificación		0	0
Clúster	[8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [20], [21]	13	61,9%
Regresión		0	0
Minería de texto		0	0
Minería de regla de asociación	[1], [2], [3], [5], [6], [7], [8], [9], [14], [17], [18], [19], [20], [21]	14	66,7%
Análisis de redes sociales		0	0
Descubrimiento con modelos		0	0
Visualización		0	0
Estadísticas	[3], [4], [6], [9], [11], [15], [17]	7	33,3%

Se observa que el 66,7% de los artículos utilizan métodos del ASI similares a la minería de reglas de asociación, que el 61,9% de los artículos utilizan métodos del ASI similares a clúster, que un 33,3% de los artículos utilizan métodos del ASI similares a la estadística.

La respuesta a la pregunta MI2 (Figura 6.6) se refiere a que los artículos que utilizan técnicas del ASI que aportan en las LA según la clasificación de Papamitsiou y Economides son: minería de reglas de asociación (66,7%), clúster (61,9%), estadísticas (33,3%), las otras técnicas no muestran ser utilizadas.

Figura 6.6.- Artículos ASI en los métodos de análisis según la clasificación de Baker e Inventado y de Papamitsiou y Economides (Baker y Inventado, 2014; Papamitsiou y Economides, 2014)

6.4 Conclusiones

A continuación, se comparan las revisiones sistemáticas de literatura realizadas en el 2016 y 2021 de los métodos de análisis ASI que aportan a LA, según la clasificación de Baker e Inventado (Baker y Inventado, 2014)y la clasificación de Papamitsiou (Papamitsiou y Economides, 2014). Según Baker e Inventado (Baker y Inventado, 2014), la minería de relaciones y el descubrimiento de estructura son los métodos en los cuales aportan más las técnicas ASI a LA (Tabla 6.12).

Tabla 6.12.- Comparativo de MI1 sobre métodos de análisis en ASI que aportan en LA según Baker e Inventado (Baker y Inventado, 2014). Ver referencias completas en Apéndice

MÉTODOS DE ANÁLISIS EN LAS LA SEGÚN BAKER E INVENTADO		2011-jun2016		2016-jun2021	
		NÚMERO_A[6]	FREQ_A	NÚMERO_D	FREQ_D
Predicción	Clasificación	0	0	0	0
	Regresión	0	0	0	0
	Estimación del Conocimiento Latente	0	0	0	0
Minería de relaciones	Minería de reglas de asociación	23	95,8%	14	66,7%
	Minería de patrones secuenciales	0	0	0	0
	Minería de correlaciones	0	0	0	0
	Minería de datos causales	1	4,2%	1	4,7%
Descubrimiento de estructura	Clúster	9	37,5%	13	61,9%
	Análisis de factores	0	0	0	0
	Descubrimiento de estructuras de dominio	0	0	0	0
Descubrimiento con modelos		0	0	0	0
Descubrimiento con modelos		0	0	0	0

[6] _A representa los resultados de la revisión sistemática de literatura entre los años 2011 hasta junio del 2016.
_D representa los resultados de la revisión sistemática de literatura entre los años 2016 hasta junio del 2021.

Según Baker e Inventado, el método clúster (61,9%) y la minería de regla de asociación (66,7%) aportan en porcentaje similar según la revisión de literatura actual, en cambio en el estudio de literatura anterior había una diferencia notable (37,5% y 95,8%). El aporte de la minería de datos causales (alrededor de 4%) se ha mantenido bajo y similar en los últimos 11 años de estudio (Figura 6.7).

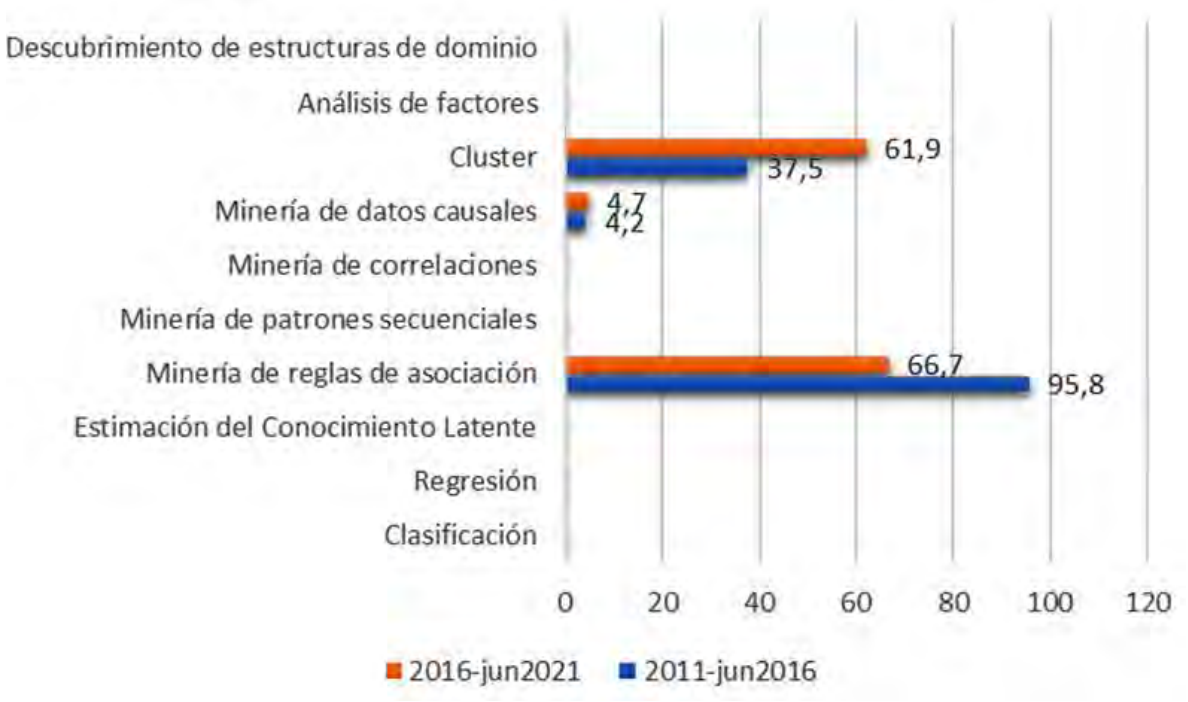


Figura 6.7.- Gráfico de Barras comparativo de MI1 sobre métodos de análisis en ASI que aportan en LA según Baker e Inventado (Baker y Inventado, 2014) de las revisiones sistemáticas realizadas desde el 2011 hasta el 2021

Según Papamitsiou (Papamitsiou y Economides, 2014), la minería de reglas de asociación fue (95,8%) y sigue siendo (66,7%) los métodos en los cuales aportan más las técnicas ASI a LA (Tabla 6.13).

Tabla 6.13.- Comparativo de MI2 sobre métodos de análisis en ASI que aportan en LA según Papamitsiou (Papamitsiou y Economides, 2014). Ver referencias completas en Apéndice

MÉTODOS DE ANÁLISIS EN LAS LA	2011-jun2016		2016-jun2021	
	NÚMERO_A	FREQ_A	NÚMERO_D	FREQ_D
Clasificación	0	0	0	0
Clúster	9	37,5%	13	61,9%
Regresión	0	0	0	0
Minería de texto	0	0	0	0
Minería de regla de asociación	23	95,8%	14	66,7%
Análisis de redes sociales	0	0	0	0
Descubrimiento con modelos	0	0	0	0
Visualización	0	0	0	0
Estadísticas	5	20,8%	7	33,3%

En la revisión de literatura actual 2021, el método clúster (61,9%) aporta casi con el doble a la revisión de literatura inicial del 2016 (37,5%). La minería de reglas de asociación ha disminuido del 95,8% al 66,7% en la revisión de literatura actual, referente a los métodos en los cuales aporta más las técnicas ASI en LA (Figura 6.8).

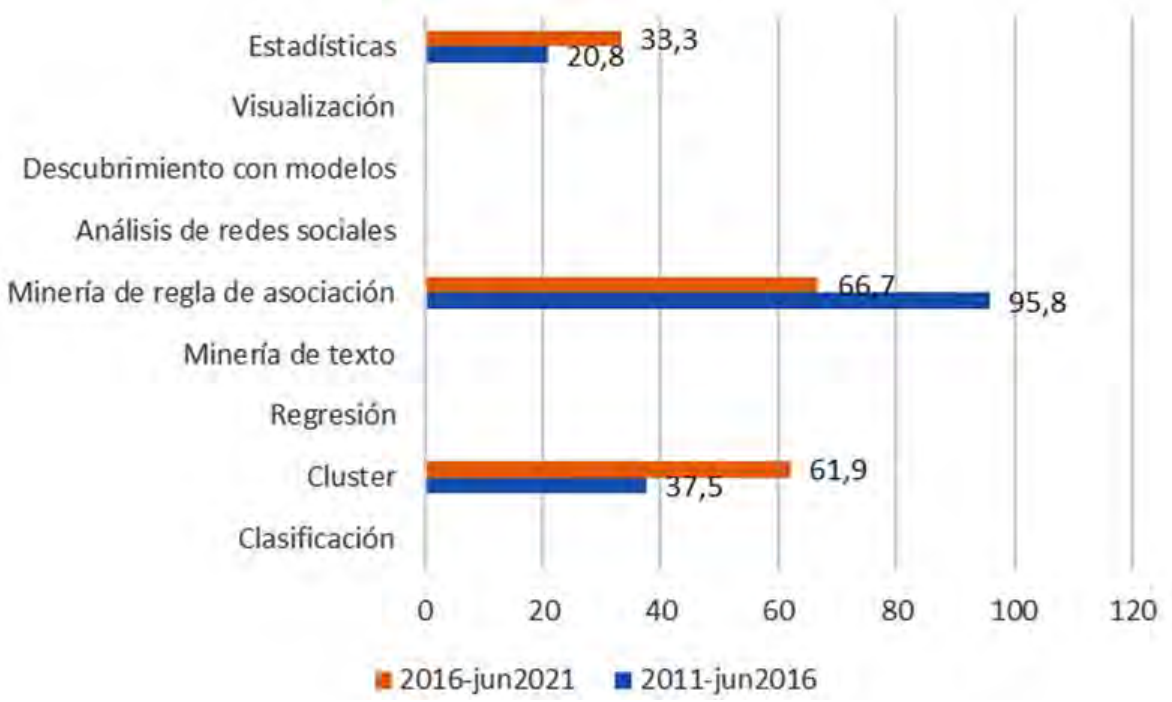


Figura 6.8.- Gráfico de Barras comparativo de MI2 sobre métodos de análisis en ASI que aportan en LA según Papamitsiou (Papamitsiou y Economides, 2014) de las revisiones sistemáticas realizadas desde el 2011 hasta el 2021

En cuanto a los resultados estadísticos referentes al análisis estadístico implicativo han aumentado, pues en la revisión sistemática de literatura del año 2016 se utilizaban en 20,8% pero en la actualización de la revisión sistemática es aproximadamente 10% más, es decir, 33,3%.

Capítulo 7mo | COMPLEJIDAD ALGORÍTMICA ENTRE TÉCNICAS CLÚSTER DE LA Y ASI

Se comparan las técnicas clúster diana, dendro.variables y hclust.vector de LA con las técnicas SimilarityTree y callHierarchyTree del ASI, desde el punto de vista de la complejidad algorítmica.

7 Capítulo.- Complejidad algorítmica entre técnicas clúster de LA y ASI

En este capítulo se comparan computacionalmente las técnicas clúster de LA y las técnicas similares proporcionadas por el ASI, es decir, se compara experimentalmente la complejidad algorítmica. Los parámetros espacio y tiempo servirán para comparar funciones entre sí, permitiendo determinar el más adecuado entre varios que solucionan un mismo problema, siendo el objetivo principal de éste y el próximo capítulo. Primeramente, se realiza un análisis respecto a la ocupación de memoria de los métodos clúster y luego se realiza un estudio del tiempo de ejecución de las mismas técnicas clúster (Naranjo et al., 2018) y (R. Pazmiño-Maji et al., 2017b). En la introducción se detalla la complejidad algorítmica, generalidades sobre las técnicas clúster y los materiales y métodos utilizados.

7.1 Introducción

El concepto de complejidad algorítmica es importante tenerlo presente pues se lo aplica en los Capítulos 7 y 8, se muestra los principios de los métodos clúster, se realiza una revisión de artículos ASI relacionados y finalmente se muestra la metodología utilizada para la comparación de la complejidad espacial y temporal. Todos los detalles sobre las técnicas clúster utilizadas se encuentran en el Apéndice D.- Manual de estadísticas utilizadas.

7.1.1 Complejidad Algorítmica

Los algoritmos deben ser capaces de resolver problemas amplios y también utilizar un menor tiempo y memoria, pero no es tan común que se encuentre algoritmos que sean capaces de cumplir estas dos características, por lo que se buscan algoritmos que intenten cumplir con esto. Al analizar un algoritmo es importante evaluar su viabilidad por lo que entramos a estudiar la complejidad del algoritmo. El estudio de complejidad suele medirse en función de dos parámetros: el espacio, es decir, la memoria que utiliza, y el tiempo, lo que tarda en ejecutarse. Ambos representan los costes que supone encontrar la solución al problema planteado, en nuestro caso encontrar grupos con elementos homogéneos y heterogéneos entre sí (Sección 7.1.2). Dichos parámetros (espacio y tiempo) van a servir además para comparar algoritmos entre sí, permitiendo determinar el más adecuado entre varios que solucionan un mismo problema, siendo este el objetivo

principal del presente y próximo capítulo. La complejidad algorítmica ayuda a describir el comportamiento de un algoritmo en términos de tiempo de ejecución, es decir, el tiempo que tarda un algoritmo en resolver un problema y por tanto permite determinar la eficiencia de dicho algoritmo, a esto se conoce como complejidad temporal (Dorta et al., 2003).

La medida del tiempo tiene que ser independiente de la máquina, del lenguaje de programación, del compilador y de cualquier otro elemento hardware o software que influya en el análisis. La complejidad temporal se expresa como T(n), en esta tesis analizamos la Tmed(n) (que expresa la complejidad temporal en el caso promedio y es una medida apropiada para la comparación) y no la Tmax(n) (que representa la complejidad temporal en el peor de los casos) ni la Tmin(n) (que trata sobre la complejidad en el mejor de los casos posibles).

La memoria requerida, es decir, la cantidad de memoria necesaria para procesar las instrucciones que solucionan dicho problema se denomina complejidad espacial. Si el tamaño de los datos es grande lo que importa es la eficiencia que tendrá el algoritmo. Una variante del algoritmo (respecto a la memoria) es memorizar solo los datos que más se repiten de forma que el espacio requerido en RAM es fijo e independiente del tamaño de los datos. El efecto en la complejidad del tiempo de ejecución es difícil de calcular pues depende del tamaño de la memoria respecto de los valores del tamaño de datos que necesitemos (Fillottrani, 2009).

La complejidad de un algoritmo se encuentra en función del tamaño del problema, pero esto no siempre sucede así, especialmente en los algoritmos de búsqueda, dentro del cual están los algoritmos de búsqueda secuencial, secuencial ordenada, y binaria y los algoritmos de ordenamiento en donde también encontramos los algoritmos cuadráticos y avanzados de búsqueda. A un conjunto de funciones que comparten un mismo comportamiento se denomina un orden de complejidad. Habitualmente estos conjuntos se denominan O, de esta manera se agrupan todas las complejidades que crecen de igual forma, es decir, que pertenecen al mismo orden.

La Tabla 7.1, muestra los órdenes de complejidad en forma creciente, es decir un algoritmo con orden O(1) será más eficiente que uno de orden O(n) y el anterior más eficiente que O(2n), esto es válido tanto en espacio como en tiempo (Vásquez, 2004).

Tabla 7.1.- Órdenes de complejidad (Vásquez, 2004)

Función	Nombre
$O(1)$	Orden constante
$O(\log n)$	Orden logarítmico
$O(n)$	Orden lineal
$O(n \log n)$	Orden cuasi-lineal
$O(n^2)$	Orden cuadrático
$O(n^3)$	Orden cúbico
$O(n^a)$	Orden polinómico
$O(2^n)$	Orden exponencial

Cuando se va a resolver un problema se debe elegir un buen algoritmo con el que podamos fijarnos en su complejidad, que no utilice tantos recursos requeridos como son el tiempo que tarda en ejecutarse y la cantidad de espacio de memoria, y así decidir cuál es mejor algoritmo. Todos los algoritmos no son iguales ni funcionarán de manera similar, cada uno tiene su función y por eso se los clasifica en su respectivo orden de complejidad (Tabla 7.1).

7.1.2 Agrupación jerárquica

La agrupación (clúster) jerárquica permite formar grupos heterogéneos de elementos homogéneos, y éstos se subdividen en dos tipos de algoritmos:

- Agrupación jerárquica (aglomerativa) en la que, cada observación se considera inicialmente como un racimo propio (nivel 0). Luego, los clústeres más similares se fusionan sucesivamente hasta que haya un solo gran clúster raíz (nivel n) (Figura 7.1).
- Agrupación jerárquica (divisiva), es una inversión de la agrupación jerárquica (aglomerativa), comienza con la raíz (nivel n), en la que todos los objetos están incluidos en un grupo (hasta el momento el más heterogéneo). Las agrupaciones se dividen sucesivamente hasta que todas las observaciones se encuentran en su propio grupo (nivel 0).

La agrupación o clúster jerárquico (aglomerativa) es una alternativa a la agrupación jerárquica (divisiva), para agrupar objetos basados en su similitud. A diferencia de la

agrupación jerárquica (divisiva), la agrupación jerárquica (aglomerativa) no requiere pre-especificar el número de clústeres que se producirán (Kassambara, 2017).

Los algoritmos de agrupación jerárquica (aglomerativa), son métodos de agrupación que se utilizan para clasificar observaciones, dentro de un conjunto de datos, en varios grupos en función de su similaridad. Algunos algoritmos comunes de agrupación particionada son:

- Agrupación de K-medias (MacQueen, 1967), en la que, cada agrupación está representada por el centro o por medio de los puntos de datos pertenecientes al clúster. El método es sensible a puntos de datos anómalos y valores atípicos.
- Agrupación de K-medias o PAM (Kaufman y Rousseeuw, 1990), en el que cada grupo está representado por uno de los objetos en el racimo. PAM es menos sensible a valores atípicos en comparación con k-medias.
- Algoritmo CLARA (Clúster Large Applications), que es una extensión de PAM adaptado para grandes conjuntos de datos(Kassambara, 2017).

El resultado de la agrupación jerárquica (aglomerativa o divisiva) es una representación basada en árboles de los objetos, también se conoce como dendrograma (Figura 7.1).

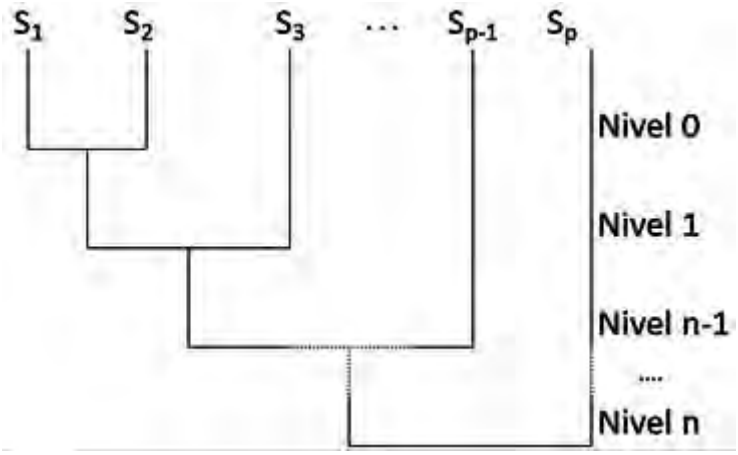


Figura 7.1.- Agrupación jerárquica aglomerativa y niveles jerárquicos

En la agrupación jerárquica, cada observación se considera inicialmente como un grupo por sí mismo (nivel 0). Luego, los conglomerados más similares se fusionan

sucesivamente (nivel 1, nivel 2, Nivel n) hasta que haya un solo gran conglomerado (nivel n). La agrupación jerárquica utiliza una forma de abajo hacia arriba, es decir, cada objeto es inicialmente considerado como un grupo de un solo elemento (hoja). En cada paso del algoritmo, los dos grupos que son los más similares se combinan en un nuevo grupo más grande (nodos). Este procedimiento se repite hasta que todos los puntos son miembros de un solo gran grupo raíz (Figura 7.1).

La inversa de la agrupación jerárquica aglomerativa es la agrupación jerárquica divisiva, que también se conoce como DIANA (DIvise ANAlysis) y comienza con la raíz, en el que todos los objetos se incluyen en un solo grupo, en cada paso de la interacción, el grupo más heterogéneo se divide (Kassambara, 2017). El proceso se repite hasta que todos los objetos están en su propio grupo menos heterogéneo (Figura 7.2).

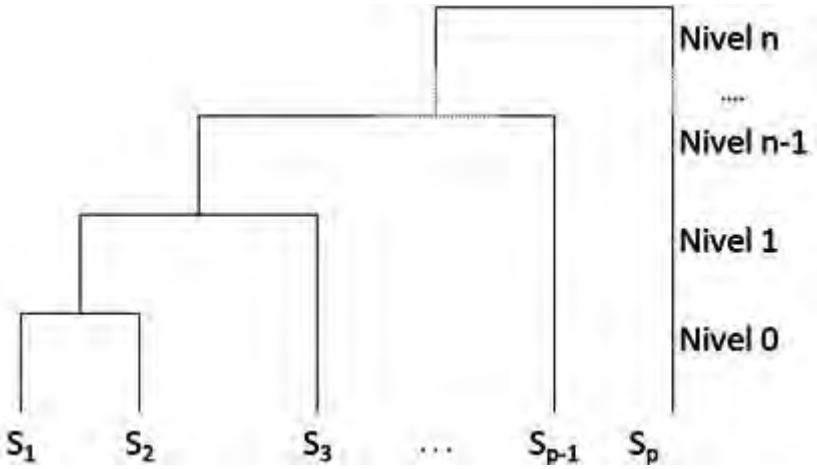


Figura 7.2.- Agrupación jerárquica divisiva y niveles jerárquicos

La agrupación jerárquica (aglomerativa) generalmente se utiliza para identificar agrupaciones pequeñas, en cambio la agrupación jerárquica (divisiva) se aplica para identificar agrupaciones grandes.

El estudio de la técnica clúster permitirá comprender de mejor manera tanto los trabajos relacionados (ver sección inmediata inferior) así como el estudio de la complejidad espacial y complejidad temporal (secciones 7.3 y 7.4).

7.1.3 Trabajos relacionados

A continuación, indicamos algunos estudios comparativos entre las técnicas ASI y otras técnicas de análisis.

Un primer estudio se basa en el artículo de (Michael et al., 2010), donde se desea conocer las características y ventajas del método implicativo del ASI y dos métodos estadísticos de análisis: la agrupación jerárquica de variables y el análisis factorial confirmatorio (CFA). Se utilizaron los resultados en la aplicación de las tres técnicas en la aprehensión operativa de la figura geométrica, se trabajó con datos de 125 alumnos de sexto curso. Mediante el Análisis Factorial Confirmatorio, se desarrolla y verifica un modelo que proporciona información sobre el papel significativo de la modificación mereológica[7], óptica y de la forma del lugar en la aprehensión operativa de la figura geométrica. Utilizando la agrupación jerárquica de las variables, se proporciona evidencia al fenómeno de la segmentación entre las modificaciones en la aprehensión operativa de los estudiantes. En general, se encontró que los resultados de los tres métodos coinciden y pueden ser complementarios (Tabla 7.2) para captar las formas en que los estudiantes utilizan los diferentes tipos de modificación de la figura.

Tabla 7.2.- Comparación entre CFA, agrupación jerárquica y el método implicativo (Michael et al., 2010)

CFA	Agrupación jerárquica	Método implicativo
Estructura factorial de la comprensión de figuras geométricas. Desarrollo de un modelo que incluye dos factores latentes para los efectos de tres tipos de modificación de la figura y un factor de segundo orden que representa la aprehensión operativa de la figura geométrica. Diferencia en la fuerza de las relaciones de los tres factores de primer orden con el factor de segundo orden.	Clasificación jerárquica y consistencia de las respuestas a las modificaciones de figuras geométricas. Agrupaciones de similitud entre las medidas observadas en las respuestas a las tres formas de modificar una figura geométrica. Agrupación separada de las variables de la modificación mereológica y la óptica. Similitud relativamente débil de las modificaciones.	Relaciones entre las respuestas de los alumnos a las modificaciones de las figuras geométricas. Implicaciones entre las variables observadas en las respuestas de los alumnos a los tres tipos de modificaciones de las figuras geométricas. Las tareas de modificación de la forma fueron más complejas que las tareas de modificación de la forma mereológica u óptica. Las tareas de modificación óptica fueron las más fáciles.

En el artículo (R. Pazmiño Maji et al., 2017), se analiza la posibilidad de que el árbol jerárquico del ASI pueda cumplir la principal función del clúster jerárquico aglomerativo

[7] La mereología es, dentro de la lógica matemática y la filosofía, el estudio de las partes de un conjunto, analizando la relación de las partes entre sí y la de las partes con el todo («Mereología | Qué es, Definición y Concepto.», 2021)

que es la de agrupar objetos (además se midió el nivel de acuerdo con las agrupaciones realizadas), a las conclusiones se llegaron mediante la observación directa realizada por 35 estudiantes universitarios. Se comprobó que el 69,14% de participantes están fuertemente de acuerdo con las agrupaciones (R. Pazmiño Maji et al., 2017).

Sobre la complejidad algorítmica entre técnicas ASI y otras técnicas clúster, se encontraron los siguientes trabajos:

El artículo científico (R. Pazmiño-Maji et al., 2017b) fija la metodología y características a utilizar para aplicar la comparación de la complejidad entre los árboles jerárquicos del ASI y el clúster jerárquico utilizado en LA. Las principales conclusiones a las que se llegaron con un nivel de error del 5% fueron: las muestras son independientes debido a que son aleatorias, la homogeneidad de varianzas es falsa (con un p-valor <2,2e-16), las muestras no han sido extraídas de una población normal (con un p-valor <2,2e-16). La diferencia en la complejidad temporal entre los algoritmos de cohesión, similaridad, agnes y hclust es altamente significativa (con p-valor <2,2e-16), son necesarias post pruebas 2 a 2 en el futuro. Además, se sugiere continuar investigando con otros sistemas operativos, que se utilicen más de 100000 datos y los diferentes métodos, métricas y opciones como factores (R. Pazmiño-Maji et al., 2017b).

La elaboración de la tesis titulada Estudio comparativo del ASI y LA en relación con el uso de las técnicas de exploración de datos educativos, fue motivada por el autor de esta tesis doctoral y elaborada juntamente con el Ing. Mauricio Naranjo. Los objetivos propuestos fueron (1) identificar las técnicas similares entre el ASI y LA, mediante la adaptación del método de estudio de similitud entre modelos y estándares (MSSS), (2) identificar el sistema operativo con mejor manejo de recursos y (3) identificar la técnica óptima en el análisis de datos educativos. Las principales conclusiones a las cuales se llegaron fueron: que existen técnicas similares de agrupación entre LA (dendro_variable, dendro_diana y hclust vector) y ASI (hrarchy y simlrty) y las técnicas similares de reglas de asociación entre LA (apriori, eclat, weclat) y ASI (implicativeGraph). El sistema operativo Ubuntu presenta mejor administración de los recursos, como la asignación de procesos a memoria, existe homogeneidad en el uso de memoria entre las técnicas reglas de asociación similares de LA y ASI, las técnicas de LA y ASI son similares entre ellas (la óptima weclat por ocupar menos memoria), pero esto no implica que sea el que tenga

menores tiempo de respuesta. Existe homogeneidad en el tiempo de ejecución entre las técnicas reglas de asociación similares de LA y ASI (la más óptima por tener menor tiempo de respuesta es implicativeGraph) y la menos óptima pero no menos importante met_apriori (Naranjo Serrano y Pazmiño Maji, 2018a).

El artículo reciente de (Fotiadis y Anastasiadou, 2019) compara las técnicas ASI (similaridad y cohesión) con el Análisis de Componentes Principales (PCA), con respecto al comportamiento del consumidor. El PCA permite el reconocimiento de patrones, es un método no supervisado, que se basa en el principio de la no existencia de información a priori (los componentes principales no se conocen de antemano), pero se logran como resultado de la aplicación del método PCA. Los componentes principales se calculan jerárquicamente. Los resultados de la aplicación de los métodos han señalado sus diferencias y similitudes (Tabla 7.3), pero también su complementariedad. Se observa que la aplicación de PCA dio como resultado una reducción de datos y mostró que hay tres componentes principales (variables latentes) que interpretan toda la variabilidad, así como los resultados del ASI en los árboles de similaridad y cohesión.

(Fotiadis y Anastasiadou, 2019) demostraron que los dos métodos (PCA y ASI) operan de manera complementaria, cada uno acentuando una dimensión diferente para la interpretación de los datos, cuya interpretación no habría sido determinante sin la participación de los especialistas en marketing.

Tabla 7.3.- Diferencias de los dos métodos (Fotiadis y Anastasiadou, 2019)

ASI	PCA
• Se basa en reglas • Se basa en un modelo probabilístico • Destaca tendencias en un conjunto de propiedades • Genera reglas de asociación • Proporciona una medida no lineal	• PCA no se basa en condiciones • Se basa en distancias espaciales métricas • Sus patrones se basan en la correlación entre variables • Proporciona una medida lineal • Visualiza correlaciones entre las variables originales y entre estas variables y los componentes • Visualiza proximidades entre unidades estadísticas • Es una técnica estadística que se emplea con frecuencia para reducción de dimensión
Propiedades entre las variables • La relación entre variables es asimétrica. • Las medidas de asociación no son lineales y se basan en probabilidades	Propiedades entre las variables • La relación entre variables es simétrica • Las medidas de asociación son lineales
	• Reducción de datos • Detección de datos y establecimiento de una estructura/ modelo • Establecimiento de variables latentes • Detección de fuentes latentes de variabilidad y covariabilidad en medidas observables • Detección de patrones
• Representado por el árbol de similaridad. • Representado por el árbol de implicaciones. • Representado por el árbol de cohesión	• Representado por plano factorial

7.2 Materiales y métodos

Para el estudio se utilizaron tres computadores con el mismo microprocesador: Intel® Core ™ i7-CPU @ 2.2 Ghz y 8Gb de memoria RAM, se ha instalado los sistemas operativos Windows 8-64 bits, Linux – Ubuntu 16.04-64 bits y MAC OS 10-64 bits. Todos los computadores y sistemas operativos trabajaron con el software estadístico libre R, versión 3.4.1; el entorno de desarrollo integrado RStudio, versión 1.0,143 y el paquete Rchic, versión 0,24. Las bases de datos se generaron aleatoriamente utilizando la función runif() perteneciente al paquete estándar de R. Los datos utilizados fueron dicotómicos

generados por la función runif() y round(). Las funciones utilizadas en LA fueron hclust_vector, dendro_variables y diana; y las utilizadas en ASI fueron: hrarchy y simlrty.

El cálculo aproximado del tamaño de la población se muestra en forma detallada en el Apéndice. Por el gran tamaño de la población, se escogió una muestra utilizando el método de muestreo aleatorio simple con parámetro de interés la media, se consideró la fórmula para el cálculo de la muestra $n = \frac{S^2}{\frac{E^2}{Z^2_{\frac{\alpha}{2}}} + \frac{S^2}{N}}$.

Para aplicar la fórmula se utilizaron los parámetros desviación estándar=1; α=5%; Z=1,96; E=10,01%; N=100000 y se generó un tamaño de la muestra de 383,2 que redondeado es 383. Las hipótesis estadísticas que se demostraron fueron normalidad, test de hipótesis de Kruskal-Wallis y su respectivo post test. Para demostrar las hipótesis se planteó un pre-experimento en la ingeniería de software de tipo RGXO1. Donde RG representa el grupo aleatorio del grupo experimental (tanto-inter como intra-grupos), X representa el tratamiento que en este caso son las 3 técnicas clúster jerárquicos utilizadas en LA (hclust_vector, dendro_variables y dendro_diana) y 2 técnicas usadas en ASI (hrarchy y simlrty). Se trabajó con un nivel significancia de α=0,05. Las variables dependientes fueron la variable memoria (para el caso de la complejidad espacial) y la variable tiempo (para el caso de la complejidad temporal), ambas de tipo numérico.

Por el paradigma de investigación es de tipo cuantitativo, por el tipo de diseño utilizado es pre-experimental, por el tiempo de estudio es transversal, el colectivo de estudio lo conforman las 100000 bases de datos aleatorias formadas por lo máximo 1000 observaciones y 100 variables, por la amplitud de estudio es de muestreo de 383 bases de datos aleatorias binarias. La población es la información sobre la muestra de estudio, tal como: nombre del archivo, número de filas que conforman la base de datos, número de columnas que conforman la base de datos, el total de datos, tiempo de ejecución, memoria utilizada y sistema operativo.

7.3 Estudio de la complejidad espacial

A continuación, se realiza el estudio espacial respecto a los métodos clúster. Primero se realiza un estudio descriptivo, para luego realizar las pruebas de hipótesis, no sin antes realizar la prueba de los supuestos de normalidad, homocedasticidad e independencia para determinar el tipo de prueba a utilizar.

A continuación, se determinaron las estadísticas descriptivas resumen entre las variables, para lo que corresponde a los métodos clúster se almacenó en una variable X y lo que corresponde a memoria en la variable Y.

7.3.1 Medidas descriptivas

La siguiente tabla resume los resultados descriptivos de la ocupación de memoria por método clúster utilizado en el análisis (Tabla 7.4).

Tabla 7.4.- Medidas descriptivas de los métodos clúster

CLÚSTER METHODS	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
dendro_diana	104,0	141,5	247,0	223,0	278,0	382,0
dendro_variables	105,0	151,0	256,0	231,3	286,0	384,0
hclust_vector	103,0	141,5	247,0	223,0	278,0	381,0
hrarchy	105,0	144,0	249,0	224,8	280,0	319,0
simlrty	105,0	248,0	143,0	223,9	279,0	319,0

El método que usa en promedio mayor memoria es el dendro_variables con un valor igual a 231,3 con un mínimo de memoria igual a 105 y un máximo de 384, el primer cuartil muestra que el 25% de la memoria es menor o igual a 151 con una mediana que indica que la mitad de la memoria empleada para el método es menor o igual a 256 y la otra mitad es mayor o igual a 256, finalmente el tercer cuartil refleja que el 75% de la memoria es menor o igual a 286. El método que utiliza menos memoria es dendro_diana y hclust_vector, con respecto al método dendro_diana se determinó que posee un valor promedio igual a 223 con un mínimo de memoria igual a 104 y un máximo de 382, el primer cuartil muestra que el 25% de la memoria es menor o igual a 141,5 con una mediana que indica que la mitad de la memoria empleada para el método es menor o igual a 247 y la otra mitad es mayor o igual a 247, el tercer cuartil refleja que el 75% de la memoria es menor o igual a 278; en lo que compete al método hclust_vector se obtuvo un valor promedio igual a 223 con un mínimo de memoria igual a 103 y un máximo de 381, el primer cuartil muestra que el 25% de la memoria es menor o igual a 141,5 con una

mediana que indica que la mitad de la memoria empleada para el método es menor o igual a 247 y la otra mitad es mayor o igual a 247, el tercer cuartil refleja que el 75% de la memoria es menor o igual a 278, con este análisis se evidenció que no existe mayor diferencia entre estos métodos a excepción del máximo y el mínimo. Para concluir, se determinó que para los métodos clúster en estudio no existe mayor diferencia en la memoria usada en promedio.

A continuación, se realiza un estudio descriptivo basado en tablas de frecuencia, medidas de posición, centralización, dispersión, posición y gráficos comparativos.
La Tabla 7.5, refleja los resultados sobre la cantidad de memoria que se empleó por cada clúster, se analizaron parámetros relevantes que se describen a continuación.

Tabla 7.5.- Cantidad de memoria por método clúster

CLÚSTER METHODS	mean	sd	IQR	cv	skewness	kurtosis	MEMORY: n
dendro_diana	222,9568	66,81823	136,5	0,2996914	-0,494546	-1,406468	3447
dendro_variables	231,3444	67,38646	135	0,2912821	-0,4921238	-1,382432	3447
hclust_vector	223,007	66,98882	136,5	0,3003889	-0,4906563	-1,408235	3447
hrarchy	224,81	66,6016	136	0,2962573	-0,5075369	-1,430878	3447
simlrty	223,8985	66,5712	136	0,2973276	-0,5083698	-1,431165	3447

En primera instancia se determinó la media, la cual indica que en promedio el método que ocupa menos memoria es Diana con un valor igual a 222,9568 y el método que en promedio usa mayor cantidad de memoria es el dendro_variables con un valor igual a 231,3444, cabe recalcar que entre hclust_vector, hrarchy y simlrty no existe mayor diferencia entre sus valores promedios obtenidos, al analizar la dispersión en la que se encuentran dichos datos con respecto al valor promedio que presentaron se obtuvo que aquel método que posee menor dispersión con 66,5712 es el simlrty y el de mayor dispersión es el dendro_variables con 67,38646, los demás métodos como dendro_diana, hclust_vector y hrarchy conservan una dispersión similar con valores de 66,81823, 66,98882 y 66,6016 respectivamente. El rango intercuartil (IQR) da a notar que el método que presenta menor variación en sus datos es el dendro_variables con un valor de 135, mientras que con una mayor variación igual a 136,5 se encuentra Diana y hclust_vector, para simlrty y hrarchy se registró una variación similar de 136; con respecto al coeficiente de variación (cv) se tiene que el método de mayor coeficiente es hclust_vector con un valor de 0,3003889% y el menor con 0,2912821% es dendro_variables. Cabe recalcar

que no existe mayor diferencia entre la dispersión de los datos de los métodos mencionados ya que la diferencia entre cada uno recae solo en los decimales. La asimetría de los datos (*skewness*) permitió notar que todos los métodos clúster poseen una asimetría negativa, es decir que los datos están sesgados a la izquierda, al evaluar cuál de los métodos clúster posee mayor asimetría se denotó que es el hclust_vector con -0,4906563 y el de menor asimetría es simlrty con -0,5083698. Los coeficientes de *Kurtosis* obtenidos para los métodos clúster son negativos por lo que la distribución que siguen estos datos es levemente platicúrtica con un valor mayor para dendro_variables (-1,382432) y menor para simlrty (-1,431165). Todos los métodos clúster se trabajaron con una muestra de 3447 bases de datos diferentes generadas de manera aleatoria.

Se procedió a realizar un gráfico de violín comparativo para cada uno de los 5 métodos analizados, éste se muestra a continuación en la Figura 7.3.

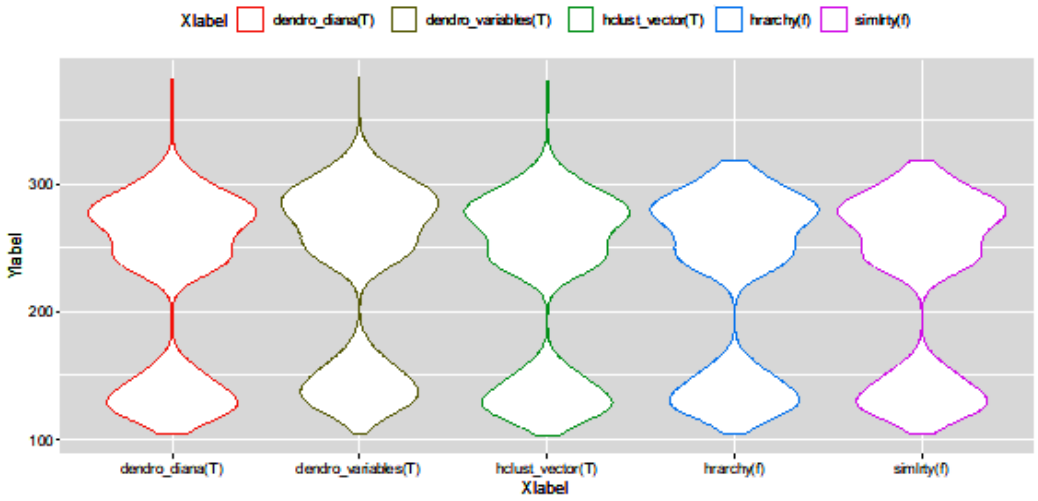


Figura 7.3.- Gráfico de violín sobre la cantidad de memoria por método clúster

Para obtener una mejor visualización del comportamiento de los datos se realizó una gráfica de violín que refleja la función de densidad de éstos, denotándose así que los datos sobre la cantidad de memoria empleada con cada método clúster poseen dos patrones de comportamiento distinto con media y variabilidad bien marcadas. Con respecto al rango se evidencia en la forma de cada violín que los métodos diana, dendro_variables y hclust_vector son similares, es decir, los datos se encuentran dispersos, poseen mayor variabilidad, mientras que para los métodos simlrty y hrarchy se

refleja menor variabilidad debido a que su distribución se encuentra más pequeña dado que sus violines no presentan bigotes. El método que refleja que usa menor cantidad de memoria es el simlrty y hrarchy y con mayor uso de memoria es el método Diana.

7.3.2 Normalidad

Para determinar la prueba apropiada a utilizar en la hipótesis se procedió a la comprobación de los supuestos. Las pruebas paramétricas tienen más potencia estadística que las pruebas no paramétricas, es por ello que se hacen esfuerzos en los estudios de normalización (Kanyongo et al., 2007).

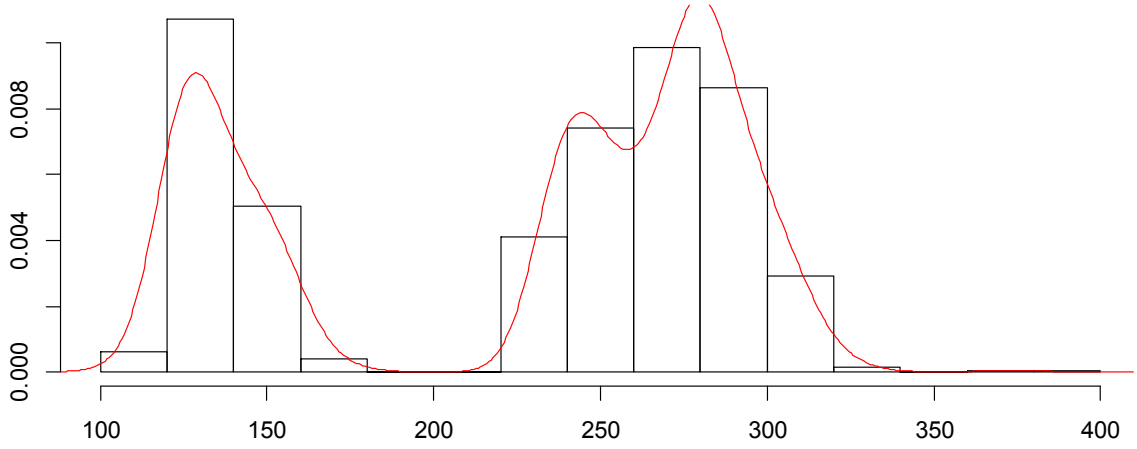


Figura 7.4.- Aproximación normal de los datos de memoria

La Figura 7.4, de aproximación normal de los datos de memoria muestra que tienen dos patrones de comportamiento totalmente marcados (histogramas bimodales), se podría considerar que tienen medias y varianzas distintas, esto se pudo haber dado debido a que el histograma se encuentra achatado considerando que puede estar presentando múltiples modas, pero hay pocas diferencias entre ellas. El histograma en el intervalo de 100 a 200 presenta un leve sesgo a la derecha por lo cual se lo considera asimétricamente positivo, mientras que del intervalo 200 a 350 posee una distribución aparentemente normal. Finalmente se destaca que posiblemente existen dos poblaciones distintas en el conjunto de datos, pudiendo ser consideradas estas poblaciones en grupos de aquellas que ocupen mayor y menor memoria.

A continuación, se muestra la gráfica de cuartiles (Figura 7.5) que provee una idea de la normalidad de los datos sobre el índice de rangos (cuartiles) en los diferentes métodos.

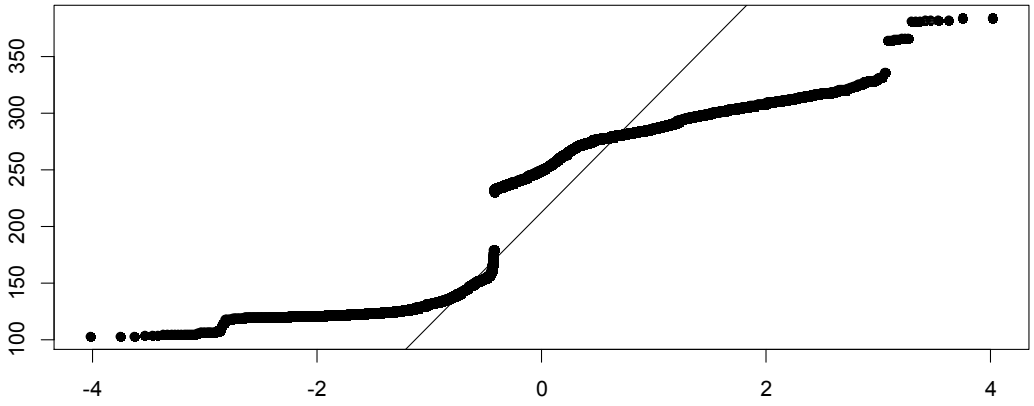


Figura 7.5.- Gráfico de cuartiles QQ para los datos de memoria

Al analizar la cercanía de la recta a la curva presentada en la Figura 7.5, sobre la cantidad de memoria usada por cada uno de los métodos clúster se observa que cierta cantidad de puntos están situados en la línea recta mientras que otros no, esto se evidencia claramente ya que sobresalen de la recta e incluso se constata lo especificado anteriormente de que posiblemente existen dos poblaciones distintas en los datos sobre la cantidad de memoria, especificado esto se puede concluir que probablemente no existe normalidad, por lo cual para verificar dicha aseveración se realizaron los respectivos test de normalidad a niveles de significancia primeramente de 0,05 que es la opción más común, luego de 0,1 que es poco significante y finalmente 0,01 que es de alta significancia

La Tabla 7.6, muestra todos los resultados obtenidos de cada prueba de normalidad (para un α=0,01, α=0,05 y α=0,1)

7.3.2.1 Paso 1A (α=0,01), 1B (α=0,05) y 1C (α=0,1): Planteamiento de Hipótesis

H_o: Ocupación de memoria ~ $N(\mu, \sigma^2)$
H_1: Ocupación de memoria $\nsim N(\mu, \sigma^2)$

7.3.2.2 Paso 2A (α=0,01), 2B (α=0,05) y 2C (α=0,1): Nivel de significancia

α=0,01, α=0,05 y α=0,1

7.3.2.3 Paso 3A (α=0,01), 3B (α=0,05) y 3C (α=0,1): Estadístico y valor p

La Tabla 7.6, muestra todos los resultados obtenidos de cada prueba de normalidad, se visualiza el valor del estadístico y el valor p para tomar su respectiva decisión de acuerdo con su nivel de significancia.

Tabla 7.6.- Resultados de las pruebas de normalidad, clúster y variable memoria: estadístico y valor p

	dendro_diana	dendro_ variables	hclust_vector	hrarchy	simlrty
Anderson-Darling normality test	A = 255,2, p-value <2,2e-16	A = 230,44, p-value <2,2e-16	A = 251,66, p-value <2,2e-16	A = 254,93, p-value <2,2e-16	A = 255,89, p-value <2,2e-16
p-values adjusted by the Holm method	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16
Lilliefors (Kolmogorov-Smirnov) normality test	D = 0,22031, p-value <2,2e-16	D = 0,18841, p-value <2,2e-16	D = 0,219, p-value <2,2e-16	D = 0,21354, p-value <2,2e-16	D = 0,21611, p-value <2,2e-16
p-values adjusted by the Holm method	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16
Cramer-von Mises normality test	W = 44,109, p-value = 7,37e-10	W = 40,257, p-value = 7,37e-10	W = 43,413, p-value = 7,37e-10	W = 43,958, p-value = 7,37e-10	W = 44,133, p-value = 7,37e-10
p-values adjusted by the Holm method	7,37e-10, 0,00000000368 5	7,37e-10, 0,00000000368 5	7,37e-10, 0,00000000368 5	7,37e-10, 0,00000000368 5	7,37e-10, 0,00000000368 5
Pearson chi-square normality test	P = 6722,1, p-value <2,2e-16	P = 4499,6, p-value <2,2e-16	P = 6690,1, p-value <2,2e-16	P = 7187,4, p-value <2,2e-16	P = 7185,7, p-value <2,2e-16
p-values adjusted by the Holm method	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16
Shapiro-Wilk normality test	W = 0,83078, p-value <2,2e-16	W = 0,84606, p-value <2,2e-16	W = 0,83247, p-value <2,2e-16	W = 0,82802, p-value <2,2e-16	W = 0,82739, p-value <2,2e-16
p-values adjusted by the Holm method	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16
Shapiro-Francia normality test	W = 0,83111, p-value <2,2e-16	W = 0,84641, p-value <2,2e-16	W = 0,83281, p-value <2,2e-16	W = 0,82848, p-value <2,2e-16	W = 0,82785, p-value <2,2e-16
p-values adjusted by the Holm method	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16	<2,22e-16, <2,22e-16

7.3.2.4 Paso 4A: Regla de decisión para α=0,01

Si el p valor es menor que 0,01 (p-value < 0,01) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla.

7.3.2.5 Paso 4B: Regla de decisión para α=0,05

Si el p valor es menor que 0,05 (p-value < 0,05) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla.

7.3.2.6 Paso 4C: Regla de decisión para α=0,1

Si el p valor es menor que 0,1 (p-value < 0,1) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla.

7.3.2.7 Paso 5A: Tabla de resultados α=0,01

Para un nivel de significancia de 0,01 se obtuvieron resultados similares a los obtenidos con los p-values de 0,05 y 0,1 en donde para las pruebas de Anderson Darling, Holm method, Lilliefors (Kolmogorov – Smirnov), Pearson chi-square, p-values adjusted by the Holm method, Shapiro Wilk y Shapiro Francia con un p-value igual a 2,2e-16 se rechaza la hipótesis nula dado que este valor es demasiado pequeño, concluyéndose que los datos no siguen una distribución normal, con respecto a los métodos Cramer Von Mises y Holm method el p-value igual a 7,37e-10 y 0,000000003685 respectivamente, tuvo una leve variación pero de igual manera llevan a la conclusión de que los datos no siguen una distribución debido a que los valores son demasiado pequeños siendo menores al nivel de significancia propuesto.

La Tabla 7.7, muestra los resultados de normalidad para α=0,01, para cada uno de los 12 métodos de normalidad utilizados.

Tabla 7.7.- Resultados de la normalización de la variable memoria para un valor de α=0,01

	dendro_diana	dendro_ variables	hclust_vector	hrarchy	simlrty
Anderson-Darling normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method	No son normales	No son normales	No son normales	No son normales	No son normales
Lilliefors (Kolmogorov-Smirnov) normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method	No son normales	No son normales	No son normales	No son normales	No son normales
Cramer-von Mises normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method	No son normales	No son normales	No son normales	No son normales	No son normales
Pearson chi-square normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method	No son normales	No son normales	No son normales	No son normales	No son normales
Shapiro-Wilk normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method	No son normales	No son normales	No son normales	No son normales	No son normales
Shapiro-Francia normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method	No son normales	No son normales	No son normales	No son normales	No son normales

7.3.2.8 Paso 5B: Tabla de resultados α=0,05

En los pasos anteriores se evidencia que el p-value obtenido para cada una de las pruebas de normalidad aplicadas a los datos sobre la cantidad de memoria es demasiado pequeño, por lo cual no se visualiza claramente la región pintada; para cada uno de estos

valores se obtuvo que caen en la zona de rechazo debido a que son menores al nivel de significancia de 0,05. De manera más detallada se tuvo 2,2e-16 para el estadístico de Anderson Darling en todos los métodos como son: dendro_diana, dendro_variables, hclust_vector, hrarchy y simlrty y se rechazó la hipótesis nula debido a que el p-value es muy pequeño concluyéndose que los datos de ocupación de memoria no siguen una distribución normal.

En el caso de Holm de igual manera se obtuvo un valor de 2,22e-16 para todos los métodos clúster en estudio, por lo cual también se rechaza la hipótesis nula dado que es menor al valor de significancia de 0,05, es decir, no existe normalidad en los datos de ocupación de memoria, lo mismo ocurre con la prueba de Kolmogorov y Smirnov con la corrección de Lilliefors la cual posee un p-value de 2,22e-16 para todos los métodos. Al analizar el estadístico Cramer-vonMises se evidenció un p-value igual a 7,37e-10 y dicho valor es inferior al nivel de significancia igual a 0,05 por lo cual se rechaza la hipótesis nula y se corrobora que se observa diferencia entre los datos de ocupación de memoria y la distribución normal. El método de Holm que es con el cual se ajustan los p-values dio como resultado un p-value final igual a 0,000000003685 para todos los métodos, dicho valor es menor a 0,05 por lo que se rechazó la hipótesis nula y se concluyó que los datos no siguen una distribución normal, con respecto a la Prueba chi cuadrada de Pearson, p-valores ajustados por el método de Holm, Shapiro Wilk y finalmente Shapiro Francia también se reincidió en su p-value de 2,2e-16 el cual es menor al nivel de significancia, por lo que al igual que en las pruebas de normalidad analizados anteriormente, también se rechazó la hipótesis nula y se concluye que se observa diferencia entre los datos de ocupación de memoria y la distribución normal.

La Tabla 7.8, muestra los resultados de normalidad para α=0,05.

Tabla 7.8.- Resultados de la normalización de la variable memoria para un valor de α=0,05

	dendro_diana	dendro_ variables	hclust_vector	hrarchy	simlrty
Anderson-Darling normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method	No son normales	No son normales	No son normales	No son normales	No son normales
Lilliefors (Kolmogorov-Smirnov) normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method	No son normales	No son normales	No son normales	No son normales	No son normales
Cramer-von Mises normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method	No son normales	No son normales	No son normales	No son normales	No son normales
Pearson chi-square normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method	No son normales	No son normales	No son normales	No son normales	No son normales
Shapiro-Wilk normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method:	No son normales	No son normales	No son normales	No son normales	No son normales
Shapiro-Francia normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method	No son normales	No son normales	No son normales	No son normales	No son normales

7.3.2.9 Paso 5C: Tabla de resultados α=0,1

Se analizan los p-values obtenidos en cada prueba de normalidad con un nivel de significancia de 0,1 concluyéndose que para las pruebas de Anderson Darling, Holm method, Lilliefors (Kolmogorov–Smirnov), Pearson chi-square, p-values adjusted by the Holm method, Shapiro Wilk y Shapiro Francia con un p-value igual a 2,2e-16 se rechaza la hipótesis nula dado que este valor es demasiado pequeño, concluyéndose que los datos

no siguen una distribución normal, con respecto a los métodos Cramer Von Mises y Holm method el p-value igual a 7,37e-10 y 0,000000003685 respectivamente, tuvo una leve variación pero de igual manera llevan a la conclusión de que los datos no siguen una distribución debido a que los valores son demasiado pequeños, siendo menores al nivel de significancia propuesto (Tabla 7.9).

Tabla 7.9.- Resultados de la normalización de la variable memoria para un valor de α=0,1

	dendro_diana	dendro_ variables	hclust_vector	hrarchy	simlrty
Anderson-Darling normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method	No son normales	No son normales	No son normales	No son normales	No son normales
Lilliefors (Kolmogorov-Smirnov) normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method	No son normales	No son normales	No son normales	No son normales	No son normales
Cramer-von Mises normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method	No son normales	No son normales	No son normales	No son normales	No son normales
Pearson chi-square normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method:	No son normales	No son normales	No son normales	No son normales	No son normales
Shapiro-Wilk normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method:	No son normales	No son normales	No son normales	No son normales	No son normales
Shapiro-Francia normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method	No son normales	No son normales	No son normales	No son normales	No son normales

La Tabla 7.7, Tabla 7.8 y Tabla 7.9, representan todos los resultados obtenidos sobre las pruebas de normalidad en donde se verificaron si los datos cumplen o no dicho supuesto de la variable memoria, corroborando así que para ninguna de las pruebas en estudio se cumplió dicha hipótesis ya que todos los p-values obtenidos resultaron menores a los niveles de significancia propuestos (0,01, 0,05 y 0,1).

7.3.3 Normalización

Se aplicaron métodos para la transformación a la normalidad de los datos en estudio usando la función en R llamada powerTransform, con el objetivo de determinar la potencia óptima a la que se debe elevar la variable de interés y así obtener normalidad en los datos, dicho método devuelve una matriz con columnas etiquetadas como "Est Power" para el valor de lambda que maximiza la probabilidad; "Pwr redondeado" para roundlam, y las columnas "Wald Lwr Bnd" y "Wald your Bnd" para un intervalo de confianza de la teoría normal de Wald del 95% para lambda calculado como la estimación más o menos 1,96 veces el error estándar*.* Para el primer caso en donde se usa la transformación "bcPower" que es el valor predeterminado para la familia de potencia Box-Cox se obtuvo en resumen lo de la Tabla 7.10.

Tabla 7.10.- Normalización por grupos

Método	p-valor	Normalización
bcPower	$2{,}22e^{-16}$	No son normales
yjPower	$2{,}22e^{-16}$	No son normales

El código utilizado en R y su salida se muestran a continuación (Tabla 7.11):

Tabla 7.11.- Normalización utilizando bcPower, clúster, complejidad espacial

NORMALITY TRANSFORMS BY GROUPS

summary(powerTransform(Y ~ X, family="bcPower"))

bcPower Transformation to Normality

	Est Power	Rounded Pwr Wald	Lwr Bnd Wald	Upr Bnd
Y1	1,8347	1,83	1,7725	1,8969

Likelihood ratio test that transformation parameter is equal to 0 (log transformation)

		LRT	df	pval
LR test,	lambda = (0)	3419	1	<2,22e-16

		LRT	df	pval
LRtest,	lambda = (1)	701,0601	1	<2,22e-16

El resumen de las transformaciones usando el método multivariado de Box-Cox permite visualizar los valores de $\hat{\lambda}$ en la columna "Est. Power" este valor es igual a 1,8347 y sugiere una transformación de $Y'=Y^2$ basándose en la Tabla 7.12 de posibles transformaciones para normalidad.

Tabla 7.12.- Posibles transformaciones para normalidad

Si ʎ =	-2	-1	-0,5	0	0,5	1	2
Transformación	$\frac{1}{Y^2}$	$\frac{1}{Y}$	$\frac{1}{Y^2}$	Log(Y)	$Y^{0,5}$	Ninguna	Y^2

En la Tabla 7.11, la potencia para Y1 no aparenta ser diferente de 1 dado que es igual a 1,8969, seguido se analizaron las pruebas de razón de verosimilitud donde se verificó que todas las potencias son cero, lo cual es firmemente rechazado ya que el valor aproximado $\chi^2(1)$ es bastante grande (3419). Finalmente, con respecto al p-value que se obtendrá realizando la transformación es igual a 2,22e-16 el cual es menor al nivel de significancia de 0,05 rechazándose la hipótesis nula de que los datos sigan una distribución normal.

Tabla 7.13.- Normalización utilizando yjPower, clúster, complejidad espacial

summary(powerTransform(Y ~ X, family="yjPower"))
yjPower Transformation to Normality

	Est Power	Rounded Pwr Wald	Lwr Bnd Wald	Upr Bnd
Y1	-3,0653	-3,07	-3,6453	-2,4852

Likelihood ratio test that transformation parameter is equal to 0

		LRT	df	pval
LR test,	lambda = (0)	1327,122	1	< 2,22e-16

En la Tabla 7.13, el resumen de las transformaciones usando el método de Yeo-Johnson permitió visualizar los valores de $\hat{\lambda}$ en la columna "Est. Power" el cual es igual a -3,0653 siendo un valor que se encuentra fuera del rango de las posibles transformaciones para normalidad, seguidamente se determinó que la potencia para Y1 aparenta ser diferente de 1 dado que es igual a -2,4852, se analizaron las pruebas de razón de verosimilitud donde se contrastó que todas las potencias son cero, lo cual es firmemente rechazado ya que el valor aproximado $\chi^2(0)$ es bastante grande (1327,122). Finalmente, con respecto al p-value que se obtendrá realizando la transformación es igual a 2,22e-16 el cual es menor al nivel de significancia de 0,05 rechazándose la hipótesis nula de que los datos sigan una distribución normal.

Se evidenció que a los datos sobre la variable memoria no se les puede realizar una normalización ya que los p-values obtenidos son demasiado pequeños.

7.3.4 Homocedasticidad

Para la prueba de homocedasticidad (homogeneidad de varianzas) se consideró el siguiente planteamiento de hipótesis.

7.3.4.1 Test de Levene

Paso 1: Planteamiento de hipótesis

H_o: $\sigma^2_{dendro\_diana} = \sigma^2_{dendro\_variables} = \sigma^2_{hclust\_vector} = \sigma^2_{hrarchy} = \sigma^2_{simlrty}$

H_1:$\exists i, j \in \{dendro\_diana, dendro\_variables, hclust\_vector, hrarchy, simlrty\}\ tal\ \ que\ i \neq j,\ \sigma_i^2 \neq \sigma_j^2$

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

Group=17230; Df=4; Fvalue=0,1513; Pr(>F) < 0,9625

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

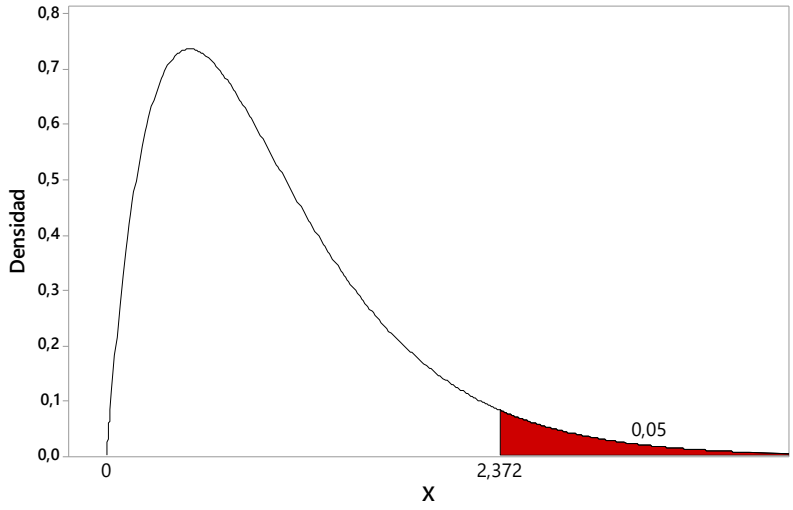


Figura 7.6.- Zonas de rechazo y aceptación para la homogeneidad de varianzas, complejidad espacial, clúster

Paso 5: Decisión

Se obtuvo un valor p igual a 0,9625 el cual es mayor al nivel de significancia de 0,05 por lo que no se rechazó la hipótesis nula (Ho) y se concluye que las varianzas de los grupos de memoria son iguales, los datos sobre memoria para cada uno de los métodos clúster son homocedásticos (Figura 7.6).

7.3.4.2 Test de Bartlett

Paso 1: Planteamiento de Hipótesis

H_o: $\sigma^2_{dendro\_diana} = \sigma^2_{dendro\_variables} = \sigma^2_{hclust\_vector} = \sigma^2_{hrarchy} = \sigma^2_{simlrty}$

H_1:$\exists i, j \in \{dendro\_diana, dendro\_variables, hclust\_vector, hrarchy, simlrty\}\ tal\ \ que\ i \neq j,\ \sigma_i^2 \neq \sigma_j^2$

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

Bartlett's K-squared=0,68473, df=4, p-value=0,9532

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5: Decisión

Mediante la prueba de Bartlett se obtuvo un valor p igual a 0,9532 el mismo que es mayor a un nivel de significancia de 0,05 por lo que no se rechaza la hipótesis nula (Ho) y se concluye que la varianza de los grupos en estudio no es diferente. Comparativamente entre las pruebas de Bartlett y Levene's se llega a la misma conclusión, aunque existe una leve variación entre el valor p de cada prueba.

7.3.5 Independencia

Test de Independencia utilizando la prueba chi cuadrado χ^2.

Paso 1: Planteamiento de Hipótesis

H_o: Técnicas clúster y memoria son independientes.

H_1: Técnicas clúster y memoria no son independientes.

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

Pearson's Chi-squared test, X-squared=147, df=12, p-value <2,2e-16

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

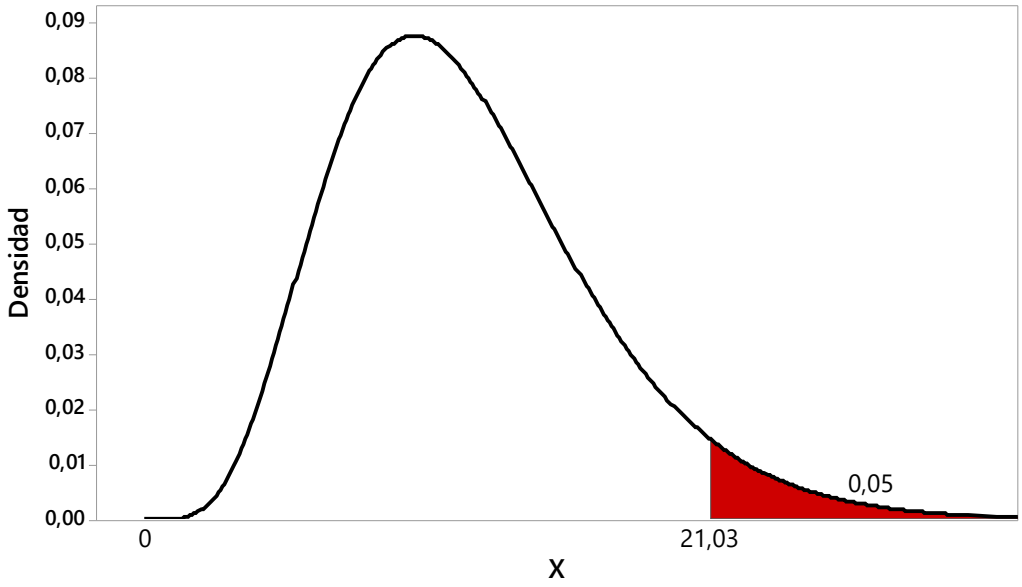


Figura 7.7.- Zonas de rechazo y aceptación para el prerrequisito de independencia

Paso 5: Decisión

El valor p obtenido en la prueba es igual a 2,2e-16 el cual es menor al nivel de significancia de 0,05 por lo que se rechaza la hipótesis nula (Ho) y se concluye que los métodos clúster y la memoria no son independientes (Figura 7.7).

7.3.6 Pruebas de hipótesis

Una vez analizados los prerrequisitos se concluyó que no se cumple normalidad ni independencia, ni tampoco se puede lograr la misma normalidad mediante las transformaciones realizadas, se debe realizar pruebas no paramétricas para muestras independientes. Para todas las pruebas de hipótesis se sigue el método de los 5 pasos propuesto por (Lind et al., 2012).

7.3.6.1 Medidas descriptivas específicas

Las medidas descriptivas reflejan los resultados sobre la cantidad de memoria que se empleó por cada método de asociación, obteniéndose así distintos parámetros de análisis para cada uno.

La Tabla 7.14, refleja los resultados sobre la cantidad de memoria que se empleó por cada técnica clúster, obteniéndose así distintos parámetros de análisis para cada uno.

Tabla 7.14.- Medidas descriptivas de las técnicas clúster

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
dendro_diana	104,0	141,5	247,0	223,0	278,0	382,0
dendro_variables	105,0	151,0	256,0	231,3	286,0	384,0
hclust_vector	103,0	141,5	247,0	223,0	278,0	381,0
hrarchy	105,0	144,0	249,0	224,8	280,0	319,0
simlrty	105,0	143,0	248,0	223,9	279,0	319,0

El método que usa en promedio mayor memoria es el dendro_variables con un valor igual a 231,3 con un mínimo de memoria igual a 105 y un máximo de 384, el primer cuartil muestra que el 25% de la memoria es menor o igual a 151 con una mediana que indica que la mitad de la memoria empleada para el método es menor o igual a 256 y la otra mitad es mayor o igual a 256, finalmente el tercer cuartil refleja que el 75% de la memoria es menor o igual a 286. El método que utiliza menos memoria es diana y hclust.vector, con respecto al método diana se determinó que posee un valor promedio igual a 223 con un mínimo de memoria igual a 104 y un máximo de 382, el primer cuartil muestra que el 25% de la memoria es menor o igual a 141,5 con una mediana que indica que la mitad de la memoria empleada para el método es menor o igual a 247 y la otra mitad es mayor o igual a 247, el tercer cuartil refleja que el 75% de la memoria es menor o igual a 278; en lo que compete al método hclust.vector se obtuvo un valor promedio igual a 223 con un mínimo de memoria igual a 103 y un máximo de 381, el primer cuartil muestra que el 25%

de la memoria es menor o igual a 141,5 con una mediana que indica que la mitad de la memoria empleada para el método es menor o igual a 247 y la otra mitad es mayor o igual a 247, el tercer cuartil refleja que el 75% de la memoria es menor o igual a 278, con este análisis se evidenció que no existe mayor diferencia entre estos métodos a excepción del máximo y el mínimo. Para concluir se determinó que para los métodos clúster en estudio no existe mayor diferencia en la memoria usada en promedio.

7.3.6.2 Kruskal Wallis H-test

La siguiente prueba no paramétrica permite comparar n medianas, en particular las 5 correspondientes a la memoria de los 5 métodos clúster. A continuación, se demuestra la hipótesis:

Paso 1: Planteamiento de Hipótesis

H_o: $\tilde{\mu}_{dendro\_diana} = \tilde{\mu}_{dendro\_variables} = \tilde{\mu}_{hclust\_vector} = \tilde{\mu}_{hrarchy} = \tilde{\mu}_{simlrty} = \tilde{\mu}_{memoria}$

H_1: $\tilde{\mu}_i \neq \tilde{\mu}_j \ para\ al\ menos\ un\ par\ de\ (i,j)$

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

Kruskal-Wallis chi-squared=138,51, df=4, p-value < 2,2e-16

Paso 4: Regla de Decisión

Si p-value < 0,05 entonces se rechaza Ho, caso contrario no se rechaza. También si el valor crítico es menor que el valor calculado se rechaza la hipótesis nula. En nuestro caso 9,488<138,51 por tanto, se rechaza la hipótesis nula.

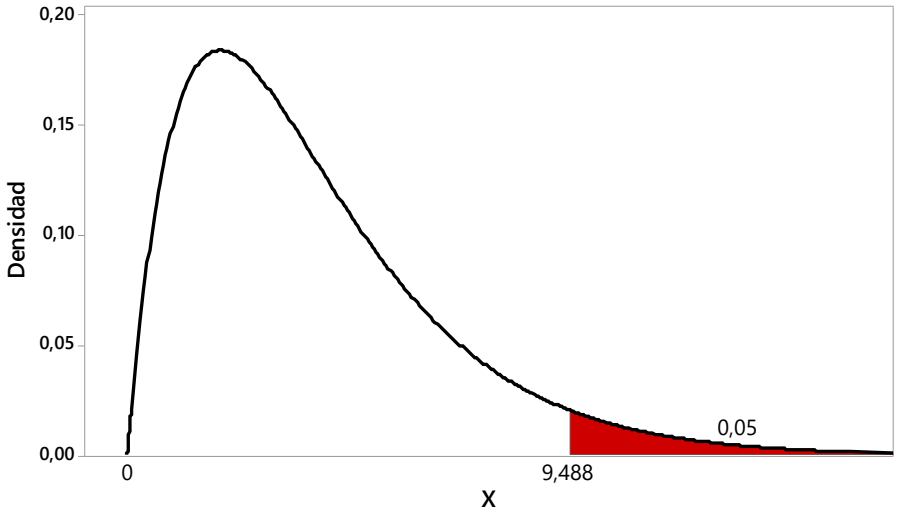


Figura 7.8.- Zonas de rechazo y aceptación para la prueba de Kruskal Wallis

Se obtuvo un valor p igual a 2,2e-16 el cual es menor a un nivel de significancia de 0,05 por lo que se rechaza la hipótesis nula y se concluye que al menos un par de los métodos clúster es diferente (Figura 7.8).

Paso 5A: Decisión

Dado que se rechaza la hipótesis nula, significa que al menos existe un par de rangos diferentes o que no todos los pares de rangos son iguales.

Para determinar cuál par de rangos son diferentes se utilizó la prueba no paramétrica para comparación de pares de dos muestras independientes de Mann Whitney Wilcoxon U-test.

7.3.6.3 Mann Whitney Wilcoxon U-test

Se utilizó Mann Whitney Wilcoxon U-test para la comparación por parejas independientes como se ve en la página oficial de R (*R: Pairwise Wilcoxon Rank Sum Tests*, 2021). Para no provocar que el error de Tipo 1 aumente, los métodos usados mediante R fueron la

corrección de Bonferroni en la cual los valores p se multiplican por el número de comparaciones y Holm que realizó correcciones menos conservadoras, los resultados obtenidos fueron los siguientes:

Paso 1: Planteamiento de Hipótesis

H_o: $No\ hay\ diferencia\ entre\ las\ memorias\ de\ las\ 2\ poblaciones\ de\ técnicas\ cluster$

H_1: $Hay\ diferencia\ entre\ las\ memorias\ de\ las\ 2\ poblaciones\ de\ técnicas\ cluster$

Paso 2: Nivel de significancia α=0,05 (dividido para el número de comparaciones)

Paso 3A: Estadístico de Prueba

Tabla 7.15.- Comparaciones múltiples Wilcoxon (Bonferroni)

	dendro_diana	dendro_variables	hclust_vector	hrarchy
dendro_variables	< 2e-16	-	-	-
hclust_vector	1,000	< 2e-16	-	-
Hrarchy	0,031	5,2e-13	0,055	-
Simlrty	1,000	< 2e-16	1,000	1,000

Se muestran los resultados obtenidos al aplicar la prueba no paramétrica Wilcoxon mediante la corrección de Bonferroni (Tabla 7.15).

Paso 4A: Regla de Decisión

Si p-value < 0,05 entonces se rechaza Ho, caso contrario no se rechaza.

Paso 5A: Decisión

Se calcularon las comparaciones por pares entre niveles de grupo con correcciones para pruebas múltiples usando el método de Bonferroni, en donde se obtuvo que los pares hrarchy – dendro_variables, hrarchy – hclust_vector, simlrty – hclust_vector son significativamente diferentes debido a que el valor p obtenido para cada par es 0,031, 5,2e-13, 2e-16 respectivamente, los cuales son menores a un nivel de significancia de 0,05.

Paso 3B: Comparaciones múltiples, estadístico de prueba de Wilcoxon (Holm)

Se muestran los resultados obtenidos al aplicar la prueba no paramétrica Wilcoxon, en este caso se usó el método de Holm (Tabla 7.16).

Tabla 7.16.- Comparaciones múltiples Wilcoxon (Holm)

	dendro_diana	dendro_variables	hclust_vector	hrarchy
dendro_variables	< 2e-16	-	-	-
hclust_vector	0,881	< 2e-16	-	-
Hrarchy	0,019	3,6e-13	0,027	-
Simlrty	0,418	< 2e-16	0,434	0,434

Paso 4B: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5B: Decisión

Mediante el método de Holm se obtuvo que los pares hrarchy-dendro_variables, simlrty-dendro_variables, hrarchy- dendro_variables, simlrty- hclust_vector y simlrty-hrarchy son significativamente diferentes debido a que su p-value de 0,019, 3,6e-13, 2e-16 y 0,434 respectivamente para cada par de método es menor al nivel de significancia de 0,05 por lo cual se rechazó la hipótesis nula. Para el par hclust_vector–dendro_variables se evidenció un p-value de 0,881 el cual es mayor al nivel de significancia de 0,05 por lo que no se rechazó la hipótesis nula y se concluye que son significativamente iguales.

7.3.6.4 ANOVA no Paramétrico

Utilizamos el paquete Rfit (Rank-Based Estimation for Linear Models) que proporciona funciones para análisis basados en rangos de modelos lineales, estimación basada en rangos, la inferencia ofrece una alternativa robusta a los mínimos cuadrados (Kloke y McKean, 2020).

Paso 1: Planteamiento de Hipótesis

$\mathbf{H_o}$: $No\ hay\ diferencia\ entre\ las\ memorias\ de\ las\ 5\ poblaciones\ de\ técnicas\ cluster$

$\mathbf{H_1}$: $Hay\ diferencia\ entre\ las\ memorias\ de\ las\ 5\ poblaciones\ de\ técnicas\ cluster$

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

F-Statistic=54,888; p-value=0,000

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5: Decisión

El valor p obtenido en el ANOVA es igual a 0 el cual es menor a un nivel de significancia de 0,05 por lo que se rechazó la hipótesis nula y se concluye que los rangos entre los métodos clúster y la memoria son distintas.

Paso 3B: Estadístico de Prueba

Tabla 7.17.- Comparaciones múltiples ANOVA no paramétrico

	dendro_variables	hclust_vector	hrarchy	simlrty
dendro_variables	< 2e-16	-	-	-
hclust_vector	0,999	< 2e-16	-	-
Hrarchy	9,6e-05	< 2e-16	9,6e-05	-
Simlrty	0,051	< 2e-16	0,051	0,051

Se muestran los resultados ANOVA, se observa que existen pares de métodos clúster cuya diferencia es significativa (Tabla 7.17)

Paso 4B: Regla de Decisión

Si el p-valor es menor que 0,05 se rechaza Ho, caso contrario no (Figura 7.9).

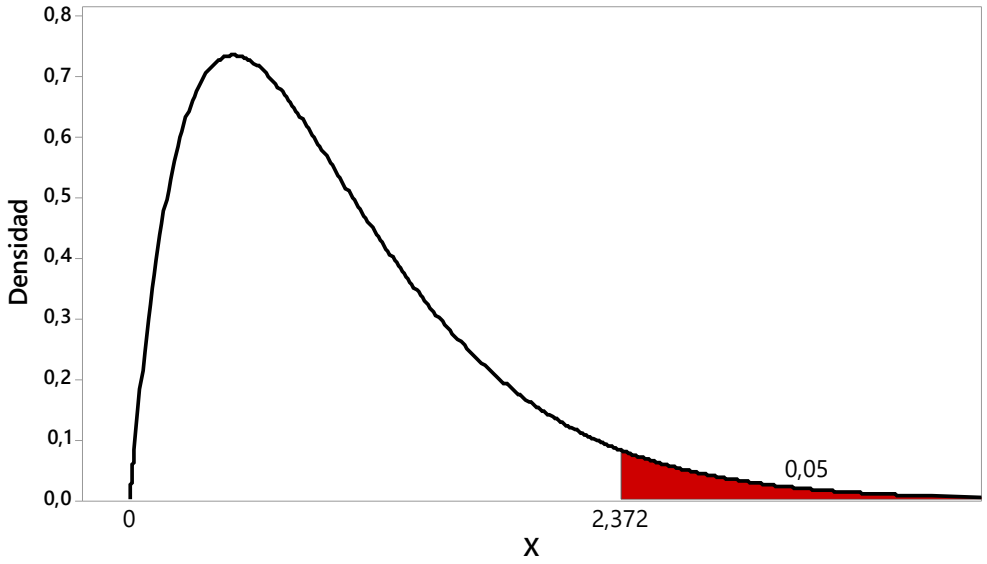


Figura 7.9.- Zonas de rechazo y aceptación ANOVA no paramétrico, complejidad espacial, clúster: F, df1=4, df2=17230, α=0,05

Paso 5B: Decisión

Los resultados obtenidos mediante el método Rfit para cada uno de los pares de métodos clúster en estudio fueron 9,6e-0,5, 2e-16, 2e-16 para hrarchy - dendro_variables, hrarchy – hclust_vector y simlrty – hclust_vector respectivamente, los mismos resultan menores al nivel de significancia de 0,05 por lo que se rechaza la hipótesis nula y se concluye que existen diferencias significativas entre estos métodos clúster, mientras que para los pares hclust_vector – dendro_variables, simlrty – dendro_variables y simlrty-hrarchy el valor p obtenido es 0,999, 0,051 y 0,051 respectivamente, estos valores son mayores al nivel de significancia de 0,05 por lo que no se rechazó la hipótesis nula y se concluye que no existen diferencias significativas entre dichos pares de métodos clúster.

Tukey

Mediante la opción de Tukey se determinaron las técnicas clúster que son significativamente diferentes con respecto a la memoria que usa cada una, para lo cual se plantearon las siguientes hipótesis.

Paso 1: Planteamiento de Hipótesis

$\mathbf{H_o}$: $No\ hay\ diferencia\ entre\ las\ memorias\ de\ las\ 2\ poblaciones\ de\ técnicas\ cluster$

$\mathbf{H_1}$: $Hay\ diferencia\ entre\ las\ memorias\ de\ las\ 2\ poblaciones\ de\ técnicas\ cluster$

Paso 2: Nivel de significancia α=0,05

Paso 3A: Estadístico de Prueba

La Tabla 7.18, muestra las comparaciones múltiples entre las diferentes técnicas clúster.

Tabla 7.18.- Comparaciones múltiples (Tukey)

	I	J	Estimate	St Err	Lower Bound CI	Upper Bound CI
1	dendro_diana	dendro_variables	8,0002	0,51257	6,60188	9,39852
2	dendro_diana	hclust_vector	0,0004	0,51257	-1,39792	1,39873
3	dendro_diana	hrarchy	2,00016	0,51257	0,60184	3,39848
4	dendro_diana	simlrty	1,00054	0,51257	-0,39778	2,39886
5	dendro_variables	hclust_vector	7,99979	0,51257	6,60147	9,39812
6	dendro_variables	hrarchy	6,00003	0,51257	4,60171	7,39836
7	dendro_variables	simlrty	6,99966	0,51257	5,60134	8,39798
8	hclust_vector	hrarchy	-1,99976	0,51257	-3,39808	-0,60144
9	hclust_vector	simlrty	-1,00013	0,51257	-2,39845	0,39819
10	hrarchy	simlrty	0,99963	0,51257	-0,3987	2,39795

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5: Decisión

Los resultados obtenidos indican que cada uno de los métodos en estudio no son significativos con respecto a la cantidad de memoria que usan, dado que se puede visualizar esta aseveración en los intervalos de confianza que están dados por valores negativos y positivos.

7.4 Estudio de la complejidad temporal

A continuación, se realiza el estudio de complejidad temporal respecto a las técnicas clúster. Primeramente, se realiza un estudio descriptivo, para luego realizar las pruebas de hipótesis, no sin antes realizar la prueba de los supuestos de normalidad, homocedasticidad e independencia para determinar el tipo de prueba a utilizar.

7.4.1 Medidas descriptivas

Las siguientes tablas resumen los resultados descriptivos del tiempo de ejecución por técnicas clúster utilizadas en el análisis. La Tabla 7.19, presenta las principales medidas de centralización, dispersión y forma.

Tabla 7.19.- Medidas descriptivas de cantidad de tiempo por técnicas clúster

TÉCNICAS CLÚSTER	mean	sd	IQR	cv	skewness	kurtosis
dendro_diana	0,3064022	0,5833718	0,1549556	1,903941	4,923436	28,345794
dendro_variables	44,2447127	56,4054589	57,4960819	1,274852	2,12315	5,651467
hclust_vector	0,2427268	0,8792874	0,0974375	3,62254	15,628728	363,12514
hrarchy	1,8069068	1,6209652	1,6909709	0,897094	2,506811	15,439837
simlrty	1,8419862	1,5502579	1,711646	0,841623	1,510456	2,26735

La Tabla 7.20, presenta los cuartiles incluida la mediana (50%), que dividen al grupo de datos en cuatro partes iguales.

Tabla 7.20.- Cuartiles de cantidad de tiempo por método clúster

TÉCNICAS CLÚSTER	0%	25%	50%	75%	100%	TIME:n
dendro_diana	0,024014612	0,09157026	0,14567665	0,2465258	6,113469	3447
dendro_variables	0,022502676	4,4838224	23,26483511	61,9799043	391,128655	3447
hclust_vector	0,007086716	0,05040263	0,08448613	0,1478401	24,25059	3447
hrarchy	0,094794627	0,69197164	1,28748489	2,3829426	23,609375	3447
simlrty	0,099339644	0,72206483	1,34384967	2,4337108	10,521396	3447

La Tabla 7.19 y La Tabla 7.20, presenta los cuartiles incluida la mediana (50%), que dividen al grupo de datos en cuatro partes iguales.

Tabla 7.20, reflejan los resultados sobre la cantidad de tiempo que se empleó por cada técnica clúster, llegándose a obtener así distintos parámetros de análisis para cada uno. La primera columna refleja el valor de la media en donde se evidenció que en promedio el

método que ocupa menos tiempo es hclust_vector con un valor de 0,2427268 y el método que en promedio usa mayor tiempo es el dendro_variables con un valor igual a 44,2447127, con relación a que tan dispersos se encuentran dichos datos analizados con respecto al valor promedio se obtuvo que el método que presenta menor dispersión es el dendro_diana con un valor igual a 0,5833718 y el método con mayor dispersión es el dendro_variables con un valor de 56,4054589. El rango intercuartil (IQR) da a notar que el método que presenta menor variación en el tiempo es el hclust_vector con un valor igual a 0,0974375 mientras que el método con mayor variación en su tiempo es el dendro_variables con un valor igual a 57,4960819; se nota que el método que posee menor coeficiente de variación (cv) es el hrarchy con un valor de 0,841623% de dispersión respecto a la media, mientras que el método con mayor coeficiente de variación es el hclust_vector con un 3,62254%. La asimetría de los datos (*skewness*) permitió notar que todos las técnicas clúster poseen una distribución simétrica, es decir que existe aproximadamente la misma cantidad de datos a los dos lados del valor de la media correspondiente, en cuanto a qué técnica clúster posee mayor asimetría es el hclust_vector con 15,628728 y con menor asimetría es el simlrty con 1,510456. El coeficiente de *kurtosis* obtenido para cada uno de los métodos dendro_diana, dendro_variables, hclust_vector, hrarchy y simlrty muestra que la distribución que siguen es leptocúrtica dado que dichos valores son positivos, lo que quiere decir que hay una mayor concentración de los datos en torno a la media, se obtuvo que la técnica simlrty posee menor coeficiente con un valor igual a 2,26735 y el hclust_vector con un coeficiente mayor, igual a 363,12514.

Finalmente se obtuvieron los cuartiles para cada técnica de regla de asociación, el primer cuartil con menor valor es para la técnica hclust_vector la cual muestra que el 25% del tiempo es menor o igual a 0,05040263 con una mediana la cual indica que la mitad del tiempo es menor o igual a 0,08448613 y la otra mitad es mayor o igual a 0,08448613, el tercer cuartil indica que el 75% de los datos es menor o igual a 0,1478401, la técnica con mayor valor es dendro_variables la cual muestra que el 25% del tiempo es menor o igual a 4,4838224 con una mediana la cual indica que la mitad del tiempo es menor o igual a 23,26483511 y la otra mitad es mayor o igual a 23,26483511, el tercer cuartil indica que el 75% de los datos es menor o igual a 61,9799043.

Mediante el gráfico BoxPlot se confirma que cada una de las técnicas analizadas muestran homogeneidad en su dispersión con una asimetría negativa o sesgada a la izquierda, debido a que la parte más larga de cada caja es la inferior a la mediana. Para ninguno de las técnicas clúster existen valores atípicos (Figura 7.10).

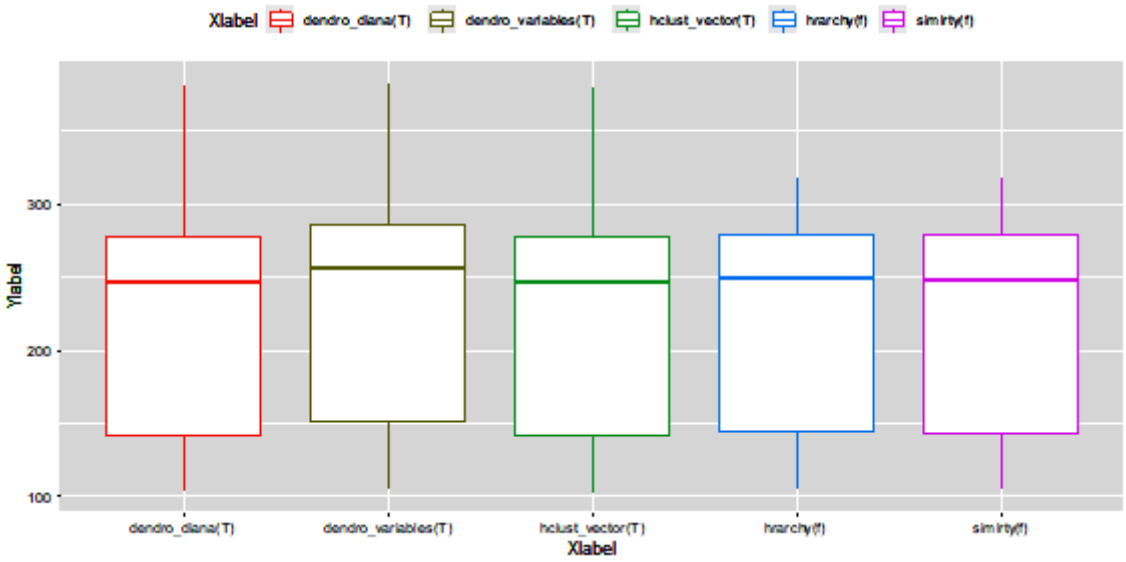


Figura 7.10.- Gráfico BoxPlot sobre el tiempo por técnica clúster

La Tabla 7.21, refleja los resultados sobre la cantidad de tiempo que se empleó por cada técnica clúster, llegándose a obtener distintos parámetros de análisis para cada uno.

Tabla 7.21.- Tiempo en las distintas técnicas clúster

TÉCNICAS CLÚSTER	mean	sd	IQR	cv	skewness	kurtosis	TIME:n
dendro_diana	0,3064022	0,5833718	0,1549556	1,903941	4,923436	28,345794	3447
dendro_variables	44,2447127	56,4054589	57,4960819	1,274852	2,12315	5,651467	3447
hclust_vector	0,2427268	0,8792874	0,0974375	3,62254	15,628728	363,12514	3447
hrarchy	1,8069068	1,6209652	1,6909709	0,897094	2,506811	15,439837	3447
simlrty	1,8419862	1,5502579	1,711646	0,841623	1,510456	2,26735	3447

La primera columna refleja el valor de la media en donde se evidenció que en promedio la técnica que ocupa menos tiempo es hclust_vector con un valor de 0,2427268 y la técnica que en promedio usa mayor tiempo es dendro_variables con un valor igual a 44,2447127, con relación a que tan dispersos se encuentran dichos datos analizados con respecto al valor promedio se obtuvo que la técnica que presenta menor dispersión es dendro_diana

con un valor igual a 0,5833718 y la técnica con mayor dispersión es dendro_variables con un valor de 56,4054589. El rango intercuartil (IQR) da a notar que la técnica que presenta menor variación en el tiempo es hclust_vector con un valor igual a 0,0974375 mientras que la técnica con mayor variación en su tiempo es dendro_variables con un valor igual a 57,4960819 ; se nota que la técnica que posee menor coeficiente de variación (cv) es simlrty con un valor de 0,841623% de dispersión respecto a la media, mientras que la técnica con mayor coeficiente de variación es hclust_vector con 3,62254%. La asimetría de los datos (*skewness*) permitió notar que todos las técnicas clúster poseen una distribución simétrica, es decir que existe aproximadamente la misma cantidad de datos a los dos lados del valor de la media correspondiente, en cuanto a qué técnica clúster posee mayor asimetría es hclust_vector con 15,628728 y con menor asimetría es simlrty con 1,510456. El coeficiente de *kurtosis* obtenido para cada una de las técnicas dendro_diana, dendro_variables, hclust_vector, hrarchy y simlrty muestra que la distribución que siguen es leptocúrtica, dado que dichos valores son positivos lo que quiere decir que hay una mayor concentración de los datos en torno a la media, se obtuvo que la técnica simlrty posee menor coeficiente con un valor igual a 2,26735 y hclust.vector con un coeficiente mayor, igual a 363,12514. Todas las técnicas de reglas de asociación se trabajaron con una muestra de 3447 datos.

7.4.2 Normalidad

Para determinar la prueba apropiada a utilizar en la hipótesis se procedió a la comprobación de los supuestos.

La Figura 7.11, de aproximación normal de los datos del tiempo de procesamiento en las técnicas clúster evidencia que el histograma es asimétrico hacia la derecha lo que quiere decir que los datos muestran un sesgo positivo, con respecto a la línea de distribución ajustada.

Se nota que las barras no la siguen por lo que no parece ofrecer un ajuste adecuado para una distribución normal, hay mayor concentración de datos en la cola derecha y escasamente se visualizan las barras ya que a partir del 100 son demasiado pequeñas.

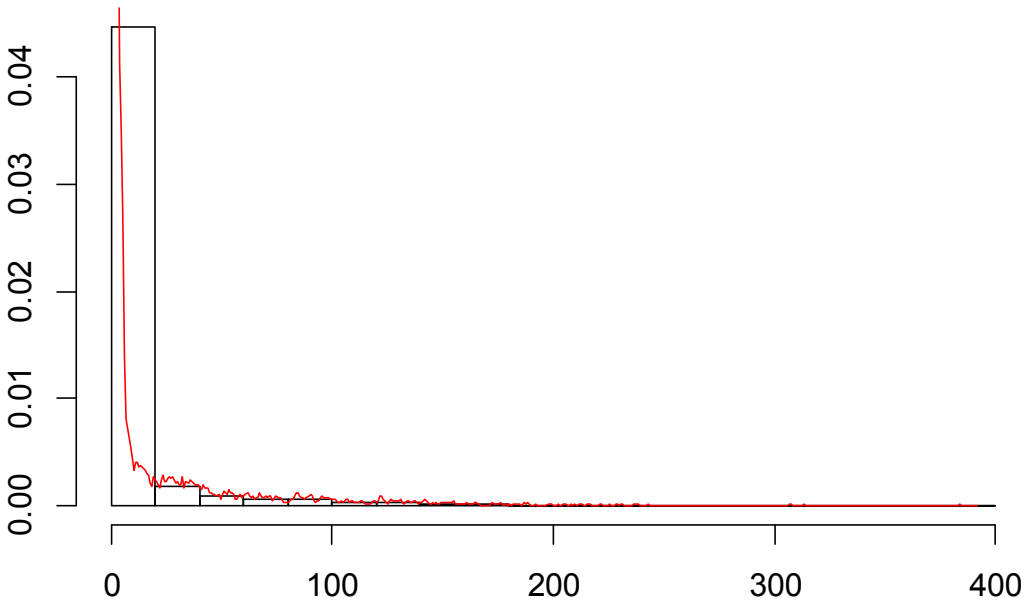


Figura 7.11.- Aproximación normal de los datos de tiempo

Al analizar la cercanía de la recta a la curva (Figura 7.12) sobre el tiempo de procesamiento usado por cada una de las técnicas clúster se corroboró que cierta cantidad de puntos se ubican en la línea recta mientras que otra cierta cantidad no, llevándonos a intuir que probablemente los datos no siguen una distribución normal.

Para verificar formalmente la aseveración de que los datos no se han extraído de una población normal, se realizaron los respectivos test de normalidad a niveles de significancia de 0,01, 0,05 y 0,1.

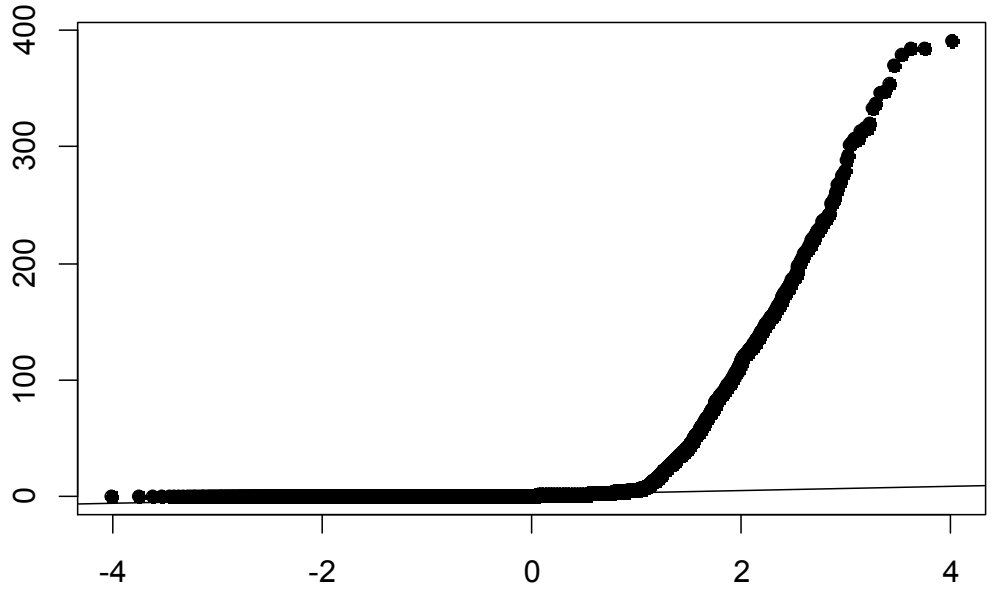


Figura 7.12.- Gráfico QQ para los datos de tiempo

La Tabla 7.22, muestra los resultados de distintas pruebas de normalidad.

7.4.2.1 Paso 1A (α=0,01), 1B (α=0,05) y 1C (α=0,1): Planteamiento de Hipótesis

H_o: Tiempo de procesamiento ~ N(μ, σ^2)

H_1: Tiempo de procesamiento $\nsim$N(μ, σ^2)

7.4.2.2 Paso 2A (α=0,01), 1B (α=0,05) y 1C (α=0,1): Nivel de significancia

α=0,01, α=0,05 y α=0,1

7.4.2.3 Paso 3A (α=0,01), 3B (α=0,05) y 3C (α=0,1): Estadístico y valor p

La Tabla 7.22, muestra todos los resultados obtenidos de cada prueba de normalidad, se visualiza el valor del estadístico y el valor p para tomar su respectiva decisión de acuerdo con su nivel de significancia.

Tabla 7.22.- Resultados de las pruebas de normalidad, clúster y variable tiempo: estadístico y valor p.

	dendro_diana	dendro_variables	hclust_vector	hrarchy	simlrty
Anderson-Darling normality test	A = 701,48, p-value < 2,2e-16	A = 251,65, p-value < 2,2e-16	A = 926,65, p-value < 2,2e-16	A = 171,67, p-value < 2,2e-16	A = 171,67, p-value < 2,2e-16
Holm technique	< 2,22e-16 < 2,22e-16	< 2,22e-16 < 2,22e-16	< 2,22e-16 < 2,22e-16	< 2,22e-16 < 2,22e-16	< 2,22e-16 < 2,22e-16
Lilliefors (Kolmogorov-Smirnov) normality test	D = 0,31969, p-value < 2,2e-16	D = 0,21652, p-value < 2,2e-16	D = 0,39435, p-value < 2,2e-16	D = 0,15308, p-value < 2,2e-16	D = 0,14697, p-value < 2,2e-16
Cramer-von Mises normality test	W = 140,73, p-value = 7,37e-10	W = 45,609, p-value = 7,37e-10	W = 195,66, p-value = 7,37e-10	W = 30,533, p-value = 7,37e-10	W = 30,533, p-value = 7,37e-10
p-values adjusted by the Holm technique:	7,37e-10 3,685e-09	7,37e-10 3,685e-09	7,37e-10 3,685e-09	7,37e-10 3,685e-09	7,37e-10 3,685e-09
Pearson chi-square normality test	P = 13431, p-value < 2,2e-16	P = 6996,9, p-value < 2,2e-16	P = 32361, p-value < 2,2e-16	P = 2324,3, p-value < 2,2e-16	P = 2258.8, p-value < 2,2e-16
p-values adjusted by the Holm technique	2,22e-16 < 2,22e-16	2,22e-16 < 2,22e-16	2,22e-16 < 2,22e-16	2,22e-16 < 2,22e-16	2,22e-16 < 2,22e-16
Shapiro-Wilk normality test	W = 0,40326, p-value < 2,2e-16	W = 0,75576, p-value < 2,2e-16	W = 0,19396, p-value < 2,2e-16	W = 0,79827, p-value < 2,2e-16	W = 0,84837, p-value < 2,2e-16
p-values adjusted by the Holm technique	2,22e-16 < 2,22e-16	2,22e-16 < 2,22e-16	2,22e-16 < 2,22e-16	2,22e-16 < 2,22e-16	2,22e-16 < 2,22e-16
Shapiro-Francia normality test	2,22e-16 < 2,22e-16	W = 0,75568, p-value < 2,2e-16	W = 0,19234, p-value < 2,2e-16	W = 0,79725, p-value < 2,2e-16	W = 0,84837, p-value < 2,2e-16
p-values adjusted by the Holm technique	2,22e-16 < 2,22e-16	2,22e-16 < 2,22e-16	2,22e-16 < 2,22e-16	2,22e-16 < 2,22e-16	2,22e-16 < 2,22e-16

7.4.2.4 Paso 4A: Regla de decisión para α=0,01

Si el p valor es menor que 0,01 (p-value < 0,01) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla.

7.4.2.5 Paso 4B: Regla de decisión para α=0,05

Si el p valor es menor que 0,05 (p-value < 0,05) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla.

7.4.2.6 Paso 4C: Regla de decisión para α=0,1

Si el p valor es menor que 0,1 (p-value < 0,1) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla.

7.4.2.7 Paso 5A: Tabla de resultados α=0,01

La Tabla 7.23, muestra los resultados de normalidad para α=0,01, para cada una de las 12 pruebas de normalidad utilizadas.

Tabla 7.23.- Resultados de la normalización de la variable tiempo para un valor de α=0,01

	dendro_diana	dendro_ variables	hclust_vector	hrarchy	simlrty
Anderson-Darling	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales
Lilliefors (Kolmogorov-Smirnov)	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm	No son normales	No son normales	No son normales	No son normales	No son normales
Cramer-von Mises	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm	No son normales	No son normales	No son normales	No son normales	No son normales
Pearson chi-square	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales
Shapiro-Wilk	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales
Shapiro-Francia	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales

7.4.2.8 Paso 5B: Tabla de resultados α=0,05

Al analizar las gráficas de distribución de probabilidad obtenidas para la variable tiempo en los pasos 4A, 4B y 4C se nota claramente que los p-values obtenidos en las distintas pruebas de normalidad son pequeños, por lo cual caen la zona de rechazo para los distintos niveles de significancia (α=0,01, α=0,05 y α=0,1).

Detalladamente se obtuvo que para el estadístico Anderson Darling el p-value es 2,2e-16 el cual es menor a los niveles de significancia propuestos (0,01, 0,05 y 0,1); rechazándose la hipótesis nula de que la variable tiempo siga una distribución normal. En el caso de Holm se nota un p-value de 2,22e-16 para todas las técnicas clúster en estudio, por lo cual también se rechaza la hipótesis nula dado que es menor a los niveles de significancia concluyéndose que no existe normalidad en los datos de la variable tiempo, esta situación se repite con la prueba de Kolmogorov Smirnov con la corrección de Lilliefors dado un p-value de 2,2e-16 para todas las técnicas clúster. Al analizar el estadístico Cramer-von Mises se evidenció como resultado un p-value igual a 7,37e-10, este valor es inferior a los niveles de significancia (0,05, 0,01 y 0,1) por lo cual se rechaza la hipótesis nula concluyéndose que se observa diferencia entre los datos de la variable tiempo y la distribución normal. La técnica de Holm que es con el cual se ajustan los p-values dio como resultado un p-value de 3,685e-09 para todas las técnicas clúster, este valor también es menor a los niveles de significancia propuestos por lo que se rechazó la hipótesis nula y se concluyó que los datos no siguen una distribución normal, con relación a la Prueba chi cuadrada de Pearson, p-valores ajustados por la técnica de Holm, Shapiro Wilk y finalmente Shapiro Francia también se obtuvo el mismo p-value de 2,2e-16 siendo este valor menor a los niveles de significancia planteados, por lo que se rechazó la hipótesis nula y se concluyó que se observa diferencia entre los datos de la variable tiempo y la distribución normal.

La Tabla 7.24, muestra los resultados de normalidad para α=0,05.

Tabla 7.24.- Resultados de la normalización de la variable tiempo para un valor de α=0,05

	dendro_diana	dendro_ variables	hclust_vector	hrarchy	simlrty
Anderson-Darling normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales
Lilliefors (Kolmogorov-Smirnov) normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales
Cramer-von Mises normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales
Pearson chi-square normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales
Shapiro-Wilk normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales
Shapiro-Francia normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales

7.4.2.9 Paso 5C: Tabla de resultados α=0,1

La Tabla 7.25, muestra los resultados de normalidad para α=0,1, para cada una de las 12 pruebas de normalidad utilizadas.

Tabla 7.25.- Resultados de la normalización de la variable tiempo para un valor de α=0,1

	dendro_diana	dendro_ variables	hclust_vector	hrarchy	simlrty
Anderson-Darling normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales
Lilliefors (Kolmogorov-Smirnov) normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales
Cramer-von Mises normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales
Pearson chi-square normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales
Shapiro-Wilk normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales
Shapiro-Francia normality test	No son normales	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales	No son normales

La Tabla 7.23, Tabla 7.24 y Tabla 7.25, representan todos los resultados obtenidos sobre las pruebas de normalidad en donde se verificaba si los datos cumplen o no el supuesto de normalidad, obteniéndose así que para ninguna de las pruebas en estudio se cumplió con dicho supuesto ya que todos los p-values obtenidos resultaron menores a los niveles de significancia de 0,01, 0,05 y 0,1.

7.4.3 Normalización

En busca de la transformación de los datos de la variable tiempo a una distribución normal se usó la función en R llamada powerTransform, con el objetivo de determinar la potencia óptima a la que se debe elevar la variable de interés y así buscar obtener normalidad en los datos.

El código utilizado en R y su salida se muestran a continuación:

Tabla 7.26.- Código R para normalización utilizando bcPower, clúster, complejidad temporal

```
# NORMALITY TRANSFORMS BY GROUPS
summary(powerTransform(Y ~ X, family="bcPower"))
bcPower Transformation to Normality
```

	Est Power	Rounded Pwr Wald	Lwr Bnd Wald	Upr Bnd
Y1	-0,1206	-0,12	-0,1264	-0,1147

Likelihood ratio test that transformation parameter is equal to 0 (log transformation)

		LRT	df	pval
LR test	lambda = (0)	1699,008	1	< 2,22e-16

Likelihood ratio test that no transformation is needed

		LRT	df	pval
LR test	lambda = (1)	113746,9	1	< 2,22e-16

Basándose en la Tabla 7.26 de posibles transformaciones para normalidad se tiene que el valor obtenido en la columna “Est. Power” sugiere una transformación de $Y' =$ Log(Y) el cual es igual a -0,1206.

La potencia para Y1 no aparenta ser diferente de 0 dado que es igual a -0,1147, seguido se analizaron las pruebas de razón de verosimilitud donde se verificó que todas las potencias son cero, lo cual es firmemente rechazado ya que el valor aproximado $\chi^2(1)$ es bastante grande (1699,008), con la segunda razón de verosimilitud igual a 1 se corrobora que no es necesario realizar transformaciones ($\lambda = 1$). Finalmente, con respecto al p-value que se obtendrá realizando la transformación es igual a 2,22e-16 este valor es menor al

nivel de significancia de 0,05 rechazándose la hipótesis nula de que los datos sigan una distribución normal.

Tabla 7.27.- Código normalización utilizando yjPower, clúster, complejidad temporal

summary(powerTransform(Y ~ X, family="yjPower"))

yjPower Transformation to Normality

	Est Power	Rounded Pwr Wald	Lwr Bnd Wald	Upr Bnd
Y1	-0,7727	-0,77	-0,7891	-0,7563

Likelihood ratio test that transformation parameter is equal to 0

		LRT	df	pval
LR test	lambda = (0)	14282,11	1	< 2,22e-16

En la Tabla 7.27, se observa que de la técnica de Yeo-Johnson se obtuvo para “Est. Power” un valor de la población diferente de cero analizando los intervalos de confianza (-0,7891; -0,7563), el cual se encuentra fuera del rango de las posibles transformaciones para normalidad, seguidamente se determinó que la potencia para Y1 aparenta ser diferente de 0 dado que es igual a -0,7727, seguido se analizaron las pruebas de razón de verosimilitud donde se contrastó que todas las potencias son cero, lo cual es firmemente rechazado ya que el valor aproximado $\chi^2(0)$ es bastante grande (14282,11). Finalmente, con respecto al p-value que se obtendría al realizar una transformación la cual no es recomendable, se tiene que este valor es 2,22e-16 el cual es menor al nivel de significancia de 0,05 rechazándose la hipótesis nula de que los datos sigan una distribución normal.

Se evidencia que a los datos sobre la variable tiempo de procesamiento no se les puede realizar una normalización ya que los p-values obtenidos son demasiado pequeños corroborándose esta aseveración tanto en el método BcPower y YjPower.

7.4.4 Homocedasticidad

7.4.4.1 Test de Levene

Paso 1: Planteamiento de Hipótesis

H₀: $\sigma^2_{dendro\_diana} = \sigma^2_{dendro\_variables} = \sigma^2_{hclust\_vector} = \sigma^2_{hrarchy} = \sigma^2_{simlrty}$

H₁:$\exists i, j \in \{dendro\_diana, dendro\_variables, hclust\_vector, hrarchy, simlrty\}\ tal\ \ que\ i \neq j,\ \sigma_i^2 \neq \sigma_j^2$

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

Group=17230; Df=4; Fvalue=2068,5; Pr(>F) < 2,2e-16 ***

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

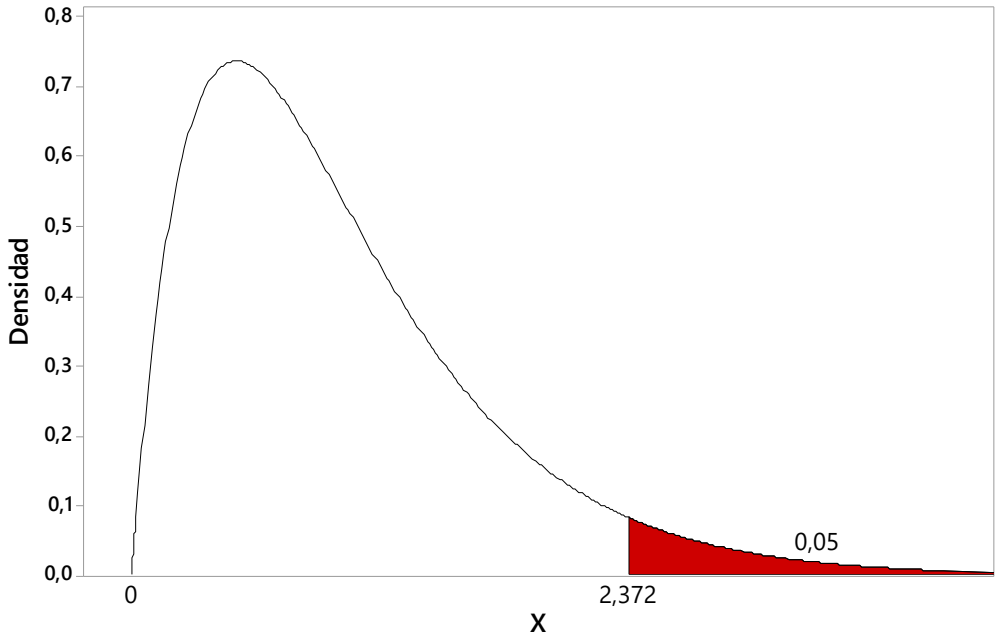


Figura 7.13.- Zonas de rechazo y aceptación para la homogeneidad de varianzas, complejidad temporal, clúster

Paso 5: Decisión

El p-value obtenido es igual a 2,2e-16 el cual es menor al nivel de significancia propuesto (0,05) por lo que se rechazó la hipótesis nula (Ho) y se concluye que las varianzas de los grupos de tiempo de procesamiento no son iguales, los datos sobre el tiempo de procesamiento para cada una de las técnicas clúster son heterocedásticos (Figura 7.13)

7.4.4.2 Test de Bartlett

Paso 1: Planteamiento de Hipótesis

H_o: $\sigma^2_{dendro\_diana} = \sigma^2_{dendro\_variables} = \sigma^2_{hclust\_vector} = \sigma^2_{hrarchy} = \sigma^2_{simlrty}$

H_1:$\exists i, j \in \{dendro\_diana, dendro\_variables, hclust\_vector, hrarchy, simlrty\}\ tal\ \ que\ i \neq j,\ \sigma_i^2 \neq \sigma_j^2$

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

Bartlett's K-squared = 81713, df = 4, p-value < 2,2e-16

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5: Decisión

Al emplear la prueba de Bartlett se obtuvo un p-value igual a 2,2e-16 siendo menor a un nivel de significancia de 0,05 por lo que se rechaza la hipótesis nula (Ho) y se concluye que la varianza de los grupos en estudio es diferente. Comparativamente entre las pruebas de Bartlett y Levene's se llega a la misma conclusión dado que el p-value para cada prueba es similar.

7.4.5 Independencia

Paso 1: Planteamiento de Hipótesis

H_o: Técnicas clúster y tiempo son independientes

H_1: Técnicas clúster y tiempo no son independientes

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

Pearson's Chi-squared test; X-squared=16658, df=12, p-value < 2,2e-16

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

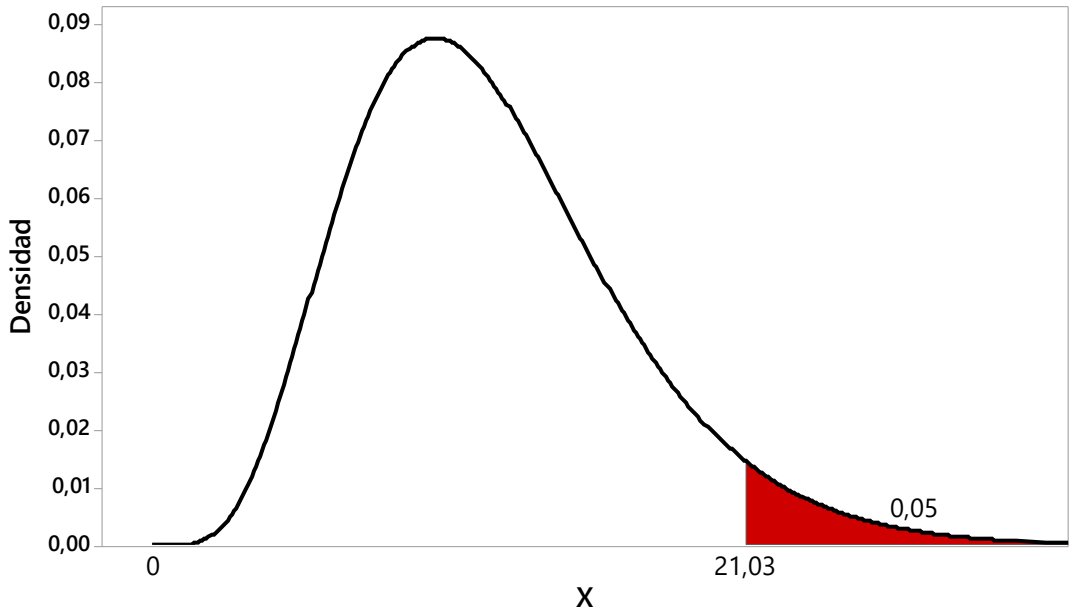


Figura 7.14.- Zonas de rechazo y aceptación para el prerrequisito de independencia

Paso 5: Decisión

El valor p obtenido en la prueba es igual a 2,2e-16 el cual es menor al nivel de significancia de 0,05 por lo que se rechaza la hipótesis nula (Ho) y se concluye que las técnicas clúster y tiempo de procesamiento no son independientes (Figura 7.14).

7.4.6 Pruebas de hipótesis

A continuación, se determinaron las estadísticas descriptivas resumen entre las variables, para lo que corresponde a las técnicas clúster se almacenó en una variable X y lo que corresponde a tiempo en la variable Y.

7.4.6.1 Medidas descriptivas específicas

Las medidas descriptivas, reflejan los resultados sobre la cantidad de memoria que se empleó por cada técnica de asociación, obteniéndose así distintos parámetros de análisis para cada uno.

La Tabla 7.28, refleja los resultados sobre la cantidad de tiempo que se empleó por cada técnica clúster, obteniéndose así distintos parámetros de análisis para cada una.

Tabla 7.28.- Medidas descriptivas de las técnicas clúster

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
dendro_diana	0,02402	0,09157	0,14568	0,30640	0,24653	6,11347
dendro_variable	0,0225	4,4838	23,2648	44,2447	61,9799	391,1287
hclust_vector	0,007087	0,050403	0,084486	0,242727	0,147840	24,250590
hrarchy	0,0948	0,6920	1,2875	1,8069	2,3829	23,6094
simlrty	0,09934	0,72206	1,34385	1,84199	2,43371	10,52140

Mediante el análisis de las medidas descriptivas se obtuvo que la técnica clúster que usa en promedio mayor tiempo de procesamiento es dendro_variables con un valor igual a 44,2447, un mínimo de tiempo de procesamiento igual a 0,0225 y un máximo de 391,1287, el primer cuartil muestra que el 25% del tiempo de procesamiento es menor o igual a 4,4838 con una mediana que indica que la mitad del tiempo empleado para la técnica clúster es menor o igual a 23,2648 y la otra mitad es mayor o igual a 23,2648, finalmente el tercer cuartil refleja que el 75% del tiempo es menor o igual a 61,9799. Al analizar la técnica que utiliza en promedio menos tiempo de procesamiento se tiene que es hclust_vector con 0,242727, posee un máximo de tiempo de 24,250590 y un mínimo de 0,007087, el primer cuartil muestra que el 25% del tiempo es menor o igual a 0,050403 con una mediana que indica que la mitad del tiempo de procesamiento empleado para la técnica clúster es menor o igual a 0,084486 y la otra mitad es mayor o igual a 0,084486, concluyendo con el tercer cuartil que refleja que el 75% del tiempo es menor o igual a 20,147840. Con respecto a las demás técnicas clúster se tiene que para dendro_diana el valor promedio del tiempo es igual a 0,30640, para hrarchy es 1,8069 y para simlrty es 1,84199 determinándose que la variación entre las últimas técnicas mencionadas es leve mientras que para dendro_diana ya es más considerable la diferencia en los resultados obtenidos.

7.4.6.2 Kruskal Wallis H-Test

Una vez analizados los prerrequisitos se concluyó que no se cumple normalidad ni independencia, ni tampoco se puede lograr la misma normalidad mediante las transformaciones realizadas (aunque si cumple homocedasticidad), se debe realizar

pruebas no paramétricas para muestras independientes. La prueba en estudio tiene como planteamiento de hipótesis las siguientes:

Paso 1: Planteamiento de Hipótesis

H_o: $\tilde{\mu}_{dendro\_diana} = \tilde{\mu}_{dendro\_variables} = \tilde{\mu}_{hclust\_vector} = \tilde{\mu}_{hrarchy} = \tilde{\mu}_{simlrty} = \tilde{\mu}_{tiempo}$

H_1: $\tilde{\mu}_i \neq \tilde{\mu}_j \ para\ al\ menos\ un\ par\ de\ (i,j)$

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

Kruskal-Wallis chi-squared=11911, df=4, p-value < 2,2e-16

Paso 4: Regla de Decisión

Si p-value < 0,05 entonces se rechaza Ho, caso contrario no se rechaza. En nuestro caso 9,488<11911, por tanto, se rechaza la hipótesis nula.

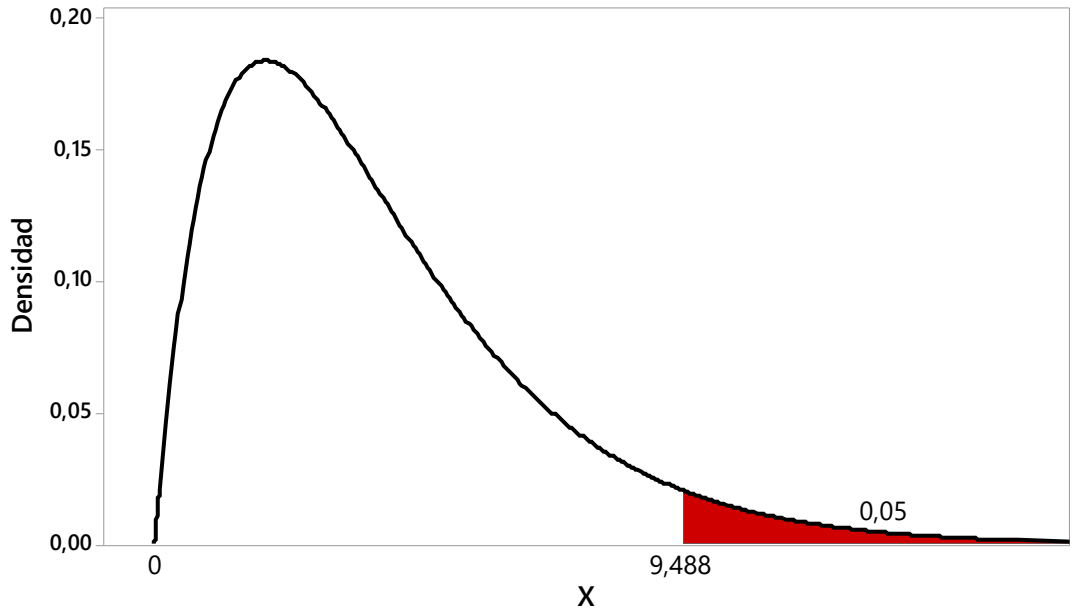


Figura 7.15.- Zonas de rechazo y aceptación para la prueba de Kruskal Wallis

Paso 5: Decisión

Se obtuvo un valor p igual a 2,2e-16 el cual es menor a un nivel de significancia de 0,05 por lo que se rechaza la hipótesis nula y se concluye que al menos un par de las técnicas clúster es diferente (Figura 7.15).

Para determinar cuál par de rangos son diferentes se utilizó la prueba no paramétrica para comparación de pares de dos muestras independientes de Mann Whitney Wilcoxon U-test.

7.4.6.3 Mann Whitney Wilcoxon U-test

Se utilizó Mann Whitney Wilcoxon U-test para la comparación por parejas independientes como se ve en la página oficial de (*R: Pairwise Wilcoxon Rank Sum Tests*, 2021). Para no provocar que el error de Tipo 1 aumente, las técnicas usadas mediante R fueron la corrección de Bonferroni en la cual los valores p se multiplican por el número de comparaciones y Holm que realizó correcciones menos conservadoras, los resultados obtenidos fueron los siguientes:

Paso 1: Planteamiento de Hipótesis

$\mathbf{H_o}$: $No\ hay\ diferencia\ entre\ los\ tiempos\ de\ las\ 2\ poblaciones\ de\ técnicas\ cluster$

$\mathbf{H_1}$: $Hay\ diferencia\ entre\ los\ tiempos\ de\ las\ 2\ poblaciones\ de\ técnicas\ cluster$

Paso 2: Nivel de significancia α=0,05 (dividido para el número de comparaciones)

Paso 3: Estadístico de Prueba

Tabla 7.29.- Comparaciones múltiples Wilcoxon (Bonferroni)

	dendro_diana	dendro_variables	hclust_vector	hrarchy
dendro_variables	< 2e-16	-	-	-
hclust_vector	< 2e-16	< 2e-16	-	-
hrarchy	< 2e-16	< 2e-16	< 2e-16	-
simlrty	< 2e-16	< 2e-16	< 2e-16	< 2e-16

Se muestran los resultados obtenidos al aplicar la prueba no paramétrica Wilcoxon mediante la corrección de Bonferroni (Tabla 7.29).

Paso 4: Regla de Decisión

Si p-value < 0,05 entonces se rechaza Ho, caso contrario no se rechaza.

Paso 5A: Decisión

Se calcularon las comparaciones por pares entre niveles de grupo con correcciones para pruebas múltiples usando el método de Bonferroni en donde se obtuvo que todos los

pares son significativamente diferentes debido a que el valor p obtenido para cada par es menor que el nivel de significancia de 0,05.

Paso 3B: Comparaciones múltiples, estadístico de prueba Wilcoxon (Holm)

Tabla 7.30.- Comparaciones múltiples Wilcoxon (Holm)

	dendro_diana	dendro_variables	hclust_vector	hrarchy
dendro_variables	< 2e-16	-	-	-
hclust_vector	< 2e-16	< 2e-16	-	-
hrarchy	< 2e-16	< 2e-16	< 2e-16	-
simlrty	< 2e-16	< 2e-16	< 2e-16	< 2e-16

Se muestran los resultados obtenidos al aplicar la prueba no paramétrica Wilcoxon en este caso se usó la técnica de Holm (Tabla 7.30).

Paso 4B: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5B: Decisión

Mediante la técnica de Holm se obtuvo que todos los pares son diferentes a un nivel de significancia α=0,05.

7.4.6.4 ANOVA no paramétrico

Utilizamos el paquete Rfit (Rank-Based Estimation for Linear Models) que proporciona funciones para análisis basados en rangos de modelos lineales, la inferencia ofrece una alternativa robusta a los mínimos cuadrados (Kloke y McKean, 2020).

Paso 1: Planteamiento de Hipótesis

$\mathbf{H_o}$: $No\ hay\ diferencia\ entre\ los\ tiempos\ de\ las\ 5\ poblaciones\ de\ técnicas\ cluster$

$\mathbf{H_1}$: $Hay\ diferencia\ entre\ los\ tiempos\ de\ las\ 5\ poblaciones\ de\ técnicas\ cluster$

Paso 2: Nivel de significancia α=0,05

Paso 3A: Estadístico de Prueba

F-Statistic=29036; p-value=0

Paso 4A: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza (Figura 7.16).

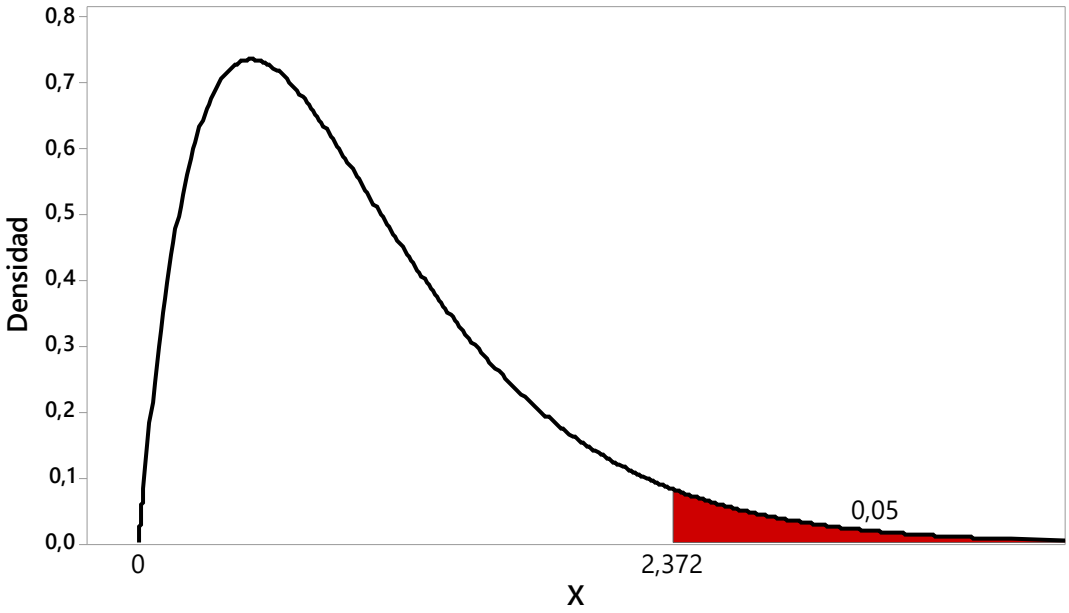


Figura 7.16.- Zonas de rechazo y aceptación ANOVA no paramétrico, complejidad temporal, clúster

Paso 5A: Decisión

El valor p obtenido en el ANOVA es igual a 0 el cual es menor a un nivel de significancia de 0,05 por lo que se rechazó la hipótesis nula y se concluye que las medias obtenidas entre las técnicas clúster y el tiempo son distintas.

Paso 3B: Estadístico de Prueba

Tabla 7.31.- Comparaciones múltiples ANOVA no paramétrico

	dendro_variables	hclust_vector	hrarchy	simlrty
dendro_variables	< 2e-16	-	-	-
hclust_vector	< 2e-16	< 2e-16	-	-
hrarchy	< 2e-16	< 2e-16	< 2e-16	-
simlrty	< 2e-16	< 2e-16	< 2e-16	< 2e-16

Se muestran los resultados obtenidos al aplicar un ANOVA mediante el paquete Rfit, se divisa que existen pares de técnicas clúster significativas y no significativas (Tabla 7.31)

Paso 4B: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5B: Decisión

Los resultados obtenidos mediante el paquete Rfit para cada uno de los pares de técnicas clúster en estudio fueron de 2e-16, para todas las parejas, por lo que todos los pares son diferentes a un nivel de significancia de 0,05.

Tukey: Mediante la opción de Tukey se determinó que las técnicas clúster son significativamente diferentes con respecto al tiempo que usa cada una, para lo cual se plantearon las siguientes hipótesis.

Paso 1: Planteamiento de Hipótesis

H_o: $No\ hay\ diferencia\ entre\ los\ tiempos\ de\ las\ 2\ poblaciones\ de\ técnicas\ cluster$

H_1: $Hay\ diferencia\ entre\ los\ tiempos\ de\ las\ 2\ poblaciones\ de\ técnicas\ cluster$

Paso 2: Nivel de significancia α=0,05

Paso 3A: Estadístico de Prueba

Tabla 7.32.- Comparaciones múltiples (Tukey)

	I	J	Estimate	St Err	Lower Bound CI	Upper Bound CI
1	dendro_diana	dendro_variables	22,78464	0,01321	22,74860	22,82068
2	dendro_diana	hclust_vector	-0,05537	0,01321	-0,09141	-0,01934
3	dendro_diana	hrarchy	1,07725	0,01321	1,04122	1,11329
4	dendro_diana	simlrty	1,12512	0,01321	1,08908	1,16115
5	dendro_variables	hclust_vector	22,84001	0,01321	22,80398	22,87605
6	dendro_variables	hrarchy	21,70739	0,01321	21,67135	21,74342
7	dendro_variables	simlrty	21,65952	0,01321	21,62349	21,69556
8	hclust_vector	hrarchy	-113,263	0,01321	-116,867	-109,659
9	hclust_vector	simlrty	-118,049	0,01321	-121,653	-114,445
10	Hrarchy	simlrty	-0,04786	0,01321	-0,08390	-0,01182

La Tabla 7.32 muestra las comparaciones múltiples entre las diferentes técnicas clúster.

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5: Decisión

Los resultados obtenidos indican que cada par de técnicas forman niveles diferentes como se muestra en la Tabla 7.33.

Tabla 7.33.- Diferentes niveles de acuerdo con el tiempo

NIVEL	TÉCNICA	TIEMPO
1	hclust_vector	0,24273
2	dendro_diana	0,30640
3	hrarchy	1,80690
4	simlrty	1,84199
5	dendro_variables	44,24470

7.5 Conclusiones

Para demostrar las hipótesis se utilizó un diseño pre-experimental del tipo RGXO1, se trabajó con un nivel de significancia de α=0,05, las variables dependientes fueron la variable memoria (para el caso de la complejidad espacial) y la variable tiempo (para el caso de la complejidad temporal), ambas de tipo numérico.

Sobre la comparación entre la complejidad espacial de las técnicas clúster de ASI (hrarchy y simlrty) y LA (hclust_vector, dendro_variables y dendro_diana) se obtuvieron los siguientes resultados:

El estudio descriptivo dio como resultado la Figura 7.17, que presenta las cuatro medidas básicas de centralización y posición de las técnicas clúster respecto a la memoria.

TÉCNICA / DESCRIPTOR	Mínimo	Mediana	Media	Máximo
dendro_diana	104,0	247,0	223,0	382,0
dendro_variables	105,0	256,0	231,3	384,0
hclust_vector	103,0	247,0	223,0	381,0
hrarchy	105,0	249,0	224,8	319,0
simlrty	105,0	248,0	223,9	319,0

Figura 7.17.- Medidas descriptivas muestrales de memoria en las técnicas clúster

El análisis de la muestra suministra la información de que la técnica que usa en promedio menor memoria es hclust_vector con un valor igual a 223,0, con un mínimo de memoria de 103,0 y un máximo de 381,0, con una mediana de 247,0. Las otras técnicas clúster se muestran bastante parecidas entre ellas sobre todo en sus medidas de centralización, las

posteriores pruebas de hipótesis verificarán si se mantiene lo observado también en la población.

Respecto al tiempo, el estudio descriptivo dio como resultado las siguientes 4 medidas básicas de centralización y posición de las técnicas clúster (Figura 7.18).

TÉCNICA / DESCRIPTOR	Mínimo	Mediana	Media	Máximo
dendro_diana	0,02402	0,09157	0,30640	6,11347
dendro_variables	0,0225	4,4838	44,2447	391,1287
hclust_vector	0,007087	0,050403	0,242727	24,250590
hrarchy	0,0948	0,6920	1,8069	23,6094
simlrty	0,09934	0,72206	1,84199	10,52140

Figura 7.18.- Medidas descriptivas muestrales de tiempo en las técnicas clúster

El análisis de la muestra suministra la información de que la técnica que usa en promedio menor tiempo es hclust_vector con un valor promedio de 0,242727, con un mínimo de memoria igual a 0,007087 y un máximo de 24,250590, con una mediana de 0,050403. Las otras técnicas clúster se muestran diferentes entre ellas, sobre todo la dendro_variables, las posteriores pruebas de hipótesis verificarán si se mantiene lo observado.

Tanto los datos de memoria como de tiempo obtenidos de las técnicas clúster no son normales para ninguna de las pruebas utilizadas (Anderson-Darling normality test, Holm technique, Lilliefors (Kolmogorov-Smirnov) normality test, Cramer-von Mises normality test, Pearson chi-square normality test, Shapiro-Wilk normality test, Shapiro-Francia normality test), a ninguno de los niveles del sesgo α=0,01, α=0,05 y tampoco α=0,1 se aplicaron técnicas de transformación a la normalidad de los datos de memoria y tiempo

usando la función powerTransform, con las opciones bcPower y yjPower, se evidenció que a los datos tanto de la variable memoria como de la variable tiempo no se les puede normalizar ya que los p-values obtenidos son bastante pequeños (< 2,22e-16 para ambos métodos).

Se comprobó que no existía homocedasticidad en los datos de memoria, pero sí en los datos de tiempo en las técnicas clúster. Al utilizar las pruebas de Bartlett y Levene, se llega a la misma conclusión que en el caso de la memoria, aunque existe una leve variación entre el valor p de cada prueba, 0,9532 y 0,9625 respectivamente. En el caso del tiempo también se llegó a las mismas conclusiones utilizando las dos pruebas y con un alto nivel de significancia dado por un p-valor menor de 2,2e-16.

El valor p obtenido en la prueba chi cuadrado $\chi 2$ es menor a 2,2e-16 en el caso de la memoria y menor a 2,2e-16 en el caso del tiempo, los cuales son menores que el nivel de significancia de α=0,05 por lo que se rechaza la hipótesis nula y se concluye que las técnicas clúster y la memoria (también las técnicas clúster y el tiempo) no son independientes.

Las pruebas de hipótesis no paramétricas aplicadas en la comparación de la complejidad espacial (memoria) entre las 5 técnicas clúster (dendro_diana, dendro_variables, hclust_vector, hrarchy y simlrty) indican que a un nivel de significancia de α=0,05 y utilizando las pruebas de Kruskal Wallis (p-value < 2,2e-16) y ANOVA no paramétrico (p-value=0,000) se rechaza la hipótesis nula con alta significancia, por lo tanto existe al menos un par de poblaciones clúster diferentes respecto a la memoria. Aplicando las pruebas posteriores de Mann Whitney Wilcoxon (con las correcciones de Bonferroni y Holm), ANOVA no paramétrico y Tukey, se obtuvieron 2 grupos de homogeneidad (A y B) que se resume en la Tabla 7.34.

Tabla 7.34.- Grupos de homogeneidad para la memoria en las técnicas clúster (las A en rojo son las técnicas con menor parámetro)

GRUPOS	dendro_diana	dendro_variables	hclust_vector	hrarchy	Simlrty
1	A		A	A	A
2		B			

Las pruebas de hipótesis aplicadas a la comparación de la complejidad temporal (tiempo) entre las 5 técnicas clúster (dendro_diana, dendro_variables, hclust_vector, hrarchy y

simlrty) indican que a un nivel de significancia de α=0,05 y utilizando las pruebas de Kruskal Wallis (p-value < 2,2e-16) y ANOVA no paramétrico (p-value=0) se rechaza la hipótesis nula con alta significancia, por lo tanto, existe al menos un par de poblaciones clúster diferentes respecto al tiempo. Aplicando las pruebas posteriores de Mann Whitney Wilcoxon (con las correcciones de Bonferroni y Holm), ANOVA no paramétrico y Tukey, se obtuvieron 5 grupos de homogeneidad que se resume en la Tabla 7.35.

Tabla 7.35.- Grupos de homogeneidad para el tiempo en las técnicas clúster (la A en rojo es la técnica con menor parámetro)

GRUPOS	dendro_diana	dendro_variables	hclust_vector	hrarchy	Simlrty
1			**A**		
2	B				
3				C	
4					D
5		E			

Considerando la complejidad espacial y temporal simultáneamente (en forma ascendente) de las técnicas clúster analizadas se obtendría la Tabla 7.36.

Tabla 7.36.- Complejidad espacial y temporal simultáneamente para las técnicas clúster.

ORDEN	TÉCNICA CLÚSTER
1	hclust_vector
2	dendro_diana
3	hrarchy
4	Simlrty
5	dendro_variables

Donde se observa que la técnica con menor ocupación de memoria y tiempo en las técnicas clúster para bases de datos de máximo 1000 observaciones y 100 variables es hclust_vector, ubicándose hrarchy y Simlrty sobre y muy cerca de la mediana.

Capítulo 8vo | COMPLEJIDAD ALGORÍTMICA ENTRE REGLAS DE ASOCIACIÓN DE LA Y ASI

Se comparan estadísticamente las técnicas de reglas de asociación apriori, eclat y weclat de LA con la técnica implicativeGraph del ASI, desde el punto de vista de la complejidad algorítmica.

8 Capítulo.- Complejidad algorítmica entre reglas de asociación de LA y ASI

En este capítulo se comparan computacionalmente las técnicas de cálculo de reglas de asociación de LA y las técnicas similares de cuasi-implicación proporcionadas por el ASI, es decir se compara experimentalmente la complejidad algorítmica. Primeramente, se realiza un análisis respecto a la ocupación de memoria y luego se realiza un estudio del tiempo de ejecución. En la introducción se detallan los materiales y métodos utilizados.

8.1 Introducción

El concepto de complejidad algorítmica es importante tenerlo presente, permite mediante la comparación del tiempo de ejecución y el espacio de memoria seleccionar el mejor entre algoritmos que tienen el mismo objetivo (Capítulo 7, Subsección 7.1.1). En este capítulo se muestran los principios de las reglas de asociación, se revisan los artículos ASI que realizan análisis comparativos con otras técnicas de reglas de asociación y finalmente se muestra la metodología utilizada para la comparación de la complejidad espacial y temporal. Todos los detalles sobre las técnicas de reglas de asociación utilizadas se encuentran en el Apéndice D.- Manual de estadísticas utilizadas.

8.1.1 Técnicas de minería de asociación

La minería de reglas de asociación es un conjunto de técnicas de las más importantes en la minería de datos, fue propuesta por primera vez por Agrawal, Imielinski y Swami (Agrawal et al., 1993). Su objetivo es extraer interesantes patrones frecuentes, asociaciones entre conjuntos de elementos en las bases de datos de transacciones. Algunas áreas de aplicación de las reglas de asociación son la bioinformática, la geo informática, la detección de instrucciones, la minería del uso de la web, las analíticas de aprendizaje, etc., en esta sección hacemos notar sus bases conceptuales (R. Pazmiño-Maji, García-Peñalvo, et al., 2019).

La minería de reglas de asociación se define de la siguiente manera: Sea $I = \{i_1, i_2, \dots, i_m\}$ un conjunto de elementos y D un conjunto de transacciones (conjunto de datos transaccionales) donde cada transacción $T \subseteq I$ está asociada con un identificador TID y m es el número de elementos. Sean A y B dos conjuntos de elementos, se dice que una transacción T contiene A si y solo si $A \subseteq T$. Una regla de asociación es una implicación en la forma $A => B$ donde $A \subset I,\ B \subset I$ y $A \cap B = \emptyset$. A se llama antecedente mientras que B

se llama consecuente; la regla significa que A implica B (R. Pazmiño-Maji, García-Peñalvo, et al., 2019).

Hay dos criterios básicos que utilizan las reglas de asociación, el soporte y la confianza. Reflejan, respectivamente, la utilidad y la certeza de las reglas descubiertas. Normalmente, las reglas de asociación se consideran interesantes si satisfacen tanto un umbral de soporte mínimo como un umbral de confianza mínimo.

El soporte, que indica la frecuencia (probabilidad) de toda la regla con respecto a D, se define como la relación entre el número de transacciones que contienen A y B y el número total de transacciones (la probabilidad de que tanto A como B coexistan en D) $soporte(A => B) = P(A \cup B)$. El soporte de un elemento es una significación estadística de una regla de asociación. Suponga que el soporte de un artículo es del 0,1%, significa que solo el 0,1% de la transacción contiene la compra de este artículo. El soporte es una medida útil debido a que, si es demasiado bajo, la regla puede ocurrir simplemente por casualidad. Además, en un entorno empresarial, una regla que cubra muy pocos casos puede no ser útil porque no tiene sentido (de tipo comercial) actuar de acuerdo con dicha regla debido a que no sería rentable.

La confianza, indica la fuerza de la implicación en la regla, se define como la relación entre el número de transacciones que contienen A y B y el número de transacciones que contienen A (condicional probabilidad de B dado A). $confianza(A => B) = P(B|A)$. Supongamos que la confianza de la regla de asociación $A => B$ es 80%, significa que el 80% de las transacciones que contienen A también contienen B juntas. Por tanto, la confianza determina la previsibilidad de la regla. Si la confianza de una regla es demasiado baja, no se puede inferir o predecir de manera confiable B a partir de A. Una regla con baja previsibilidad es de uso limitado.

La tarea de la minería de reglas de asociación es descubrir reglas sólidas en grandes bases de datos. El problema de las reglas de asociación minera se puede descomponer en dos partes: la primera es descubrir los conjuntos de elementos grandes, es decir, los conjuntos de elementos que tienen soporte de transacciones por encima de un umbral mínimo predeterminado y la segunda consiste en utilizar los conjuntos de elementos grandes para generar reglas de asociación para la base de datos que tengan una confianza c por encima de un umbral mínimo (R. Pazmiño-Maji et al., 2017a).

8.1.2 Trabajos relacionados

El siguiente trabajo titulado “Association rules with SIA in *B-learning* Courses: A mapping review”, determina las reglas de asociación con el ASI, aplicados a cursos de *B-learning* en la Facultad de Ciencias de la ESPOCH. Se utilizó la revisión sistemática en cursos de Blended-Learning utilizados en los últimos 5 años (2012 a 2016) desarrollados en el LMS institucional. Se inició con 3350 cursos y finalmente 13 tenían todos los criterios de calidad. Después de la revisión de literatura el único trabajo al respecto en este ámbito se debe al autor de esta tesis (R. Pazmiño-Maji et al., 2017a), este artículo también describe una experiencia sobre las reglas de asociación con ASI aplicados a cursos Blended-Learning en los últimos cinco años.

No se han encontrado otros estudios sobre complejidad algorítmica entre técnicas ASI y minerías de asociación.

8.2 Materiales y métodos

El cálculo y una cota inferior del tamaño de la población se muestra en forma detallada en el Apéndice. Por el gran tamaño de la población, se escogió una muestra utilizando el método de muestreo aleatorio simple con parámetro de interés la media, se consideró esta fórmula para el cálculo de la muestra $n = \frac{S^2}{\frac{E^2}{Z^2_{\frac{\alpha}{2}}} + \frac{S^2}{N}}$.

Para aplicar la fórmula se utilizaron los parámetros desviación estándar=1; α=5%; Z=1,96; E=10,01%; N=100000 y se generó un tamaño de la muestra de 383,2 que redondeado es 383. Se utilizaron tres computadores con el mismo microprocesador: Intel® Core ™ i7-CPU @ 2,2 Ghz y 8Gb de memoria RAM, se han instalado los sistemas operativos Windows 8-64 bits, Linux – Ubuntu 16.04-64 bits y MAC OS 10-64 bits. Todos los computadores y sistemas operativos trabajaron con el software estadístico libre R, versión 3.4.1; el entorno de desarrollo integrado libre RStudio, versión 1.0.143 y el paquete Rchic, versión 0.24. Las bases de datos se generaron aleatoriamente utilizando la función runif() perteneciente al paquete estándar de R. Los datos utilizados fueron dicotómicos generados por la función runif() y round(). Las funciones para utilizar fueron en LA: apriori, weclat, eclat; y en ASI: implicativeGraph. Las hipótesis estadísticas que se demostraron fueron normalidad según test de Anderson-Darling, test de hipótesis de

Kruskal-Wallis y su respectivo post test. Para demostrar las hipótesis se planteó un pre-experimento en la ingeniería de software de tipo RGXO1. Donde RG representa el grupo aleatorio del grupo experimental (tanto-inter como intra-grupos), X representa el tratamiento que en este caso son las 3 técnicas clúster jerárquicas utilizadas en LA (hclust_vector, dendro_variables y dendro_diana) y 2 técnicas usadas en ASI (hrarchy y simlrty). Se trabajó con un nivel de significancia del 95%. Las variables dependientes fueron la variable memoria (para el caso de la complejidad espacial) y la variable tiempo (para el caso de la complejidad temporal), ambas de tipo numérico.

Por el paradigma de investigación es de tipo cuantitativo, por el tipo de diseño utilizado es pre-experimental, por el tiempo de estudio es transversal, el colectivo de estudio lo conforman las 100 000 bases de datos aleatorias formadas por lo máximo 1000 observaciones y 100 variables, por la amplitud de estudio es de muestreo de 383 bases de datos aleatorias binarias. La población es la información sobre la muestra de estudio, tal como: nombre del archivo, número de filas que conforman la base de datos, número de columnas que conforman la base de datos, el total de datos, tiempo de ejecución, memoria utilizada y sistema operativo.

8.3 Estudio de la complejidad espacial

A continuación, se realiza el estudio de complejidad espacial respecto a las reglas de asociación. Primeramente, se realiza un estudio descriptivo, para luego realizar las pruebas de hipótesis no sin antes realizar la prueba de los supuestos de normalidad, homocedasticidad e independencia para determinar el tipo de prueba a utilizar.

8.3.1 Medidas descriptivas

Las medidas descriptivas reflejan los resultados sobre la cantidad de memoria que se empleó por cada método de asociación, obteniéndose así distintos parámetros de análisis para cada uno (Tabla 8.1).

Tabla 8.1.- Cantidad de memoria por técnicas de reglas de asociación

ASSOCIATION RULES TECHNIQUES	mean	sd	IQR	cv	skewness	kurtosis	n
met_apriori	157,0569	12,15024	15,4	0,07736205	1,466647	5,996377	3447
met_ASI	158,4308	12,20701	15,5	0,07704951	1,45902	5,935088	3447
met_eclat	156,9651	12,157	15,5	0,07745033	1,463304	5,970295	3447
met_weclat	156,9296	12,16103	15,5	0,07749353	1,463439	5,964025	3447

Primero se determinó la media, la cual indica que la técnica que ocupa menos memoria es weclat con un valor de 156,9296 y la técnica que en promedio usa mayor cantidad de memoria es el ASI con un valor igual a 158,4308, al analizar la dispersión en la que se encuentran dichos datos analizados con respecto al valor promedio que presentaron se obtuvo que aquella técnica que presenta menor dispersión es apriori con un valor igual a 12,1504 y la técnica con mayor dispersión es el ASI con un valor de 12,20701. El rango intercuartil (IQR) da a notar que la técnica que presenta menor variación en sus datos es apriori con un valor igual a 15,4 mientras que las demás técnicas presentan la misma variación de 15,5; se nota que la técnica que posee menor coeficiente de variación (cv) es el ASI con un 7,704951% de dispersión respecto a la media mientras que la técnica con mayor coeficiente de variación es weclat con 7,749353%. Cabe recalcar que no existe mayor diferencia entre la dispersión de los datos de las técnicas mencionadas ya que la diferencia entre cada una recae solo en los decimales.

La asimetría de los datos (*skewness*) permitió notar que todas las reglas de asociación poseen una distribución simétrica, es decir que existe aproximadamente la misma cantidad de datos a los dos lados del valor de la media correspondiente, en cuanto a qué técnica de regla de asociación posee mayor asimetría es apriori con 1,466647 y con menor asimetría es ASI con 1,45902. Cada coeficiente de *kurtosis* obtenido para cada una de las técnicas muestra que la distribución que siguen es leptocúrtica dado que dichos valores son positivos, lo que quiere decir que hay una mayor concentración de los datos en torno a la media siendo la técnica a priori la que posee mayor coeficiente y ASI el que tiene menor valor. Todas las técnicas de reglas de asociación se trabajaron con una muestra de 3447 datos.

Se realizó un gráfico de Violín comparativo para cada uno de los 5 métodos analizados, que se muestra en la Figura 8.1.

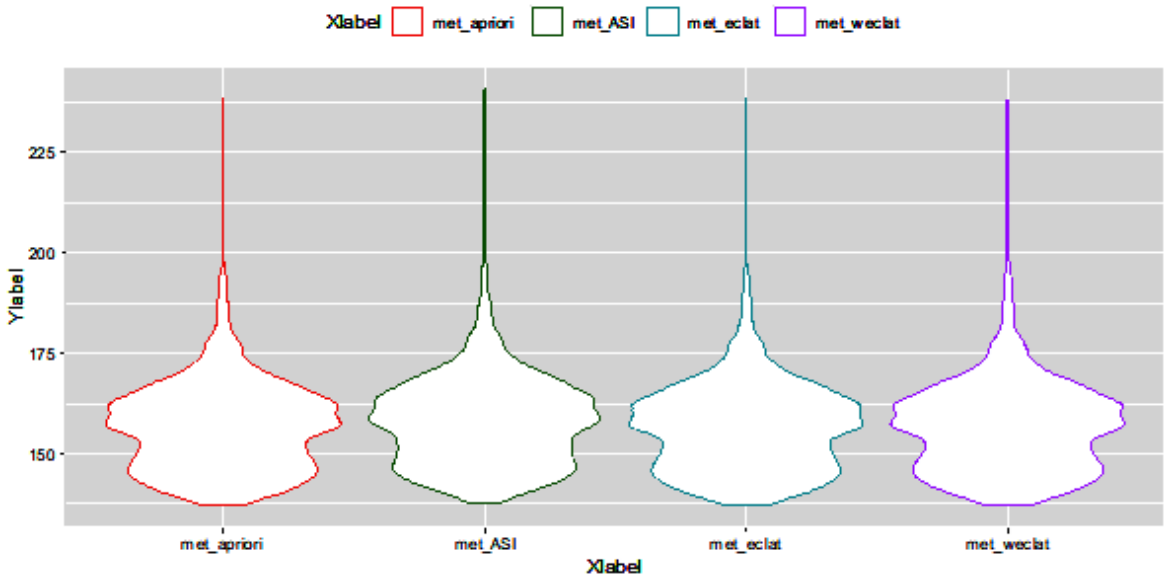


Figura 8.1.- Gráfico de violín sobre la cantidad de memoria por técnica de asociación

Debido a que los datos en estudio son de gran volumen se realizó una gráfica de violín de acuerdo con la cantidad de memoria que se está usando por cada regla de asociación, dado que este tipo de gráfico permite una comprensión más profunda de la densidad de los datos en estudio, corroborando que no existe mayor diferencia en cómo están distribuidos los datos sobre la cantidad de memoria de cada técnica, se nota que existe una gran semejanza entre cada uno de los violines. Con el análisis de cada violín se nota que la forma general y la distribución de las puntas son muy similares en las técnicas apriori, eclat y weclat, es decir, poseen cuartiles muy cercanos entre sí; en la técnica ASI se evidenció cierta variación en la punta ya que está más angosta en comparación a la de las demás técnicas, en cuanto a la punta superior de cada violín es muy parecida para todas las técnicas, por lo que es razonable concluir que existe una cantidad similar de datos atípicos en cada técnica. En cuestión a la forma general del violín, la técnica ASI es la que presenta rasgos en su distribución que se diferencian a los violines de las demás técnicas de reglas de asociación.

8.3.2 Normalidad

Para determinar la prueba apropiada a utilizar en la hipótesis se procedió a la comprobación de los supuestos.

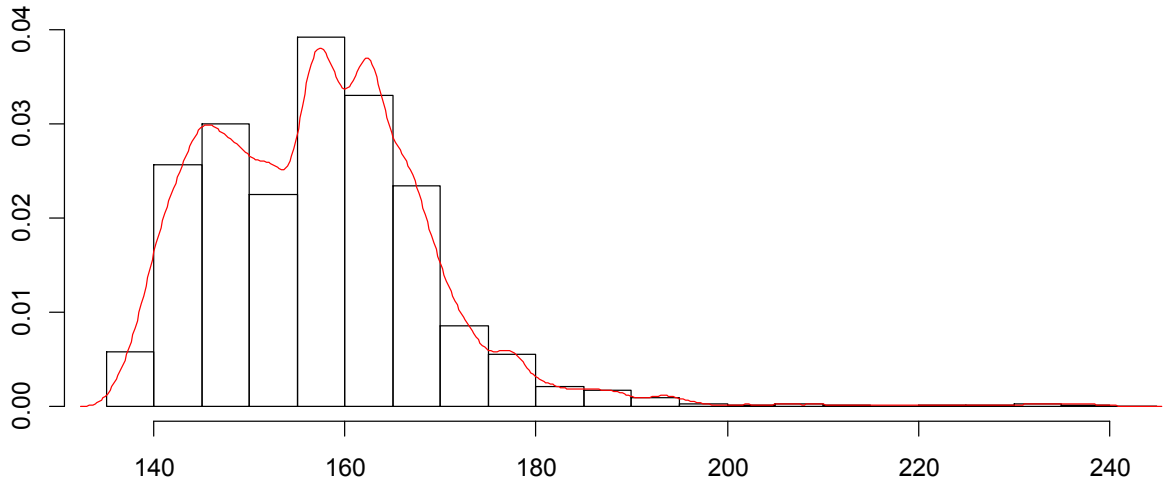


Figura 8.2.- Aproximación normal de los datos de memoria

La Figura 8.2, de aproximación normal de los datos de memoria evidencia que el histograma es asimétrico hacia la derecha, lo que quiere decir que los datos muestran un sesgo positivo respecto a la línea de distribución ajustada, se nota que ciertas barras no la siguen muy de cerca por lo que no parece ofrecer un ajuste adecuado para una distribución normal, hay más datos de lo esperado en la cola derecha.

A continuación, se muestra la gráfica de cuartiles (Figura 8.3) que provee una idea gráfica de la normalidad de los datos sobre el índice de rangos (cuartiles) en las diferentes técnicas.

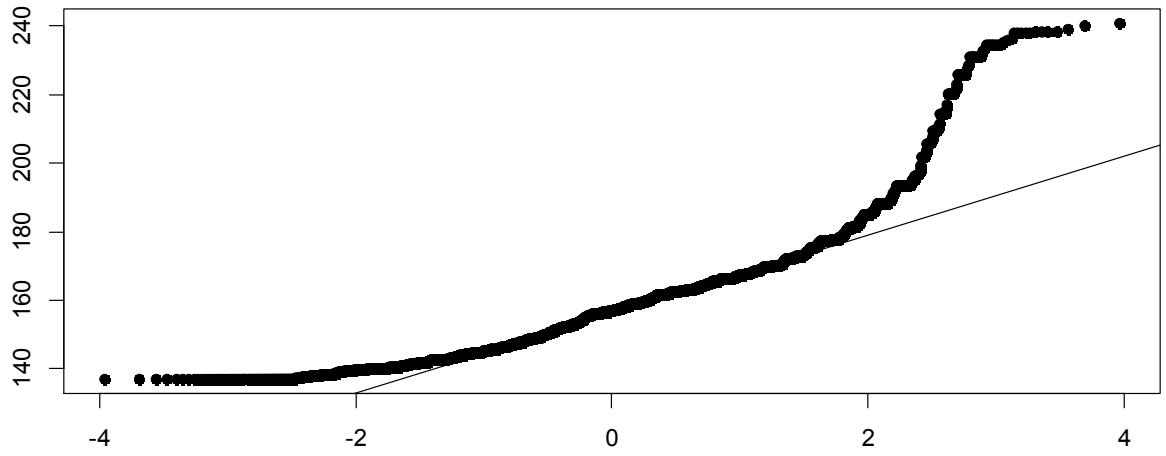


Figura 8.3.- Gráfico de cuartiles QQ para los datos de memoria

Al analizar la cercanía de la recta a la curva presentada en la Figura 8.3, sobre la cantidad de memoria usada por cada una de las técnicas de reglas de asociación, se observa que cierta cantidad de puntos están situados en la línea recta mientras que otros no y se evidencia claramente, ya que dichos puntos sobresalen de la recta indicando que probablemente no existe normalidad, para verificar dicha aseveración se realizaron los respectivos test de normalidad a niveles de significancia de 0,5, 0,01 y 0,1.

La Tabla 8.2, refleja los resultados sobre la cantidad de memoria que se empleó por cada regla de asociación, obteniéndose así distintos parámetros de análisis para cada una. Primero se determinó la media la cual indica que la técnica que ocupa menos memoria es weclat con un valor de 156,9296 y la técnica que en promedio usa mayor cantidad de memoria es ASI con un valor igual a 158,4308, al analizar la dispersión en la que se encuentran dichos datos analizados con respecto al valor promedio que presentaron se obtuvo que la técnica que presenta menor dispersión es apriori con un valor igual a 12,1504 y la técnica con mayor dispersión es ASI con un valor de 12,20701. El rango intercuartil (IQR) da a notar que la técnica que presenta menor variación en sus datos es apriori con un valor igual a 15,4 mientras que las demás técnicas presentan la misma

variación de 15,5; se nota que la técnica que posee menor coeficiente de variación (cv) es el ASI con 7,7% de dispersión respecto a la media mientras que la técnica con mayor coeficiente de variación es weclat con 7,749353%. Cabe recalcar que no existe mayor diferencia entre la dispersión de los datos de las técnicas mencionadas ya que la diferencia entre cada una recae solo en los decimales.

Tabla 8.2.- Cantidad de memoria por técnicas de reglas de asociación

ASSOCIATION RULES TECHNIQUES	mean	sd	IQR	cv	skewness	kurtosis	n
met_apriori	157,0569	12,15024	15,4	0,07736205	1,466647	5,996377	3447
met_ASI	158,4308	12,20701	15,5	0,07704951	1,45902	5,935088	3447
met_eclat	156,9651	12,157	15,5	0,07745033	1,463304	5,970295	3447
met_weclat	156,9296	12,16103	15,5	0,07749353	1,463439	5,964025	3447

La asimetría de los datos (*skewness*) permitió notar que todas las reglas de asociación poseen una distribución simétrica, es decir que existe aproximadamente la misma cantidad de datos a los dos lados del valor de la media correspondiente, en cuanto a qué técnica de regla de asociación posee mayor asimetría es la apriori con 1,466647 y con menor asimetría es ASI con 1,45902. Cada coeficiente de *kurtosis* obtenido para cada una de las técnicas muestra que la distribución que siguen es leptocúrtica dado que dichos valores son positivos, lo que quiere decir que hay una mayor concentración de los datos en torno a la media siendo la técnicaa apriori (5,996377) la que posee mayor coeficiente y ASI (5,935088) la que tiene menor valor (Tabla 8.2). Todas las técnicas de reglas de asociación se trabajaron con una muestra de 3447 datos.

La selección de la prueba de hipótesis adecuada depende del cumplimiento de ciertos supuestos como normalidad, homocedasticidad y dependencia. A continuación, se prueba el supuesto de normalidad probado con valores de α=0,01, α=0,05 y α=0,1 (en ese orden), buscando la más alta significatividad.

8.3.2.1 Paso 1A (α=0,01), 1B (α=0,05) y 1C (α=0,1): Planteamiento de Hipótesis

H_o: Ocupación de memoria ~ $N(\mu, \sigma^2)$
H_1: Ocupación de memoria $\nsim N(\mu, \sigma^2)$

8.3.2.2 Paso 2A (α=0,01), 2B (α=0,05) y 2C (α=0,1): Nivel de significancia

α=0,01, α=0,05 y α=0,1

8.3.2.3 Paso 3A (α=0,01), 3B (α=0,05) y 3C (α=0,1): Estadístico y valor p

La Tabla 8.3, muestra los resultados obtenidos de cada prueba de normalidad, se visualiza el valor del estadístico y el valor p para tomar su respectiva decisión de acuerdo con su nivel de significancia.

Tabla 8.3.- Resultados de las pruebas de normalidad, reglas de asociación y variable memoria: estadístico y valor p

	met_apriori	met_ASI	met_eclat	met_weclat
Anderson-Darling normality test	A = 28,547, p-value < 2,2e-16	A = 26,421, p-value < 2,2e-16	A = 28,569, p-value < 2,2e-16	A = 28,763, p-value < 2,2e-16
Holm technique	met_apriori < 2,22e-16 < 2,22e-16	met_ASI < 2,22e-16 < 2,22e-16	met_eclat < 2,22e-16 < 2,22e-16	met_weclat < 2,22e-16 < 2,22e-16
Lilliefors (Kolmogorov-Smirnov) normality test	D = 0,066071, p-value < 2,2e-16	D = 0,056365, p-value < 2,2e-16	D = 0,064666, p-value < 2,2e-16	D = 0,065732, p-value < 2,2e-16
Holm technique	met_apriori < 2,22e-16 < 2,22e-16	met_ASI < 2,22e-16 < 2,22e-16	met_eclat < 2,22e-16 < 2,22e-16	met_weclat < 2,22e-16 < 2,22e-16
Cramer-von Mises normality test	W = 3,2498, p-value= 7,37e-10	W = 2,9112, p-value= 7,37e-10	W = 3,2395, p-value= 7,37e-10	W = 3,2723, p-value= 7,37e-10
Holm technique	met_apriori 7,37e-10 2,948e-09	met_ASI 7,37e-10 2,948e-09	met_eclat 7,37e-10 2,948e-09	met_weclat 7,37e-10 2,948e-09
Pearson chi-square normality test	P = 1539, p-value < 2,2e-16	P = 766,14, p-value < 2,2e-16	P = 1390,7, p-value < 2,2e-16	P = 1668.7, p-value < 2,2e-16
p-values adjusted by the Holm technique	unadjusted adjusted met_apriori < 2,22e-16 < 2,22e-16	met_ASI < 2,22e-16 < 2,22e-16	met_eclat < 2,22e-16 < 2,22e-16	met_weclat < 2,22e-16 < 2,22e-16
Shapiro-Wilk normality test	W = 0,91256, p-value < 2,2e-16	W = 0,91493, p-value < 2,2e-16	W = 0,9126, p-value < 2,2e-16	W = 0,91248, p-value < 2,2e-16
p-values adjusted by the Holm technique	met_apriori < 2,22e-16 < 2,22e-16	met_ASI < 2,22e-16 < 2,22e-16	met_eclat < 2,22e-16 < 2,22e-16	met_weclat < 2,22e-16 < 2,22e-16
Shapiro-Francia normality test	W = 0,91236, p-value < 2,2e-16	W = 0,91471, p-value < 2,2e-16	W = 0,9124, p-value < 2,2e-16	W = 0,91228, p-value < 2,2e-1
p-values adjusted by the Holm technique	unadjusted adjusted met_apriori < 2,22e-16 < 2,22e-16	met_ASI < 2,22e-16 < 2,22e-16	met_eclat < 2,22e-16 < 2,22e-16	met_weclat < 2,22e-16 < 2,22e-16

8.3.2.4 Paso 4A: Regla de decisión para α=0,01

Si el p valor es menor que 0,01 (p-value < 0,01) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla.

8.3.2.5 Paso 4B: Regla de decisión para α=0,05

Si el p valor es menor que 0,05 (p-value < 0,05) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla.

8.3.2.6 Paso 4C: Regla de decisión para α=0,1

Si el p valor es menor que 0,1 (p-value < 0,1) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla.

8.3.2.7 Paso 5C: Tabla de resultados α=0,01

La Tabla 8.4, muestra los resultados de normalidad para α=0,01 para cada una de las 12 pruebas de normalidad utilizadas.

Tabla 8.4.- Resultados de la normalidad de la variable tiempo para un valor de α=0,01

	met_apriori	met_ASI	met_eclat	met_weclat
Anderson-Darling	No son normales	No son normales	No son normales	No son normales
Holm method	No son normales	No son normales	No son normales	No son normales
Lilliefors (Kolmogorov-Smirnov)	No son normales	No son normales	No son normales	No son normales
Cramer-von Mises	No son normales	No son normales	No son normales	No son normales
Holm method	No son normales	No son normales	No son normales	No son normales
Pearson chi-square	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method:	No son normales	No son normales	No son normales	No son normales
Shapiro-Wilk	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method:	No son normales	No son normales	No son normales	No son normales
Shapiro-Francia	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm method:	No son normales	No son normales	No son normales	No son normales

8.3.2.8 Paso 5A: Tabla de resultados α=0,05

Dado que el p-value obtenido es pequeño (Tabla 8.3), cae en la zona de rechazo, por lo que se rechaza la hipótesis nula para cada una de las pruebas de normalidad a un nivel de significancia de 0,01, 0,05 y 0,1 para el estadístico de Anderson Darling. Un valor p de 2,2e-16 para todas las técnicas en estudio como son: apriori, ASI, eclat y weclat; dado que dicho valor mencionado anteriormente es menor a 0,01, 0,05 y 0,1 por lo que al rechazar

la hipótesis nula se concluye que los datos de ocupación de memoria no siguen una distribución normal para los otras pruebas de normalidad.

En el caso de Holm se obtuvo también un valor de 2,22e-16 para todas las técnicas de reglas de asociación en estudio, por lo cual también se rechaza la hipótesis nula dado que es menor al valor de significancia de 0,05, 0,01 y 0,1 es decir no existe normalidad en los datos de ocupación de memoria, lo mismo ocurre con la prueba de Kolmogorov y Smirnov con la corrección de Lilliefors la cual posee un valor p de 2,22e-16 para todas las técnicas. Al analizar el estadístico Cramer-von Mises se evidenció como resultado un valor p igual a 7,37e-10 para las técnicas de reglas de asociación, dicho valor es inferior al nivel de significancia igual a 0,05,0,01 y 0,1 por lo cual se rechaza la hipótesis nula concluyendo que se observa diferencia entre los datos de ocupación de memoria y la distribución normal. La técnica de Holm con la cual se ajustan los p valores dio como resultado un valor p final igual a 2,948e-09 para todas las técnicas, dicho valor es menor a 0,05, 0,01 y 0,1 por lo que se rechazó la hipótesis nula y se concluyó que los datos no siguen una distribución normal, con respecto a la prueba chi cuadrada de Pearson, p-valores ajustados por la técnica de Holm, Shapiro-Wilk y finalmente Shapiro-Francia también se repitió en su valor de p igual 2,2e-16 el cual es menor al nivel de significancia, por lo que al igual que en las pruebas de normalidad analizadas anteriormente también se rechazó la hipótesis nula y se concluye que existe diferencia entre los datos de ocupación de memoria y la distribución normal.

La Tabla 8.5 muestra los resultados de normalidad para α=0,05.

Tabla 8.5.- Resultados de la normalidad de la variable memoria para un valor de α=0,05

	met_apriori	met_ASI	met_eclat	met_weclat
Anderson-Darling normality test	No son normales	No son normales	No son normales	No son normales
Holm method	No son normales	No son normales	No son normales	No son normales
Lilliefors (Kolmogorov-Smirnov) normality test	No son normales	No son normales	No son normales	No son normales
Holm technique	No son normales	No son normales	No son normales	No son normales
Cramer-von Mises normality test	No son normales	No son normales	No son normales	No son normales
Holm technique	No son normales	No son normales	No son normales	No son normales
Pearson chi-square normality test	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales
Shapiro-Wilk normality test	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales
Shapiro-Francia normality test	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales

8.3.2.9 Paso 5B: Tabla de resultados α=0,1

La Tabla 8.6, muestra los resultados de normalidad para α=0,1, para cada una de las 12 pruebas de normalidad utilizadas.

Tabla 8.6.- Resultados de la normalidad de la variable memoria para un valor de α=0,1

	met_apriori	met_ASI	met_eclat	met_weclat
Anderson-Darling normality test	No son normales	No son normales	No son normales	No son normales
Holm technique	No son normales	No son normales	No son normales	No son normales
Lilliefors (Kolmogorov-Smirnov) normality test	No son normales	No son normales	No son normales	No son normales
Holm technique	No son normales	No son normales	No son normales	No son normales
Cramer-von Mises normality test	No son normales	No son normales	No son normales	No son normales
Holm technique	No son normales	No son normales	No son normales	No son normales
Pearson chi-square normality test	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales
Shapiro-Wilk normality test	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales
Shapiro-Francia normality test	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales

La Tabla 8.4, Tabla 8.5 y Tabla 8.6, representan todos los resultados obtenidos sobre las pruebas de normalidad en donde se verificaba si los datos cumplen o no el supuesto de normalidad, obteniéndose así que para ninguna de las pruebas en estudio se cumplió con dicho supuesto ya que todos los p-values obtenidos resultaron menores a los niveles de significancia de 0,01, 0,05 y 0,1.

8.3.3 Normalización

Se aplicaron técnicas para transformar a la normalidad los datos en estudio usando la función en R llamada powerTransform, con el objetivo de determinar la potencia óptima a la que se debe elevar la variable de interés y así obtener normalidad en los datos, dicha

técnica devuelve una matriz con columnas etiquetadas como "Est Power" para el valor de lambda que maximiza la probabilidad; "Pwr redondeado" para roundlam, y las columnas "Wald Lwr Bnd" y "Wald your Bnd" para un intervalo de confianza de la teoría normal de Wald del 95% para lambda calculado como la estimación más o menos 1,96 veces el error estándar. Para el primer caso en donde se usa la transformación "bcPower" que es el valor predeterminado para la familia de potencia Box-Cox se obtuvo en resumen lo de la Tabla 8.7:

Tabla 8.7.- Normalización por grupos

Método	p-valor	Normalización
bcPower	2,22e-16	No son normales
yjPower	2,22e-16	No son normales

Las salidas se muestran a continuación (Tabla 8.8):

Tabla 8.8.- Normalización utilizando bcPower, reglas de asociación, complejidad espacial.

# NORMALITY TRANSFORMS BY GROUPS
summary(powerTransform(Y ~ X, family="bcPower"))
bcPower Transformation to Normality

	Est Power	Rounded Pwr Wald	Lwr Bnd Wald	Upr Bnd
Y1	-3,0398	-3,04	-3.1818	-2,8978

Likelihood ratio test that transformation parameter is equal to 0 (log transformation)

		LRT	df	pval
LR test	lambda = (0)	1321,392	1	< 2,22e-16

Likelihood ratio test that no transformation is needed

		LRT	df	pval
LR test	lambda = (1)	2508,97	1	< 2,22e-16

Dado que el valor EstPower (lambda) obtenido en la prueba es negativo (-3,04) se lo toma como cero, obteniéndose así un valor p igual a 2,22e-16 el mismo que es menor a un nivel de significancia de 0,05 por lo cual se concluyó que al asignar una lambda igual a 0 a los datos, éstos no siguen una distribución normal, lo mismo sucede cuando lambda es igual a 1 ya que también se obtiene un valor p igual a 2,22e-16.

Se usó yjPower (Tabla 8.9) o también llamado método de Yeo-Johnson en donde se obtuvo un valor para EstPower (lambda) igual a -3,07 el cual es negativo por lo que se lo toma como cero, el valor p obtenido con lambda igual a 0 es 2,22e-16 el cual es menor a un nivel de significancia de 0,05 por lo que se concluyó que al asignar una lambda igual a

0 a los datos, éstos no siguen una distribución normal. Se evidencia que a los datos en estudio sobre la variable memoria no se les puede realizar una normalización, los resultados se observan de mejor manera en la Tabla 8.9, en donde lambda=(1) representa que si se puede realizar la normalización de los datos y lambda=(0) que no se puede realizar la normalización.

Tabla 8.9.- Normalización utilizando yjPower, reglas de asociación, complejidad espacial.

summary(powerTransform(Y ~ X, family="yjPower"))

yjPower Transformation to Normality

	Est Power	Rounded Pwr Wald	Lwr Bnd Wald	Upr Bnd
Y1	-3,0653	-3,07	-3.6453	-2,4852

Likelihood ratio test that transformation parameter is equal to 0

		LRT	df	pval
LR test	lambda = (0)	1327,122	1	< 2,22e-16

8.3.4 Homocedasticidad

Para la prueba de homogeneidad de varianzas (Homocedasticidad) se consideró el siguiente planteamiento de hipótesis.

8.3.4.1 Test de Levene

Paso 1: Planteamiento de Hipótesis

H₀: $\sigma^2_{met\_apriori} = \sigma^2_{met\_eclat} = \sigma^2_{met\_weclat} = \sigma^2_{met\_ASI}$

H₁:$\exists i, j \in \{met\_apriori, met\_eclat, met\_weclat, met\_ASI\}\ tal\ que\ i \neq j,\ \sigma_i^2 \neq \sigma_j^2$

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

Group=13784; Df=3; Fvalue=0,0326; Pr(>F) =0,9921

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 (2,606<0,0326) entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

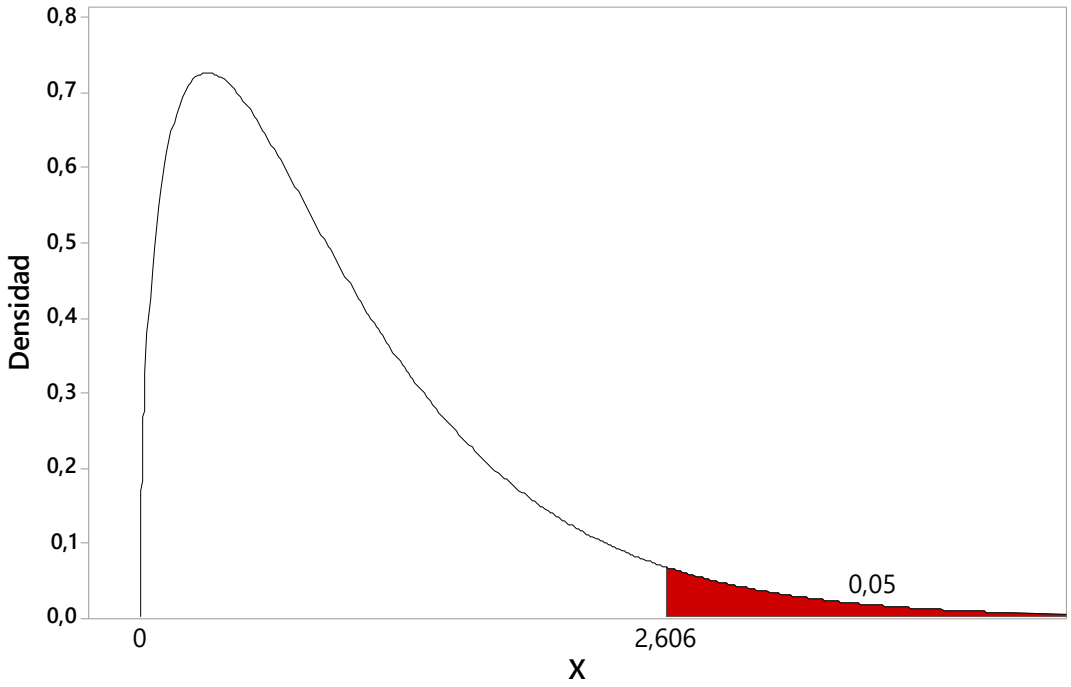


Figura 8.4.- Zonas de rechazo y aceptación para la homogeneidad de varianzas, complejidad espacial, reglas de asociación

Dado un nivel de significancia del 5% se obtuvo un valor p igual a 0,9921 el mismo que es mayor al nivel de significancia por lo cual no se rechaza Ho y se concluye que las varianzas de los grupos de memoria son iguales, los datos sobre memoria para cada una de las técnicas clúster en estudio son homocedásticos (Figura 8.4).

8.3.4.2 Test de Bartlett

Paso 1: Planteamiento de Hipótesis

H$_o$: $\sigma^2_{met\_apriori} = \sigma^2_{met\_eclat} = \sigma^2_{met\_weclat} = \sigma^2_{met\_ASI}$

H$_1$:$\exists i, j \in \{met\_apriori, met\_eclat, met\_weclat, met\_ASI\}\ tal\ que\ i \neq j,\ \sigma_i^2 \neq \sigma_j^2$

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

Bartlett's K-squared=0,093207, df=3, p-value=0,9926

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 (o 16,92<0,093207) entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5: Decisión

Mediante la prueba de Bartlett se obtuvo un valor p igual a 0,9926 el mismo que es mayor a un nivel de significancia de 0,05 por lo que no se rechaza Ho y se concluye que la varianza de los grupos en estudio no es diferente. Comparativamente entre las pruebas de Bartlett y Levene's se nota que se llega a la misma conclusión, aunque existe una leve variación entre el valor p de cada una.

8.3.5 Independencia

Test de Independencia utilizando la prueba chi.

Paso 1: Planteamiento de Hipótesis

$\mathbf{H_o}$: Técnicas de reglas de asociación y memoria son independientes.

$\mathbf{H_1}$: Técnicas de reglas de asociación y memoria no son independientes.

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

Pearson's Chi-squared test; X-squared=91,739; df=9; p-value=7,285e-16

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 (o 91,739) entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

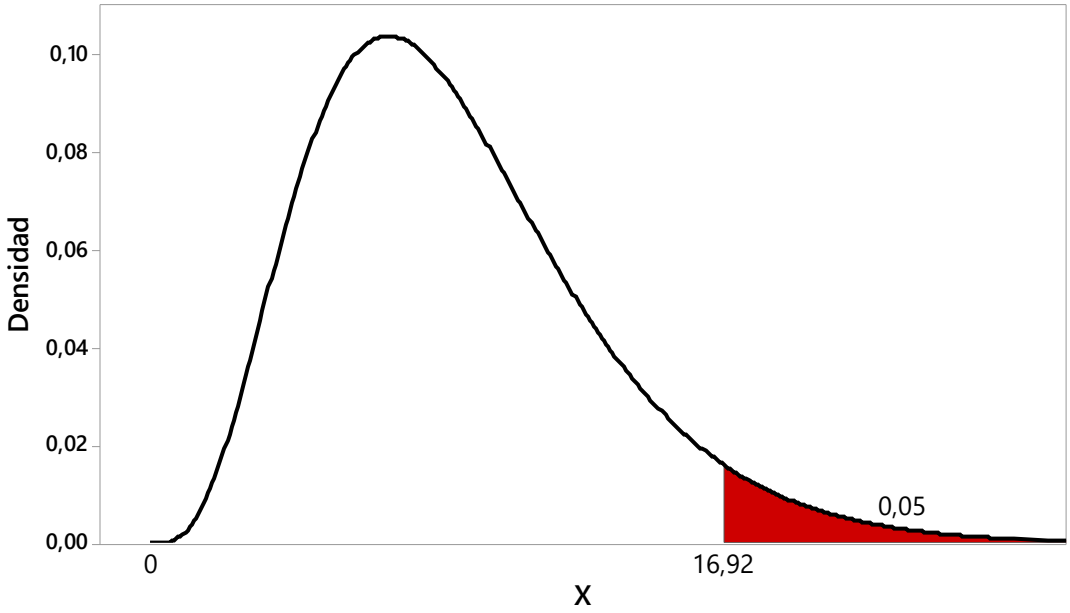


Figura 8.5.- Zonas de rechazo y aceptación para el prerrequisito de independencia

Paso 5: Decisión

El valor p obtenido en la prueba es igual a 7,285e-16 el cual es menor al nivel de significancia de 0,05 por lo que se rechaza Ho y se concluye que las técnicas de reglas de asociación y la memoria no son independientes (Figura 8.5).

8.3.6 Pruebas de hipótesis

A continuación, se determinaron las estadísticas descriptivas resumen entre las variables, para lo que corresponde a las técnicas de reglas de asociación se almacenó en una variable X y lo que corresponde a memoria en la variable Y.

8.3.6.1 Medidas descriptivas específicas

Se procedió a calcular las medidas descriptivas para cada uno de los grupos formados por las técnicas clúster, obteniéndose así los resultados que se reflejan en la Tabla 8.10.

Tabla 8.10.- Medidas descriptivas de las técnicas de reglas de asociación

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
met_apriori	136,9	147,7	156,7	157,1	163,1	238,3
met_ASI	137,5	149,2	158,1	158,4	164,7	240,9
met_eclat	136,9	147,6	156,6	157,0	163,1	238,3
met_weclat	136,9	147,6	156,6	156,9	163,1	238,2

La Tabla 8.10 además, muestra que la técnica que usa en promedio mayor memoria es ASI con un valor igual 158,4 con un mínimo de memoria igual a 137,5 y un máximo de 240,9, el primer cuartil muestra que el 25% de la memoria es menor o igual a 149,2 con una mediana que indica que la mitad de la memoria empleada para el método es menor o igual a 158,1 y la otra mitad es mayor o igual a 158,1, por último, el tercer cuartil refleja que el 75% de la memoria es menor o igual a 164,7. La técnica que utiliza menos memoria es weclat con un valor promedio igual a 156,9 con un mínimo de memoria igual a 136,9 y un máximo de 238,2, el primer cuartil muestra que el 25% de la memoria es menor o igual a 147,6 con una mediana que indica que la mitad de la memoria empleada para la técnica es menor o igual a 156,6 y la otra mitad es mayor o igual a 156,6, el tercer cuartil refleja que el 75% de la memoria es menor o igual a 163,1. Finalmente se determinó que para las técnicas en estudio no existe mayor diferencia en la memoria usada en promedio.

8.3.6.2 Kruskal Wallis H-test

Debido a que no se cumple normalidad ni independencia, ni tampoco se logra la misma normalidad mediante las transformaciones realizadas (aunque si cumple homocedasticidad), se deben realizar pruebas no paramétricas para muestras independientes. La siguiente prueba en estudio tiene como planteamiento de hipótesis las siguientes:

Paso 1: Planteamiento de Hipótesis

H_0: $\tilde{\mu}_{APRIORI} = \tilde{\mu}_{ASI} = \tilde{\mu}_{ECLAT} = \tilde{\mu}_{WECLAT} = \tilde{\mu}_{memoria}$

H_1: $\tilde{\mu}_i \neq \tilde{\mu}_j \ para\ al\ menos\ un\ par\ de\ (i,j)$

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

Kruskal-Wallis chi-squared=50,455, df=3, p-value=6,391e-11

Paso 4: Regla de Decisión

Si p-value <0,05 (o 7,815<50,455) entonces se rechaza Ho, caso contrario no se rechaza. En nuestro caso, se rechaza la hipótesis nula.

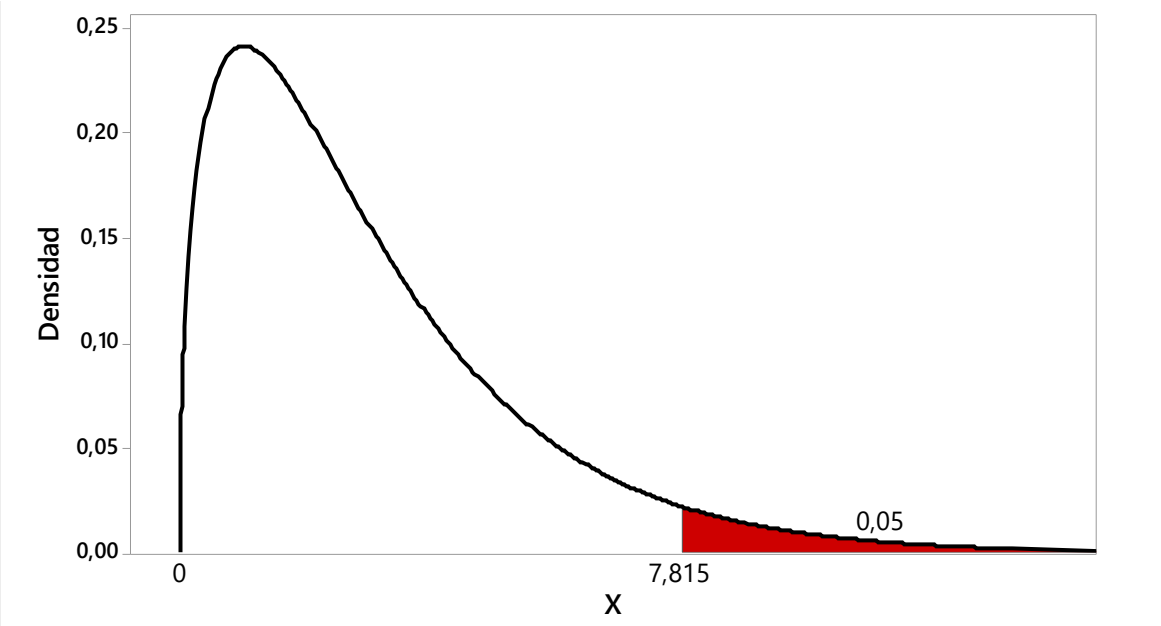


Figura 8.6.- Zonas de rechazo y aceptación para la prueba de Kruskal Wallis

Paso 5: Decisión

Se obtuvo un valor p igual a 6,391e-11 el cual es menor a un nivel de significancia de 0,05 por lo que se rechaza la hipótesis nula y se concluye que al menos una de las técnicas de reglas de asociación y la memoria son diferentes (Figura 8.6).

Para determinar cuál par de rangos son diferentes, se utilizó la prueba no paramétrica para comparación de pares de dos muestras independientes de Mann Whitney Wilcoxon U-test.

8.3.6.3 Mann Whitney Wilcoxon U-test

Se utilizó Mann Whitney Wilcoxon U-test para la comparación por parejas independientes como se ve en la página oficial de R (*R: Pairwise Wilcoxon Rank Sum Tests*, 2021). Para no provocar que el error de Tipo 1 aumente, las técnicas usadas mediante R fueron la corrección de Bonferroni en la cual los valores p se multiplican por el número de comparaciones y Holm que realizó correcciones menos conservadoras, los resultados obtenidos fueron los siguientes:

Paso 1: Planteamiento de Hipótesis

$\mathbf{H_o}$: $No\ hay\ diferencia\ entre\ la\ memoria\ de\ las\ 2\ poblaciones\ de\ técnicas\ de\ reglas\ de\ asociación$

$\mathbf{H_1}$: $Hay\ diferencia\ entre\ la\ memoria\ de\ las\ 2\ poblaciones\ de\ técnicas\ de\ reglas\ de\ asociación$

Paso 2: Nivel de significancia α=0,05 (dividido para el número de comparaciones)

Paso 3A: Estadístico de Prueba

Tabla 8.11.- Comparaciones múltiples Wilcoxon (Bonferroni)

	met_apriori	met_ASI	met_eclat
met_ASI	2,5e-07	-	-
met_eclat	1	3,2e-08	-
met_weclat	1	1,4e-08	1

Se muestran los resultados obtenidos al aplicar la prueba no paramétrica Wilcoxon mediante la corrección de Bonferroni (Tabla 8.11).

Paso 4A: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5A: Decisión

Se calcularon las comparaciones por pares entre niveles de grupo con correcciones para pruebas múltiples usando el método de Bonferroni en donde se obtuvo que los pares met_ASI - met_apriori, met_eclat – met_ASI y met_weclat – met_ASI son significativamente diferentes dado que el valor p obtenido para cada par es 2,5e-07, 3,2e-08, 1,4e-08 respectivamente, los cuales son menores a un nivel de significancia de 0,05 (Tabla 8.11).

Paso 3B: Comparaciones múltiples, estadístico de prueba, Wilcoxon (Holm)

Se muestran los resultados obtenidos al aplicar la prueba no paramétrica Wilcoxon, en este caso se usó la técnica de Holm (Tabla 8.12).

Tabla 8.12.- Comparaciones múltiples Wilcoxon (Holm)

	met_apriori	met_ASI	met_eclat
met_ASI	1,7e-07	-	-
met_eclat	1	2,7e-08	-
met_weclat	1	1,4e-08	1

Paso 4B: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5B: Decisión

Las comparaciones por pares usando el método de Holm dio como resultado que el met_ASI – met_apriori, met_eclat - met_ASI, met_weclat – met_ASI obtuvieron un valor p igual a 1,7e-07, 2,7e-08,1,4e-08 respectivamente, dichos valores son menores a un nivel de significancia de 0,05 por lo cual se concluye que estas técnicas poseen medianas significativamente diferentes.

8.3.6.4 ANOVA no paramétrico

Paso 1: Planteamiento de Hipótesis

H_o: $No\ hay\ diferencia\ entre\ la\ memoria\ de\ las\ 4\ poblaciones\ de\ técnicas\ de\ reglas\ de\ asociación$

H_1: $Hay\ diferencia\ entre\ la\ memoria\ de\ las\ 4\ poblaciones\ de\ técnicas\ de\ reglas\ de\ asociación$

Paso 2: Nivel de significancia α=0,05

Paso 3A: Estadístico de Prueba

F-Statistic=2,1501e+01; p-value=6,8834e-14

Paso 4A: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza (Figura 8.7).

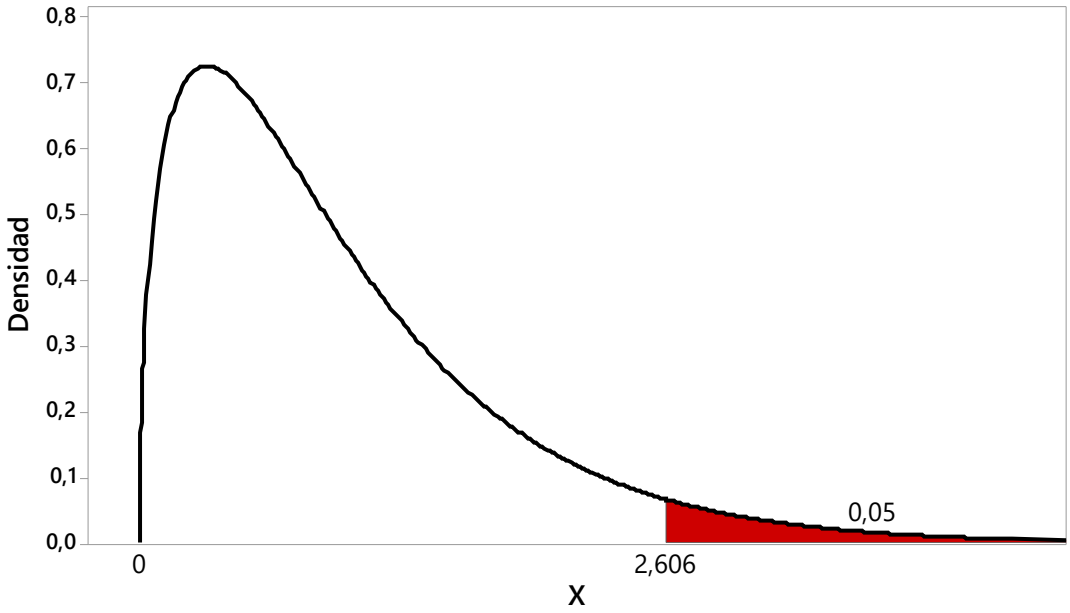


Figura 8.7.- Zonas de rechazo y aceptación ANOVA no paramétrico, complejidad espacial, reglas de asociación

Paso 5A: Decisión

Se obtuvo un valor p igual a 6,8834e-14 el cual es menor a un nivel de significancia de 0,05 por lo cual se rechaza la hipótesis nula y se concluye que hay al menos un grupo estadísticamente diferente de los otros grupos en el conjunto de datos en estudio.

Paso 3B: Estadístico de Prueba

Tabla 8.13.- Comparaciones múltiples ANOVA no paramétrico

	met_apriori	met_ASI	met_eclat
met_ASI	1,7e-12	-	-
met_eclat	0,61	3,8e-14	-
met_weclat	0,61	3,8e-14	1,00

Se muestran los resultados obtenidos al aplicar el ANOVA no paramétrico, evidenciándose así valores significativos y no significativos para el nivel de significancia empleado en esta prueba (Tabla 8.13).

Paso 4B: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5B: Decisión

Mediante el método ANOVA, se obtuvo que los pares met_eclat – met_apriori, met_weclat – met_apriori obtuvieron un p-value igual a 0,61 para ambos casos mencionados, dichos valores son mayores a un nivel de significancia de 0,05 por lo cual se concluye que estas técnicas poseen medias iguales, con respecto al par met_weclat – met-ASI se obtuvo un p-value de 3,8e-14 el cual es menor para un nivel de significancia de 0,05 por lo que se concluye que al menos uno posee una media significativamente diferente, dicho resultado obtenido se asemeja a los resultados dados con otras técnicas en análisis.

Tukey

Mediante la opción de Tukey se determinó las técnicas de reglas de asociación que son significativamente diferentes con respecto a la memoria que usa cada uno, para lo cual se plantearon las siguientes hipótesis:

Paso 1: Planteamiento de Hipótesis

$\mathbf{H_o}$: $No\ hay\ diferencia\ entre\ la\ memoria\ de\ las\ 2\ poblaciones\ de\ técnicas\ de\ reglas\ de\ asociación$

$\mathbf{H_1}$: $Hay\ diferencia\ entre\ la\ memoria\ de\ las\ 2\ poblaciones\ de\ técnicas\ de\ reglas\ de\ asociación$

Paso 2: Nivel de significancia α=0,05

Paso 3A: Estadístico de Prueba

Tabla 8.14.- Comparaciones múltiples (Tukey)

	I	J	Estimate	St Err	Lower Bound CI	Upper Bound CI
1	met_apriori	met_ASI	1,37199	0,19429	0,8728	1,87119
2	met_apriori	met_eclat	-0,09989	0,19429	-0,59909	0,39931
3	met_apriori	met_weclat	-0,10016	0,19429	-0,59936	0,39904
4	met_ASI	met_eclat	1,47189	0,19429	0,97269	1,97108
5	met_ASI	met_weclat	1,47216	0,19429	0,97296	1,97136
6	met_eclat	met_weclat	0,00027	0,19429	-0,49893	0,49947

La Tabla 8.14 muestra las comparaciones múltiples entre las diferentes técnicas de reglas de asociación.

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5: Decisión

Los resultados obtenidos indican que cada una de las técnicas en estudio no son significativas con respecto a la cantidad de memoria que usan, dado que se puede visualizar esta aseveración en los intervalos de confianza que están dados por valores negativos y positivos.

8.4 Estudio de la complejidad temporal

A continuación, se realiza el estudio de complejidad temporal respecto a las reglas de asociación. Primeramente, se realiza un estudio descriptivo, para luego realizar las pruebas de hipótesis, no sin antes realizar la prueba de los supuestos de normalidad, homocedasticidad e independencia para determinar el tipo de prueba a utilizar.

8.4.1 Medidas descriptivas

Las siguientes tablas resumen los resultados descriptivos del tiempo de ejecución por técnica de reglas de asociación utilizada en el análisis.

La Tabla 8.15, presenta las principales medidas de centralización, dispersión y forma. Se reflejan los resultados sobre la cantidad de tiempo que se empleó por cada técnica de asociación, llegándose a obtener distintos parámetros de análisis para cada uno. La primera columna refleja el valor de la media en donde se evidenció que en promedio la técnica que ocupa menos tiempo es el eclat con un valor de 0,982917 y la técnica que en promedio usa mayor tiempo es el ASI con un valor igual a 23,085733. Con respecto a que tan dispersos se encuentran dichos datos analizados con respecto al valor promedio que presentaron se obtuvo que la técnica que presenta menor dispersión es eclat con un valor igual a 6,106949 y la técnica con mayor dispersión es el ASI con un valor de 28,436968. El rango intercuartil (IQR) da a notar que la técnica que presenta menor variación en el tiempo es eclat con un valor igual a 0,3498121 mientras que el método con mayor variación es su tiempo es el ASI con un valor igual a 48,1853426; se nota que la técnica que posee menor coeficiente de variación (cv) es ASI con 1,231798% de dispersión respecto a la media mientras que la técnica con mayor coeficiente de variación es eclat con un 6,213087%.

Tabla 8.15.- Medidas descriptivas de cantidad de tiempo por técnica de regla de asociación

ASSOCIATION RULES TECHNIQUES	mean	sd	IQR	cv	skewness	kurtosis
met_apriori	2,205058	12,526872	0,404452	5,680972	9,493657	102,7038195
met_ASI	23,085733	28,436968	48,1853426	1,231798	0,985527	-0,7405681
met_eclat	0,982917	6,106949	0,3498121	6,213087	14,300674	239,489486
met_weclat	2,475675	13,010339	0,8339193	5,255268	11,709648	155,0926409

La Tabla 8.16, presenta los cuartiles incluida la mediana (50%), que dividen al grupo de datos en cuatro partes iguales.

Tabla 8.16.- Cuartiles de cantidad de tiempo por método de asociación

ASSOCIATION RULES TECHNIQUES	0%	25%	50%	75%	100%	TIME:n
met_apriori	0,04118032	0,11458315	0,2479032	0,5190351	204,4119	3447
met_ASI	0,0720016	2,89824595	6,0061767	51,0835885	114,6095	3447
met_eclat	0,03404434	0,09100568	0,2003153	0,4408178	115,9971	3447
met_weclat	0,03475832	0,30504326	0,5785784	1,1389626	212,6889	3447

En la Tabla 8.16 se refleja la asimetría de los datos (*skewness*) permitió notar que todas las reglas de asociación poseen una distribución simétrica, es decir que existe aproximadamente la misma cantidad de datos a los dos lados del valor de la media correspondiente, en cuanto a qué técnica de regla de asociación posee mayor asimetría es el eclat con 14,300674 y con menor asimetría es el ASI con 0,985527. El coeficiente de *kurtosis* obtenido para cada una de las técnicas muestra que la distribución que siguen las técnicas apriori, eclat y weclat es leptocúrtica dado que dichos valores son positivos, lo que quiere decir que hay una mayor concentración de los datos en torno a la media, a excepción de la técnica ASI que sigue una distribución platicúrtica dado que sus valores son negativos, se obtuvo que el método ASI posee menor coeficiente con un valor igual a –0,7405681 y el eclat con un coeficiente mayor igual a 239,489486. Se obtuvieron los cuartiles para cada técnica de regla de asociación, el primer cuartil con mayor valor es para la técnica ASI la cual muestra que el 25% del tiempo es menor o igual a 2,89824595 con una mediana la cual indica que la mitad del tiempo es menor o igual a 6,0061767 y la otra mitad es mayor o igual a 6,0061767 la cual corresponde al método ASI que es el mayor valor , finalmente el tercer cuartil indica que el 75% de los datos es menor o igual a 51,0835885 que también corresponde al método ASI que presentó el valor más alto. Todas las técnicas de reglas de asociación se trabajaron con una muestra de 3447 datos. Mediante el gráfico BoxPlot se observa la dispersión de cada una de las técnicas (Figura 8.8).

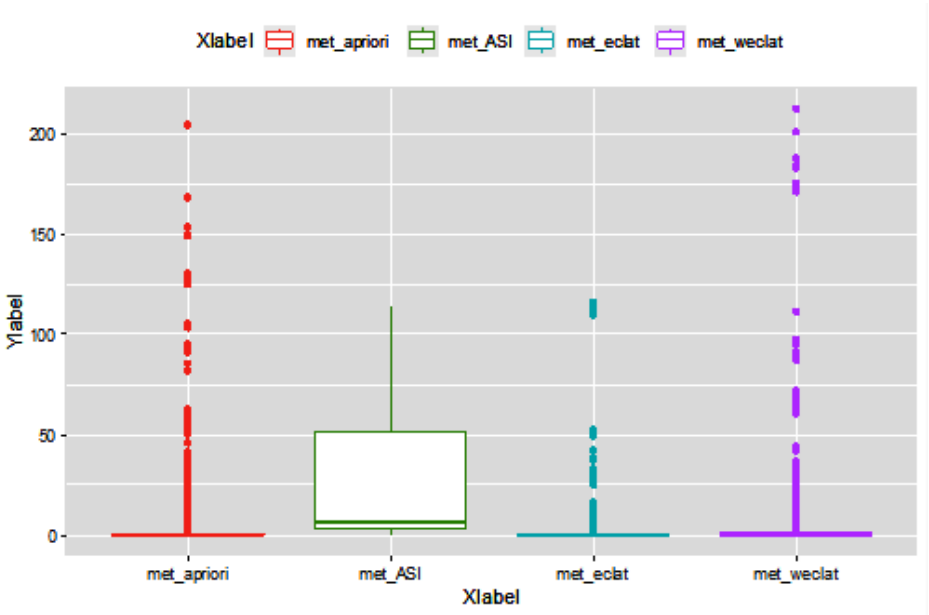


Figura 8.8.- Gráfico BoxPlot sobre el tiempo por técnica de asociación

La Tabla 8.17, presenta las principales medidas de centralización, dispersión y forma.

Tabla 8.17.- Medidas descriptivas del tiempo en las distintas técnicas de reglas de asociación

ASSOCIATION RULES TECHNIQUES	mean	sd	IQR	cv	skewness	kurtosis
met_apriori	2,205058	12,526872	0,404452	5,680972	9,493657	102,7038195
met_ASI	23,085733	28,436968	48,1853426	1,231798	0,985527	-0,7405681
met_eclat	0,982917	6,106949	0,3498121	6,213087	14,300674	239,489486
met_weclat	2,475675	13,010339	0,8339193	5,255268	11,709648	155,0926409

Se realizó el análisis descriptivo sobre la cantidad de tiempo que se empleó por cada técnica de reglas de asociación, llegándose a obtener distintos parámetros de análisis para cada una. En promedio la técnica que ocupa menos tiempo es eclat con un valor de 0,982917 y la técnica que en promedio usa mayor tiempo es el ASI con un valor igual a 23,085733, con respecto a qué tan dispersos se encuentran dichos datos analizados con respecto al valor promedio que presentaron se obtuvo que la técnica que presenta menor dispersión es eclat con un valor igual a 6,106949 y la técnica con mayor dispersión es el ASI con un valor de 28,436968. El rango intercuartil (IQR) da a notar que la técnica que presenta menor variación en el tiempo es eclat con un valor igual a 0,3498121 mientras que la técnica con mayor variación en el tiempo es la correspondiente al ASI con un valor de 48,1853426; se nota que la técnica que posee menor coeficiente de variación (cv) es el ASI con 1,231798% de dispersión respecto a la media mientras que la técnica con mayor coeficiente de variación es eclat con 6,213087%.

La asimetría de los datos (*skewness*) permitió notar que todas las reglas de asociación poseen una distribución simétrica, es decir que existe aproximadamente la misma cantidad de datos a los dos lados del valor de la media correspondiente, en cuanto a qué técnica de regla de asociación posee mayor asimetría es eclat con 14,300674 y con menor asimetría es el ASI con 0,985527. El coeficiente de *kurtosis* obtenido para cada una de las técnicas muestra que la distribución que siguen apriori, eclat y weclat es leptocúrtica dado que dichos valores son positivos, lo que quiere decir que hay una mayor concentración de los datos en torno a la media, a excepción de la técnica ASI que sigue una distribución platicúrtica dado que sus valores son negativos, se obtuvo que la técnica del ASI posee menor coeficiente con un valor de –0,7405681 y eclat con un coeficiente

mayor de 239,489486. Todas las técnicas de reglas de asociación se trabajaron con una muestra de 3447 datos (Tabla 8.17).

La Tabla 8.18, presenta los cuartiles incluida la mediana (50%), que dividen al grupo de datos en cuatro partes iguales.

Tabla 8.18.- Cuartiles del tiempo en las distintas técnicas de reglas de asociación

ASSOCIATION RULES METHODS	0%	25%	50%	75%	100%	TIME: n
met_apriori	0,04118032	0,11458315	0,2479032	0,5190351	204,4119	3447
met_ASI	0,0720016	2,89824595	6,0061767	51,0835885	114,6095	3447
met_eclat	0,03404434	0,09100568	0,2003153	0,4408178	115,9971	3447
met_weclat	0,03475832	0,30504326	0,5785784	1,1389626	212,6889	3447

8.4.2 Normalidad

La Figura 8.9, de aproximación normal de los datos del tiempo de procesamiento en las reglas de asociación evidencia que el histograma es asimétrico hacia la derecha lo que quiere decir que los datos muestran un sesgo positivo, con respecto a la línea de distribución ajustada se nota que las barras no la siguen por lo que no parece ofrecer un ajuste adecuado para una distribución normal, hay mayor concentración de datos en la cola derecha.

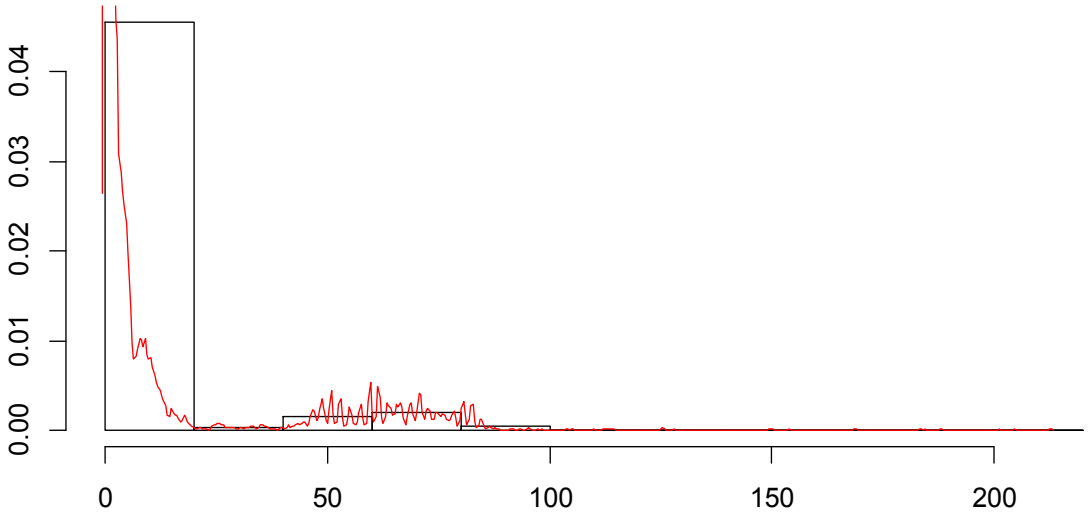


Figura 8.9.- Aproximación normal de los datos de tiempo

Al analizar la cercanía de la recta a la curva (Figura 8.10) sobre el tiempo de procesamiento usado por cada una de las técnicas de reglas de asociación se corroboró que cierta cantidad de puntos están situados en la línea recta mientras que otros no, llevando a intuir que probablemente los datos no siguen una distribución normal, por lo cual, para verificar dicha aseveración se realizaron los respectivos test de normalidad a niveles de significancia de 0,01, 0,05 y 0,1.

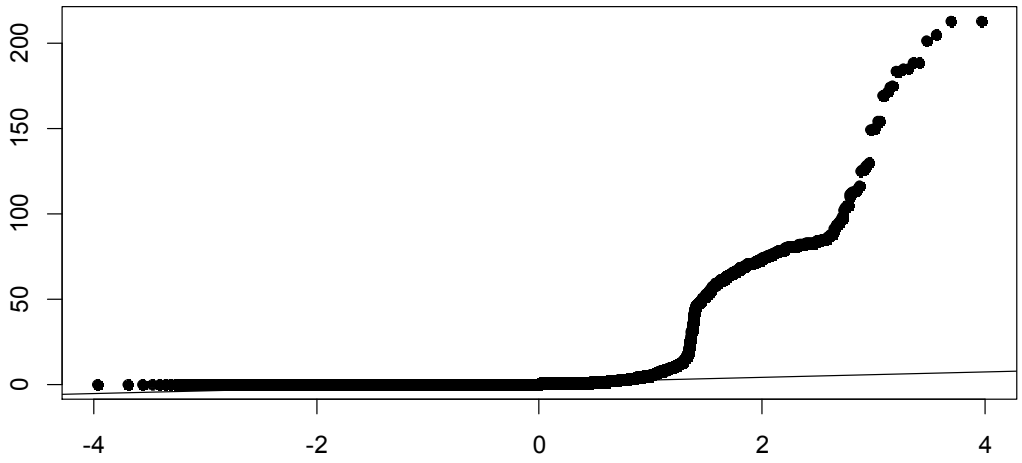


Figura 8.10.- Gráfico QQ para los datos de tiempo

La Tabla 8.19 muestra los resultados de distintas pruebas de normalidad aplicadas a los datos.

8.4.2.1 Paso 1A (α=0,01), 1B (α=0,05) y 1C (α=0,1): Planteamiento de Hipótesis

H_o: Ocupación de memoria ~ $N(\mu, \sigma^2)$
H_1: Ocupación de memoria $\nsim N(\mu, \sigma^2)$

8.4.2.2 Paso 2A (α=0,01), 2B (α=0,05) y 2C (α=0,1): Nivel de significancia

α=0,01, α=0,05 y α=0,1

8.4.2.3 Paso 3A (α=0,01), 3B (α=0,05) y 3C (α=0,1): Estadístico y valor p

La Tabla 8.19 muestra todos los resultados obtenidos de cada prueba de normalidad, se visualiza el valor del estadístico y el valor p para tomar su respectiva decisión de acuerdo con su nivel de significancia.

Tabla 8.19.- Resultados de las pruebas de normalidad, reglas de asociación y variable tiempo: estadístico y valor p

	met_apriori	met_ASI	met_eclat	met_weclat
Anderson-Darling normality test	A = 1125,6, p-value < 2,2e-16	A = 443,5, p-value < 2,2e-16	A = 1106,2, p-value < 2,2e-16	A = 1053,6, p-value < 2,2e-16
Holm technique	met_apriori < 2,22e-16 < 2,22e-16	met_ASI < 2,22e-16 < 2,22e-16	met_eclat < 2,22e-16 < 2,22e-16	met_weclat < 2,22e-16 < 2,22e-16
Lilliefors (Kolmogorov-Smirnov) normality test	D = 0,45609, p-value < 2,2e-16	D = 0,3091, p-value < 2,2e-16	D = 0,43858, p-value < 2,2e-16	D = 0,42559, p-value < 2,2e-16
Holm technique	met_apriori < 2,22e-16 < 2,22e-16	met_ASI < 2,22e-16 < 2,22e-16	met_eclat < 2,22e-16 < 2,22e-16	met_weclat < 2,22e-16 < 2,22e-16
Cramer-von Mises normality test	W = 241,34, p-value = 7,37e-10	W = 85,434, p-value = 7,37e-10	W = 235,77, p-value = 7,37e-10	W = 222,52, p-value = 7,37e-10
Holm technique	met_apriori 7,37e-10 2,948e-09	met_ASI 7,37e-10 2,948e-09	met_eclat 7,37e-10 2,948e-09	met_weclat 7,37e-10 2,948e-09
Pearson chi-square normality test	P = 73370, p-value < 2,2e-16	P = 13006, p-value < 2,2e-16	P = 70977, p-value < 2,2e-16	P = 45690, p-value < 2,2e-16
p-values adjusted by the Holm technique	met_apriori < 2,22e-16 < 2,22e-16	met_ASI < 2,22e-16 < 2,22e-16	met_eclat < 2,22e-16 < 2,22e-16	met_weclat < 2,22e-16 < 2,22e-16
Shapiro-Wilk normality test	W = 0,14963, p-value < 2,2e-16	W = 0,72545, p-value < 2,2e-16	W = 0,11161, p-value < 2,2e-16	W = 0,1383, p-value < 2,2e-16
p-values adjusted by the Holm technique	met_apriori < 2,22e-16 < 2,22e-16	met_ASI < 2,22e-16 < 2,22e-16	met_eclat < 2,22e-16 < 2,22e-16	met_weclat < 2,22e-16 < 2,22e-16
Shapiro-Francia normality test	W = 0,14886, p-value < 2,2e-16	W = 0,72573, p-value < 2,2e-16	W = 0,11078, p-value < 2,2e-16	W = 0,13754, p-value < 2,2e-16
p-values adjusted by the Holm technique	met_apriori < 2,22e-16 < 2,22e-16	met_ASI < 2,22e-16 < 2,22e-16	met_eclat < 2,22e-16 < 2,22e-16	met_weclat < 2,22e-16 < 2,22e-16

8.4.2.4 Paso 4A: Regla de decisión para α=0,01

Si el p valor es menor que 0,01 (p-value < 0,01) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla.

8.4.2.5 Paso 4B: Regla de decisión para α=0,05

Si el p valor es menor que 0,05 (p-value < 0,05) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla.

8.4.2.6 Paso 4C: Regla de decisión para α=0,1

Si el p valor es menor que 0,1 (p-value < 0,1) se rechaza la hipótesis nula Ho, caso contrario no existe evidencia suficiente para rechazarla.

8.4.2.7 Paso 5C: Tabla de resultados α=0,01

La Tabla 8.20 muestra los resultados de normalidad para α=0,01 para cada una de las once pruebas de normalidad utilizadas.

Tabla 8.20.- Resultados de la normalidad de la variable tiempo para un valor de α=0,01

	met_apriori	met_ASI	met_eclat	met_weclat
Anderson-Darling	No son normales	No son normales	No son normales	No son normales
Holm technique	No son normales	No son normales	No son normales	No son normales
Lilliefors (Kolmogorov-Smirnov)	No son normales	No son normales	No son normales	No son normales
Cramer-von Mises	No son normales	No son normales	No son normales	No son normales
Holm technique	No son normales	No son normales	No son normales	No son normales
Pearson chi-square	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales
Shapiro-Wilk	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales
Shapiro-Francia	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales

8.4.2.8 Paso 5A: Tabla de resultados α=0,05

Al analizar los valores de la Tabla 8.19, a partir de la variable tiempo en los pasos 4A, 4B y 4C se evidenció que los p-values obtenidos en las distintas pruebas de normalidad aplicadas son demasiado pequeños, por lo cual caen en la zona de rechazo tanto en el nivel de significancia de 0,01, 0,05 y 0,1. De manera más detallada se tuvo que para el estadístico de Anderson Darling el p-value es 2,2e-16 el cual es menor a los niveles de significancia 0,01, 0,05 y 0,1; rechazándose la hipótesis nula de que la variable tiempo siga una distribución normal. En el caso de Holm se obtuvo también un valor de 2,22e-16

para todas las técnicas de reglas de asociación en estudio, por lo cual también se rechaza la hipótesis nula dado que es menor al valor de significancia de 0,05, 0,01 y 0,1, no existe normalidad en los datos de la variable tiempo, dicha situación se repite con la prueba de Kolmogorov Smirnov con la corrección de Lilliefors la cual posee un valor p de 2,22e-16 para todas las técnicas. Al analizar el estadístico Cramer-von Mises se evidenció como resultado un p-value igual a 7,37e-10, este valor es inferior a los niveles de significancia de 0,05, 0,01 y 0,1 por lo cual se rechaza la hipótesis nula, concluyéndose que se observa diferencia entre los datos de la variable tiempo y la distribución normal.

Tabla 8.21.- Resultados de la normalidad de la variable tiempo para un valor de α=0,05

	met_apriori	met_ASI	met_eclat	met_weclat
Anderson-Darling normality test	No son normales	No son normales	No son normales	No son normales
Holm technique	No son normales	No son normales	No son normales	No son normales
Lilliefors (Kolmogorov-Smirnov) normality test	No son normales	No son normales	No son normales	No son normales
Holm technique	No son normales	No son normales	No son normales	No son normales
Cramer-von Mises normality test	No son normales	No son normales	No son normales	No son normales
Holm technique	No son normales	No son normales	No son normales	No son normales
Pearson chi-square normality test	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales
Shapiro-Wilk normality test	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales
Shapiro-Francia normality test	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales

La técnica de Holm que es con la cual se ajustan los p-values dio como resultado un p-value final igual a 2,948e-09 para todos las técnicas de reglas de asociación, este valor también es menor a 0,05, 0,01 y 0,1 por lo que se rechazó la hipótesis nula y se concluyó que los datos no siguen una distribución normal, con respecto a la prueba chi cuadrada de Pearson, p-valores ajustados por la técnica de Holm, Shapiro Wilk y finalmente Shapiro

Francia también se obtuvo el mismo p-value de 2,2e-16 siendo menor a los niveles de significancia planteados, por lo que se rechazó la hipótesis nula y se concluye que se observa diferencia entre los datos de la variable tiempo y la distribución normal. La Tabla 8.21 muestra los resultados de normalidad para α=0,05.

8.4.2.9 Paso 5B: Tabla de resultados α=0,1

La Tabla 8.22 muestra los resultados de normalidad para α=0,1, para cada uno de los 12 métodos de normalidad utilizados.

Tabla 8.22.- Resultados de la normalidad de la variable tiempo para un valor de α=0,1

	met_apriori	met_ASI	met_eclat	met_weclat
Anderson-Darling normality test	No son normales	No son normales	No son normales	No son normales
Holm technique	No son normales	No son normales	No son normales	No son normales
Lilliefors (Kolmogorov-Smirnov) normality test	No son normales	No son normales	No son normales	No son normales
Holm technique	No son normales	No son normales	No son normales	No son normales
Cramer-von Mises normality test	No son normales	No son normales	No son normales	No son normales
Holm technique	No son normales	No son normales	No son normales	No son normales
Pearson chi-square normality test	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales
Shapiro-Wilk normality test	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales
Shapiro-Francia normality test	No son normales	No son normales	No son normales	No son normales
p-values adjusted by the Holm technique	No son normales	No son normales	No son normales	No son normales

La Tabla 8.20, Tabla 8.21 y Tabla 8.22 representan todos los resultados obtenidos sobre las pruebas de normalidad en donde se verificaba si los datos cumplen o no el supuesto de normalidad, obteniéndose así que para ninguna de las pruebas en estudio se cumplió con dicho supuesto ya que todos los p-values obtenidos resultaron menores a los niveles de significancia de 0,01, 0,05 y 0,1.

8.4.3 Normalización

En busca de la transformación de los datos de la variable tiempo a una distribución normal se usó la función en R llamada powerTransform, con el objetivo de determinar la potencia óptima a la que se debe elevar la variable de interés y así buscar obtener normalidad en los datos.

Tabla 8.23.- Normalización utilizando bcPower, reglas de asociación, complejidad temporal

# NORMALITY TRANSFORMS BY GROUPS
summary(powerTransform(Y ~ X, family="bcPower"))
bcPower Transformation to Normality

	Est Power	Rounded Pwr Wald	Lwr Bnd Wald	Upr Bnd
Y1	-0,2254	-0,23	-0,2345	-0,2162

Likelihood ratio test that transformation parameter is equal to 0
(log transformation)

		LRT	df	pval
LR test	lambda = (0)	2536,571	1	< 2,22e-16

Likelihood ratio test that no transformation is needed

		LRT	df	pval
LR test	lambda = (1)	72218,01	1	< 2,22e-16

Dado que el valor EstPower (lambda) obtenido en la prueba es negativo (-0,2254) se lo toma como cero, obteniéndose así un valor p igual a 2,22e-16 el mismo que es menor a un nivel de significancia de 0,05 por lo cual se concluyó que al asignar una lambda igual a 0 a los datos de la variable tiempo, éstos no siguen una distribución normal, lo mismo sucede cuando lambda es igual a 1 ya que también se obtiene un p-value pequeño (2,22e-16).

Tabla 8.24.- Normalización utilizando yjPower, reglas de asociación, complejidad temporal

summary(powerTransform(Y ~ X, family="yjPower"))
yjPower Transformation to Normality

	Est Power	Rounded Pwr Wald	Lwr Bnd Wald	Upr Bnd
Y1	-0,9155	-0,92	-0,9369	-0,8942

Likelihood ratio test that transformation parameter is equal to 0

		LRT	df	pval
LR test	lambda = (0)	10780,44	1	< 2,22e-16

Se usó el método yjPower dando como resultado un valor para EstPower (lambda) igual a -0,9155 el cual es negativo por lo que se lo toma como cero, el valor p obtenido con

lambda igual a 0 es 2,22e-16 el cual es menor a un nivel de significancia de 0,05 por lo que se concluyó que al asignar una lambda igual a 0 a los datos de la variable tiempo, éstos no siguen una distribución normal.

8.4.4 Homocedasticidad

8.4.4.1 Test de Levene

Paso 1: Planteamiento de Hipótesis

H$_o$: $\sigma^2_{met\_apriori} = \sigma^2_{met\_eclat} = \sigma^2_{met\_weclat} = \sigma^2_{met\_ASI}$

H$_1$:$\exists i, j \in \{met\_apriori, met\_eclat, met\_weclat, met\_ASI\}\ tal\ que\ i \neq j,\ \sigma_i^2 \neq \sigma_j^2$

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

Group=13785; Df=2; Fvalue=1176,5; Pr(>F) < 2,2e-16 ***

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

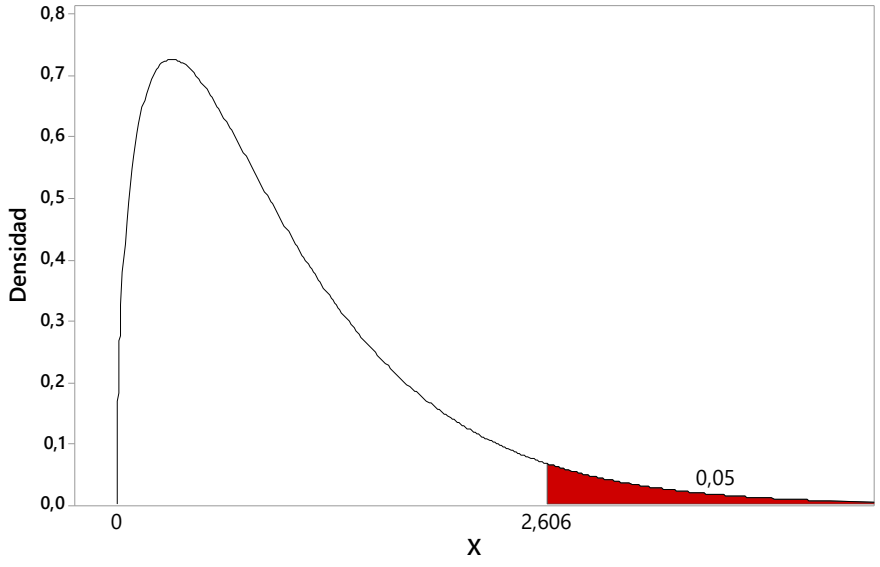


Figura 8.11.- Zonas de rechazo y aceptación para la homogeneidad de varianzas, complejidad temporal, reglas de asociación

Paso 5: Decisión

Dado un nivel de significancia del 5% se obtuvo un p-value igual a 2,2e-16 siendo menor, por lo cual se rechaza Ho y se concluye que las varianzas de la variable tiempo no son iguales es decir son heterocedásticos (Figura 8.11).

8.4.4.2 Test de Bartlett

Paso 1: Planteamiento de Hipótesis

H_o: $\sigma^2_{met\_apriori} = \sigma^2_{met\_eclat} = \sigma^2_{met\_weclat} = \sigma^2_{met\_ASI}$

H_1:$\exists i, j \in \{met\_apriori, met\_eclat, met\_weclat, met\_ASI\}\ tal\ que\ i \neq j,\ \sigma_i^2 \neq \sigma_j^2$

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

Bartlett's K-squared=20520, df=2, p-value < 2,2e-16

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5: Decisión

Mediante la prueba de Bartlett se obtuvo un p-value igual a 2,2e-16 el mismo que es menor a un nivel de significancia de 0,05 por lo que se rechaza Ho y se concluye que la varianza de los grupos en estudio de la variable tiempo es diferente. Comparativamente entre las pruebas de Bartlett y Levene's se nota que se llega a la misma conclusión.

8.4.5 Independencia

Paso 1: Planteamiento de Hipótesis

H_o: Técnicas de reglas de asociación y tiempo son independientes

H_1: Técnicas de reglas de asociación y tiempo no son independientes

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

X-squared=3072,7, df=6, p-value < 2,2e-16

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

X-squared=3072,7, df=6, p-value < 2,2e-16

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

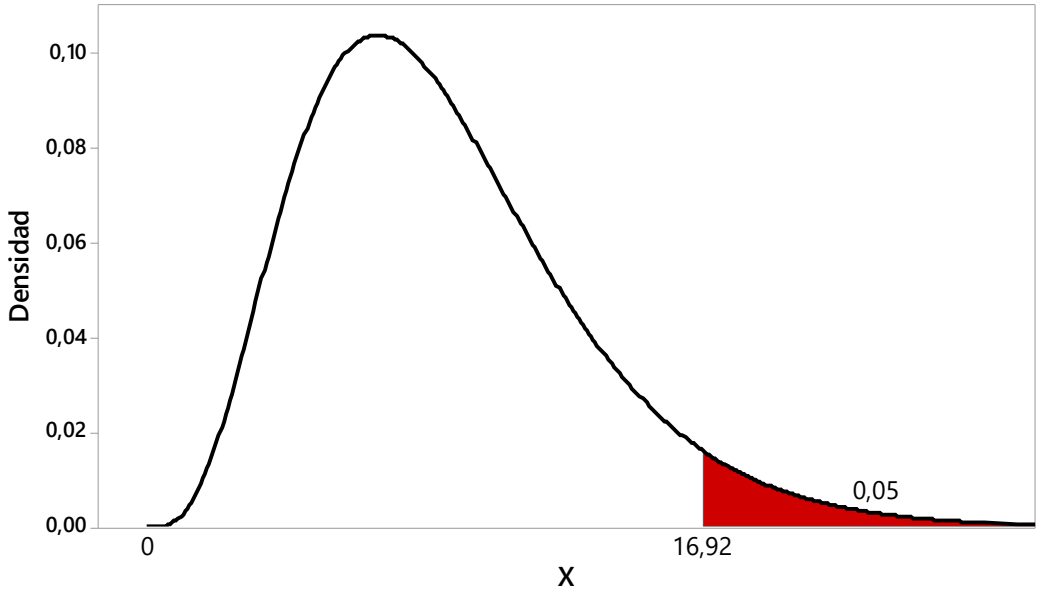


Figura 8.12.- Zonas de rechazo y aceptación para el prerrequisito de independencia

Paso 5: Decisión

El valor p obtenido en la prueba es igual a 2,2e-16 siendo menor al nivel de significancia de 0,05 por lo que se rechaza Ho y se concluye que las reglas de asociación y el tiempo no son independientes (Figura 8.12).

8.4.6 Pruebas de hipótesis

Una vez analizados los prerrequisitos se concluyó que no se cumple normalidad ni independencia, ni tampoco se puede lograr la misma normalidad mediante las

transformaciones realizadas, se debe realizar pruebas no paramétricas para muestras independientes. Para todas las pruebas de hipótesis se sigue el método de los 5 pasos propuesto por (Douglas y Marchal, 2018).

8.4.6.1 Medidas descriptivas específicas

Se presentan los resultados sobre la cantidad de memoria que se empleó por cada técnica de reglas de asociación, obteniéndose así distintos parámetros de análisis para cada una.

La Tabla 8.25 refleja los resultados sobre la cantidad de tiempo que se empleó por cada técnica de reglas de asociación, obteniéndose así distintos parámetros de análisis para cada una.

Tabla 8.25.- Medidas descriptivas de las técnicas de reglas de asociación en el tiempo

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
met_apriori	0,04118	0,11458	0,24790	2,20506	0,51904	204,41186
met_ASI	0,072	2,898	6,006	23,086	51,084	114,610
met_eclat	0,03404	0,09101	0,20032	0,98292	0,44082	115,99705
met_weclat	0,3476	0,30504	0,57858	2,47568	1,13896	212,68886

La técnica que usa en promedio mayor tiempo en el sistema operativo es el ASI con un valor igual a 23,086 con un mínimo de tiempo igual a 0,072 y un máximo de 114,610, el primer cuartil muestra que el 25% del tiempo es menor o igual a 2,898 con una mediana que indica que la mitad de tiempo empleado para la técnica es menor o igual a 6,006 y la otra mitad es mayor o igual a 6,006, por último, el tercer cuartil refleja que el 75% del tiempo es menor o igual a 51,084. La técnica que utiliza menos tiempo es eclat con un valor promedio igual a 0,98292 con un mínimo de tiempo igual a 0,03404 y un máximo de 115,99705, el primer cuartil muestra que el 25% del tiempo es menor o igual a 0,09101 con una mediana que indica que la mitad de tiempo empleado para el método es menor o igual a 0,20032 y la otra mitad es mayor o igual a 0,20032, el tercer cuartil refleja que el 75% del tiempo es menor o igual a 0,44082. Se determinó que para las técnicas en estudio no existe mayor diferencia en el tiempo promedio utilizado ya que para la técnica apriori es 2,20506 y para la técnica weclat es 2,47568. Finalmente se estableció que para las técnicas en estudio no existe mayor diferencia en el tiempo promedio utilizado, ya que para apriori es 2,20506 y para la técnica weclat es 2,47568.

8.4.6.2 Kruskal-Wallis H test

Paso 1: Planteamiento de Hipótesis

H_o: $\tilde{\mu}_{APRIORI} = \tilde{\mu}_{ASI} = \tilde{\mu}_{ECLAT} = \tilde{\mu}_{WECLAT} = \tilde{\mu}_{tiempo}$

H_1: $\tilde{\mu}_i \neq \tilde{\mu}_j \ para\ al\ menos\ un\ par\ de\ (i, j)$

Paso 2: Nivel de significancia α=0,05

Paso 3: Estadístico de Prueba

Kruskal-Wallis chi-squared=7338,4, df=3, p-value < 2,2e-16

Paso 4: Regla de Decisión

Si p-value < 0,05 entonces se rechaza Ho, caso contrario no se rechaza. En nuestro caso 7,815<7338,4 por tanto, se rechaza la hipótesis nula.

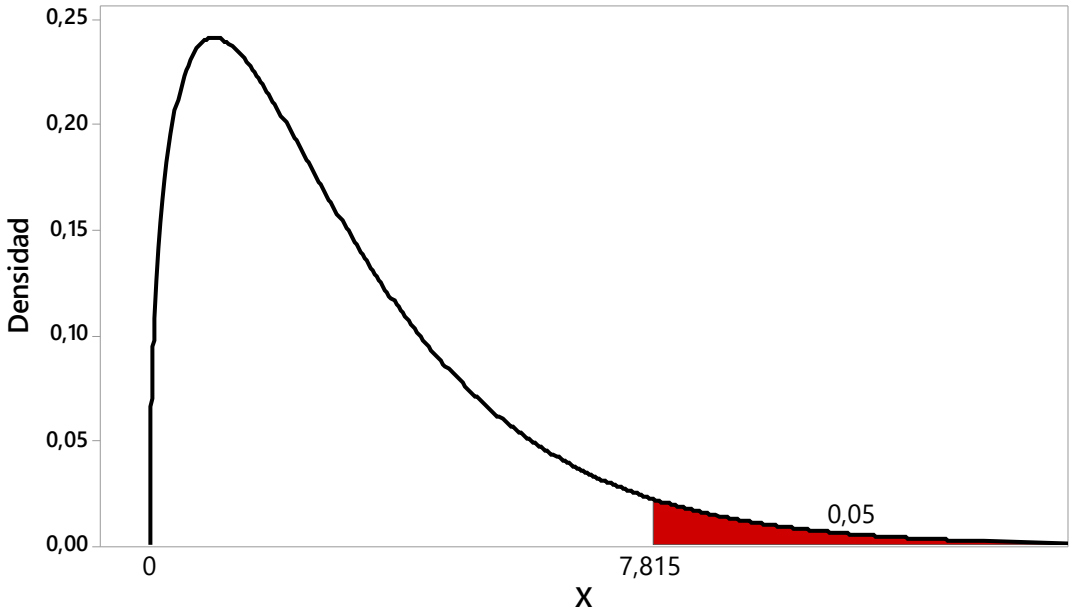


Figura 8.13.- Zonas de rechazo y aceptación para la prueba de Kruskal Wallis, factor técnica de reglas de asociación.

Paso 5: Decisión

Como el valor p es igual a 2,2e-16 el cual es menor a un nivel de significancia de 0,05 se concluye que existen diferencias significativas entre los tiempos de las diferentes técnicas de reglas de asociación (Figura 8.13).

Para determinar cuál par de rangos son diferentes se utilizó la prueba no paramétrica para comparación de pares de dos muestras independientes de Mann Whitney Wilcoxon U-test.

8.4.6.3 Mann Whitney Wilcoxon U-test

Se utilizó Mann Whitney Wilcoxon U-test para la comparación por parejas independientes como se puede ver en la página oficial de R (*R: Pairwise Wilcoxon Rank Sum Tests*, 2021). Para no provocar que el error de Tipo 1 aumente, las técnicas usadas mediante R fueron la corrección de Bonferroni en la cual los valores p se multiplican por el número de comparaciones y Holm que realizó correcciones menos conservadoras, los resultados obtenidos fueron los siguientes:

Paso 1: Planteamiento de Hipótesis

$\mathbf{H_o}$: $No\ hay\ diferencia\ entre\ los\ tiempos\ de\ las\ 2\ poblaciones\ de\ técnicas\ de\ reglas\ de\ asociación$

$\mathbf{H_1}$: $Hay\ diferencia\ entre\ los\ tiempos\ de\ las\ 2\ poblaciones\ de\ técnicas\ de\ reglas\ de\ asociación$

Paso 2: Nivel de significancia α=0,05 (dividido para el número de comparaciones)

Paso 3A: Estadístico de Prueba

Tabla 8.26.- Comparaciones múltiples Wilcoxon (Bonferroni)

	met_apriori	met_ASI	met_eclat
met_ASI	<2e-16	-	-
met_eclat	<2e-16	<2e-16	-
met_weclat	<2e-16	<2e-16	<2e-16

Se muestran los resultados obtenidos al aplicar la prueba no paramétrica Wilcoxon mediante la corrección de Bonferroni (Tabla 8.26). Al analizar los valores en la Tabla 8.26 con respecto a las comparaciones múltiples de Wilcoxon se tuvo que para todos los pares el valor p obtenido de <2e-16 es menor al nivel de significancia de 0,05 por lo que todos estos pares (met_ASI -, met_apriori; met_eclat – met_apriori; met_weclat – met_apriori;

met_eclat – met_ASI; met_weclat – met_ASI y met_weclat – met_eclat) son significativamente diferentes.

Paso 4A: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5A: Decisión

A través de la opción Bonferroni se obtuvo que los pares met_ASI – met_apriori , met_eclat – met_a priori , met_weclat – met_a priori, met_eclat – met_ASI, met_weclat – met_ASI y met_weclat – met_eclat presentan diferencias significativas entre los pares de técnicas de reglas de asociación, dado que el valor p obtenido para cada par es 2e-16 el cual es menor a un nivel de significancia de 0,05 (Tabla 8.26).

Paso 3B: Estadístico de Prueba

Tabla 8.27.- Comparaciones múltiples Wilcoxon (Holm)

	met_apriori	met_ASI	met_eclat
met_ASI	<2e-16	-	-
met_eclat	<2e-16	<2e-16	-
met_weclat	<2e-16	<2e-16	<2e-16

Se muestran los resultados obtenidos al aplicar la prueba no paramétrica Wilcoxon, en este caso se usó la técnica de Holm (Tabla 8.27).

Paso 4B: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5B: Decisión

Mediante la técnica de Holm se obtuvo que los pares met_ASI – met_apriori, met_eclat – met_a priori, met_weclat – met_a priori, met_eclat – met_ASI, met_weclat – met_ASI y met_weclat – met_eclat son significativamente diferentes ya que obtuvieron un valor p de 2e-16 para cada par de métodos, el cual resultó menor al nivel de significancia de 0,05 por lo que se rechazó la hipótesis nula.

8.4.6.4 ANOVA no paramétrico

Utilizamos el paquete Rfit (Rank-Based Estimation for Linear Models) que proporciona funciones para análisis basados en rangos de modelos lineales, la inferencia ofrece una alternativa robusta a los mínimos cuadrados (Kloke y McKean, 2020).

Paso 1: Planteamiento de Hipótesis

H_o: $No\ hay\ diferencia\ entre\ los\ tiempos\ de\ las\ 4\ poblaciones\ de\ técnicas\ de\ reglas\ de\ asociación$

H_1: $Hay\ diferencia\ entre\ los\ tiempos\ de\ las\ 4\ poblaciones\ de\ técnicas\ de\ reglas\ de\ asociación$

Paso 2: Nivel de significancia α=0,05

Paso 3A: Estadístico de Prueba

F-Statistic=15464; p-value=0

Paso 4A: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza (Figura 8.14).

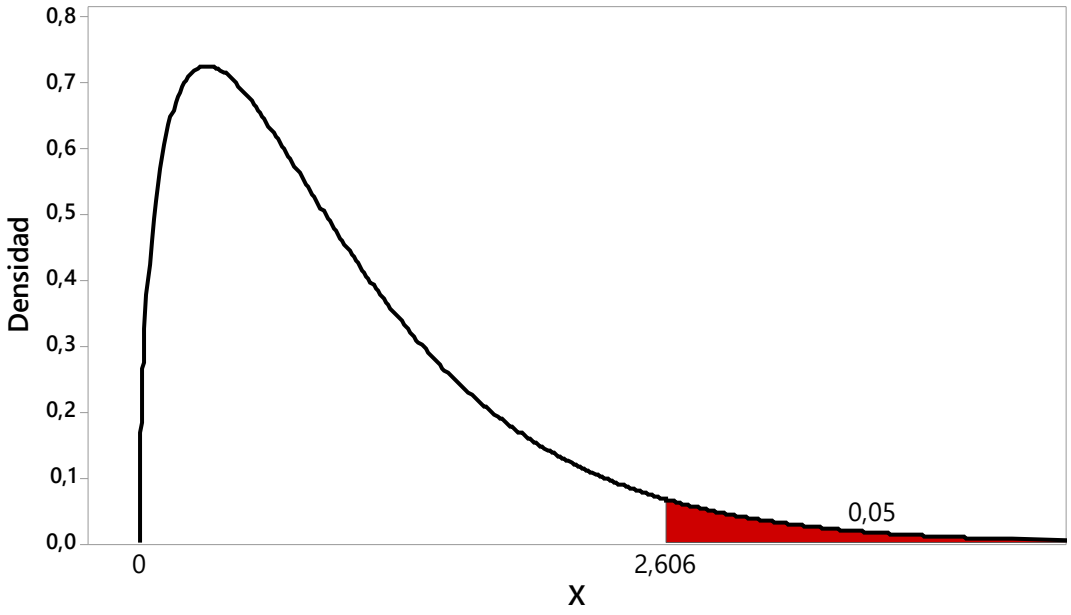


Figura 8.14.- Zonas de rechazo y aceptación ANOVA no paramétrico, complejidad temporal, reglas de asociación

Paso 5A: Decisión

El valor p obtenido en el ANOVA es igual a 0 el cual es menor a un nivel de significancia de 0,05 por lo que se rechazó la hipótesis nula y se concluye que las medias obtenidas entre las técnicas de reglas de asociación y el tiempo son distintas.

Paso 3B: Estadístico de Prueba

Tabla 8.28.- Comparaciones múltiples ANOVA no paramétrico

	met_apriori	met_ASI	met_eclat
met_ASI	< 2e-16	-	-
met_eclat	0,00078	< 2e-16	-
met_weclat	< 2e-16	< 2e-16	< 2e-16

Paso 4B: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Paso 5B: Decisión

Los resultados obtenidos en el paquete Rfit para cada uno de los pares de técnicas en estudio fueron 2e-16 y 0,00078 los mismos que son menores a un nivel de significancia de 0,05 por lo que se rechazó la hipótesis nula concluyéndose que existe diferencias significativas entre estos métodos.

Tukey

Mediante la prueba de Tukey se determinaron las técnicas de reglas de asociación que son significativamente diferentes con respecto a la cantidad de memoria que usa cada uno, para lo cual se plantearon las siguientes hipótesis:

Paso 1: Planteamiento de Hipótesis

$\mathbf{H_o}$: $No\ hay\ diferencia\ entre\ los\ tiempos\ de\ las\ 2\ poblaciones\ de\ técnicas\ de\ reglas\ de\ asociación$

$\mathbf{H_1}$: $Hay\ diferencia\ entre\ los\ tiempos\ de\ las\ 2\ poblaciones\ de\ técnicas\ de\ reglas\ de\ asociación$

Paso 2: Nivel de significancia α=0,05

Paso 3A: Estadístico de Prueba

Tabla 8.29.- Comparaciones múltiples (Tukey)

	I	J Estimate	St Err	Lower Bound CI	Upper Bound CI
1	met_apriori met_ASI	5,41753	0,0108	5,38979	5,44527
2	met_apriori met_eclat	-0,03629	0,0108	-0,06402	-0,00855
3	met_apriori met_weclat	0,27822	0,0108	0,25048	0,30596
4	met_ASI met_eclat	5,45382	0,0108	5,42608	5,48156
5	met_ASI met_weclat	5,13931	0,0108	5,11157	5,16705
6	met_eclat met_weclat	-0,31451	0,0108	-0,34224	-0,28677

Paso 4: Regla de Decisión

Si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

Los resultados obtenidos indican que cada una de las técnicas de reglas de asociación en estudio son significativamente diferentes con respecto al tiempo empleado, dicha aseveración se nota en los intervalos de confianza que están dados por valores negativos y positivos.

8.5 Conclusiones

Para demostrar las hipótesis se utilizó un diseño pre-experimental del tipo RGXO1, se trabajó con un nivel de confianza del 95% (α=0,05), las variables dependientes fueron para el caso de la complejidad espacial, la memoria y para el caso de la complejidad temporal, el tiempo, ambas son de tipo numérico.

Sobre la comparación de la complejidad espacial entre las técnicas de reglas de asociación del ASI: el método cuasi-implicativo (met_ASI) y de LA: apriori, eclat y weclat (met_apriori, met_eclat y met_weclat) se obtuvieron los siguientes resultados:

El estudio descriptivo de la memoria en las técnicas de reglas de asociación dio como resultado las siguientes 4 medidas básicas (Ver Figura 8.5).

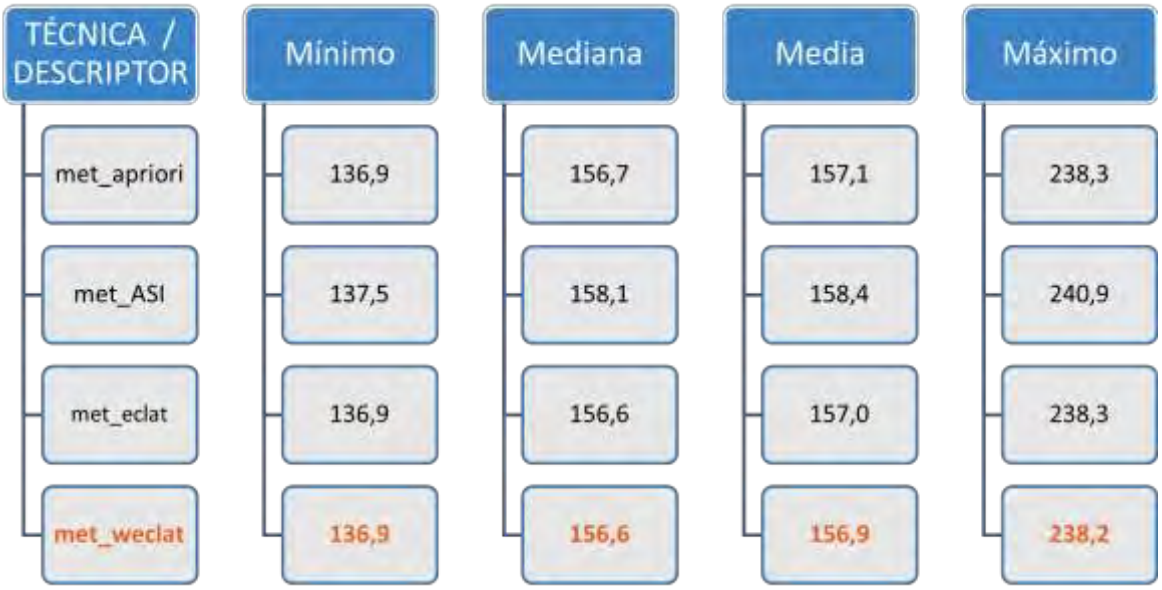


Figura 8.15.- Principales medidas descriptivas muestrales de memoria en las técnicas de reglas de asociación

El análisis de la muestra suministra la información de que la técnica que usa menor memoria es met_weclat con un valor igual a 156,9, con un mínimo de memoria igual a 136,9 y un máximo de 238,2, con una mediana de 156,6. Las otras técnicas de reglas de asociación se muestran bastante parecidas entre ellas sobre todo en sus medidas de centralización, las posteriores pruebas de hipótesis verificarán si se mantiene esta tendencia también en la población.

El estudio descriptivo del tiempo de ejecución dio como resultado las siguientes 4 medidas básicas de centralización y posición de las técnicas de reglas de asociación (Figura 8.16).

Figura 8.16.- Principales medidas descriptivas muestrales de tiempo en las técnicas de reglas de asociación

El análisis de la muestra suministra la información de que la técnica que usa en promedio menor tiempo es met_eclat con un valor igual a 0,98292, con un mínimo de tiempo igual a 0,03404 y un máximo de 115,99705, con una mediana de 0,20032. Las otras técnicas se muestran bastante diferentes entre ellas sobre todo en sus medidas de centralización, las posteriores pruebas de hipótesis verificarán si se mantiene lo observado también en la población.

Tanto los datos de memoria como de tiempo obtenidos de las técnicas de reglas de asociación no son estadísticamente normales por ninguna de las pruebas utilizadas (Anderson-Darling normality test, Lilliefors (Kolmogorov-Smirnov) normality test, Holm technique, Cramer-von Mises normality test, Pearson chi-square normality test, Shapiro-Wilk normality test, Shapiro-Francia normality test) a ninguno de los niveles del sesgo α=0,01, α=0,05 y tampoco α=0,1. Se aplicaron técnicas de transformación a la normalidad de los datos de memoria y tiempo usando la función powerTransform; con las técnicas

bcPower y yjPower, se evidenció que a los datos de las variables memoria y tiempo no se les puede normalizar ya que los p-values obtenidos son demasiado pequeños.

Se comprobó que existía heterocedasticidad tanto para los datos de memoria como para los datos de tiempo (con un alto nivel de significancia 2,2e-16), generados por las técnicas de reglas de asociación, para comprobarlo se utilizaron las pruebas de Bartlett y Levene.

El valor p obtenido en la prueba chi cuadrado $\chi 2$ de independencia fue 7,285e-16 en el caso de la memoria y menor a 2,2e-16 en el caso del tiempo, los cuales son menores que el nivel de significancia de $\alpha=0,05$ por lo que se rechaza la hipótesis nula y se concluye que las técnicas de reglas de asociación y la memoria (también las técnicas de reglas de asociación y el tiempo) no son independientes.

La prueba de hipótesis no paramétricas aplicadas para la comparación de la complejidad espacial (memoria) entre las 4 técnicas de reglas de asociación (met_apriori, met_eclat, met_weclat y met_ASI) indican que a un nivel de significancia de $\alpha=0,05$ y utilizando la prueba deKruskal Wallis (p-value=6,391e-11) y ANOVA no paramétrico (p-value=6,8834e-14) se rechaza la hipótesis nula con alta significancia, por lo tanto existe al menos un par de poblaciones de reglas de asociación diferentes respecto a la memoria. Aplicando las pruebas posteriores de Mann Whitney Wilcoxon (con las correcciones de Bonferroni y Holm), ANOVA no paramétrico y Tukey, se obtuvieron 2 grupos de homogeneidad (A y B) que se resume en la Tabla 8.30.

Tabla 8.30.- Grupos de homogeneidad para la memoria en las técnicas de reglas de asociación (las A en rojo son las técnicas con menor parámetro)

GRUPOS	met_apriori	met_ASI	met_eclat	met_weclat
1	A		A	A
2		B		

Las pruebas de hipótesis aplicadas para la comparación de la complejidad temporal (tiempo) entre las 4 técnicas de reglas de asociación (met_apriori, met_eclat, met_weclat y met_ASI) indican que a un nivel de significancia de $\alpha=0,05$ y utilizando las pruebas de Kruskal Wallis (p-value < 2,2e-16) y ANOVA no paramétrico (p-value=0), se rechaza la hipótesis nula con alta significancia, por lo tanto, existe al menos un par de poblaciones de reglas de asociación diferentes respecto al tiempo. Aplicando las pruebas posteriores de Mann Whitney Wilcoxon (con las correcciones de Bonferroni y Holm), ANOVA no

paramétrico y Tukey, se obtuvieron 4 grupos de homogeneidad que se resume en la Tabla 8.31.

Tabla 8.31.- Grupos de homogeneidad para el tiempo en las técnicas de reglas de asociación (la A en rojo es la técnica con menor parámetro)

GRUPOS	met_apriori	met_ASI	met_eclat	met_weclat
1			A	
2	B			
3				C
4		D		

Considerando la complejidad espacial y temporal simultáneamente (en forma ascendente) de las técnicas de reglas de asociación analizadas se obtendría la Tabla 8.32.

Tabla 8.32.- Complejidad espacial y temporal simultáneamente para las técnicas de reglas de asociación

ORDEN	TÉCNICA DE REGLAS DE ASOCIACION
1	met_eclat
2	met_apriori
3	met_weclat
4	met_ASI

Donde se observa que la técnica con menor ocupación de memoria y tiempo en las técnicas de reglas de asociación para bases de datos de máximo 1000 observaciones y 100 variables es met_eclat, ubicándose met_ASI al final de la tabla.

Capítulo 9no | APORTES FACTIBLES DESDE LAS OPCIONES ADICIONALES DEL ASI

Se presenta las opciones adicionales a las técnicas del ASI, que podrían ser aportes factibles a LA.

9 Capítulo.- Aportes factibles desde las opciones adicionales del ASI

La mayoría de las técnicas de análisis clúster y de reglas de asociación no tienen opciones adicionales que ayuden a profundizar el proceso de análisis como lo tiene el ASI, algunas de ellas han sido y podrían seguir siendo de gran utilidad en LA. A continuación, se detallan.

9.1 Introducción

Se resaltan las opciones adicionales del Análisis Estadístico Implicativo que podrían incrementar el aporte del ASI a las Analíticas de Aprendizaje. El ASI permite bases de datos multivariadas con valores binarios, modales (categóricos ordinales), frecuenciales (numéricos discretos), intervalo (numéricos continuos), utiliza variables suplementarias, tiene la opción de cálculo entrópico que permite trabajar con gran número de datos, está apoyada con opciones adicionales tales como nodos significativos, la tipicalidad y la contribución, permite visualizaciones sencillas de interpretar como grafos implicativos, conos implicativos, dendrogramas simétricos, dendrogramas asimétricos y permite la creación de escenarios de análisis y experimentación a través de la interacción con las variables.

9.2 Tipo de datos amplio

El ASI acepta los dos tipos de datos universalmente considerados dentro del análisis estadístico, los de tipo atributo y los de tipo numérico.

Los datos de tipo atributo pueden ser nominales u ordinales, dentro de los datos nominales los datos binarios y dentro de los datos ordinales los datos modales, pero siempre trasladados al intervalo [0,1].

Los datos de tipo numérico se clasifican en tipo intervalo y tipo razón. Los datos de tipo razón están totalmente cubiertos por los de datos de tipo intervalo del ASI y en particular los datos de tipo frecuencial del ASI son los datos numéricos discretos. Los datos de tipo intervalo se pueden considerar todos, pero sin cero relativo coincidente con el cero del intervalo [0,1].

9.3 Variables suplementarias

Las variables suplementarias son variables cualitativas como el género, el nivel educativo o la categoría económica. Es posible conocer cuáles son los sujetos o clases de sujetos con más responsabilidad en las implicaciones calculadas. Su función es dar información descriptiva sobre la formación de las categorías, esta información se observará únicamente al utilizar las opciones de tipicalidad y contribución (Orús et al., 2005). Se indican con el nombre de la variable seguida de un espacio y la letra s por ejemplo casado s. Si por ejemplo se desea conocer si una determinada implicación (método de estudio -> rendimiento) está influenciada por la edad de los estudiantes, se ingresará una nueva variable edad s, que contendrá información de tipo 0, 0,25, 0,5, 0,75 y 1 que corresponderá a las edades 18 años, 19 años, 20 años, 21 años, 12 o más años. Pueden ser variables de tipo binario o modal (Gras et al., 2002).

9.4 Nodos significativos

Los árboles de similaridad y de cohesión muestran en sus respectivos dendrogramas ciertos nodos de color rojo que son llamados nodos significativos. Conceptualmente los nodos significativos son nodos correspondientes a una clasificación compatible lo mejor posible con los valores y calidad de la agrupación obtenida.

Los nodos significativos son nodos particulares tanto de un árbol de similaridad (jerárquico) o un árbol cohesivo, son los nodos correspondientes a una clasificación compatible lo mejor posible con los valores y la calidad de los valores de similaridad. Un nodo significativo es un nodo que se quiere destacar de los demás, por reunir en su seno a variables que mantienen una similaridad (cuando se trabaja por parejas) mejor que en otros niveles (Valls, 2014).

Se llama preorden incial y global Ω sobre AxA, al preorden inducido por la aplicación S (similaridad) sobre AxA. $G_s(\Omega) = \{((a,b);(c,d)) : s(a,b) < s(c,d)\}$. Sea $S\Pi_k$ el conjunto de pares separadas al nivel k y $R\Pi_k$ el conjunto de pares que ya se han reunido hasta este nivel k.$G_s(\Omega) \cap [S\Pi_k x R\Pi_k]$ está formado por los pares de parejas que al nivel k respetan el preorden inicial. Por ejemplo, si se tiene: s(e,f)<s(a,b) entonces((e,f);(a,b))$\epsilon G_s(\Omega)$ y si al nivel k, e y f están aún separados mientras que a y b se reúnen en la clase formada, la pareja (e,f) (a,b) $\in G_s(\Omega)$ $[S\Pi_k x R\Pi_k]$. El cardinal de este último conjunto es función de este

nivel k, y es un indicador del acuerdo entre el preorden inicial l Ω y el preorden Π_k inducido.

A la cardinalidad de $G_s(\Omega) \cap [\mathrm{S}\Pi_k x \mathrm{R}\Pi_k]$ se le asocia el índice aleatorio $G_s(\Omega^*) \cap [\mathrm{S}\Pi_k x \mathrm{R}\Pi_k]$. Donde Ω^* es un preorden aleatorio en general, provisto de una probabilidad uniforme, de todos los preordenes del mismo tipo cardinal que Ω. Este índice tiene por esperanza $\frac{1}{2}S_k r_k$, por varianza $\frac{S_k r_k (S_k + r_k + 1)}{12}$, siendo Card$[\mathrm{S}\Pi_k] = S_k$ y Card $[\mathrm{R}\Pi_k] = r_k$. El índice centrado se define como $S(\Omega, K) = \frac{Card[G(\Omega) \cap [\mathrm{S}\Pi_k x \mathrm{R}\Pi_k]] - \frac{1}{2}S_k r_k}{\sqrt{\frac{S_k r_k (S_k + r_k + 1)}{12}}}$.

Se llama nivel significativo a todo nivel que corresponde a un máximo local de S(Ω,k) durante la construcción de la jerarquía. En este caso se dirá que la división Π_k está en resonancia parcial con Ω. Si, además, $G_s(\Omega) \cap [\mathrm{S}\Pi_k x \mathrm{R}\Pi_k]$= $[\mathrm{S}\Pi_k x \mathrm{R}\Pi_k]$,diremos que la división Π_k está en resonancia total con Ω.

Se llama nodo significativo cualquier nodo formado a un nivel que corresponde a un máximo local de (Ω k,), donde: v(Ω ,k) = S(Ω,K)−S (Ω,K-1).

9.5 Entropía y conjuntos grandes de datos

Cuando se necesita analizar conjuntos grandes de datos el ASI mediante CHIC y Rchic ofrece la opción de cálculo llamada entropía (Lerman, Gras, et al., 1981). La entropía se utiliza para analizar una muestra grande de datos. Me permito hacer referencia textual a lo indicado en un artículo científico reciente (Gras et al., 2015) escrito por el creador de la teoría ASI, Regis Gras (Blanchard et al., 2003):

> “Elle vise à se substituer à la modélisation entropique jusqu’alors utilisée et qui présente un caractère jugé trop ad-hoc par les familiers de l’A.S.I. Elle va donc être construite contre une connaissance antérieure comme le dit G. Bachelard dans (Bachelard G.1967). Cependant, cette précédente modélisation était loin de déplaire aux utilisateurs qui en appréciaient la capacité à accepter plus facilement la grande taille de l’échantillon des sujets considéré. D’où son intérêt pour ce que l’on appelle les «big data»”.

Que traducido significa: “Su objetivo es reemplazar el modelado entrópico utilizado hasta entonces y que presenta un carácter considerado demasiado ad-hoc por

aquellos familiarizados con A.S.I. Por lo tanto, se construirá contra el conocimiento previo, como lo afirma G. Bachelard en (Bachelard G, 1967). Sin embargo, este modelo anterior (El modelo entrópico) estuvo lejos de desagradar a los usuarios que apreciaron su capacidad para aceptar más fácilmente el gran tamaño de muestra de los sujetos considerados. De ahí su interés por lo que se denomina "big data".

Que nos indica que el modelo entrópico del ASI es útil en conjuntos grandes de datos.

La opción versión entrópica en CHIC, permite trabajar con cantidades grandes de datos. A continuación, la formalizamos matemáticamente.

Recordemos que la intensidad de implicación de la regla de asociación $\mathrm{a} \Rightarrow \mathrm{b}$ está definida por $\varphi\,(\mathrm{a} \Rightarrow \mathrm{b}) = 1 - \Pr(\tilde{\mathrm{N}}_{X\cap Y} \leq \tilde{\mathrm{n}}_{A\cap B})$ si $n_B \neq n$; caso contrario $\varphi\,(\mathrm{a} \Rightarrow \mathrm{b}) = 0$ la regla se mantiene para un umbral dado $1 - \sigma$ si $\varphi(\mathrm{a} \Rightarrow \mathrm{b}) \geq 1 - \sigma$.

La definición anterior mide la importancia de la regla $\mathrm{a} \Rightarrow \mathrm{b}$. Sin embargo, tener en cuenta el contrapositivo $\bar{b} \Rightarrow \bar{a}$ podría reforzar la afirmación de la implicación entre a y b. Además, podría mejorar la calidad de discriminación de φ cuando el conjunto de transacciones T aumenta: si A y B son pequeños en comparación con T, sus conjuntos complementarios son grandes y viceversa

Por estas razones, aquí presentamos una versión ponderada de la intención de implicación $\varphi\,(\mathrm{a} \Rightarrow \mathrm{b})$= $(\phi\ (\mathrm{a} \Rightarrow \mathrm{b}).\mathrm{T}\,(\mathrm{a},\mathrm{b}))^{\frac{1}{2}}$ donde $\mathrm{T}(\mathrm{a},\mathrm{b})$ que mide el desequilibrio entre $n_{A\cap\bar{B}}$ y $n_{\bar{A}\cap\bar{B}}$ asociado con $\mathrm{a} \Rightarrow \mathrm{b}$ y el desequilibrio entre $n_{A\cap\bar{B}}$ y $n_{\bar{A}\cap\bar{B}}$ asociado con su contrapositivo. Intuitivamente, la sorpresa inducida por la regla medida por φ debe suavizarse (respectivamente confirmarse) cuando el número de contraejemplos $n_{A\cap\bar{B}}$ es alto (respectivamente pequeño) para la regla y su contrapositivo considerando los números observados n_a y $n_{\bar{b}}$. Aquí, seguimos un enfoque axiomático de la medición de desequilibrios.

Un índice bien conocido para tener en cuenta los desequilibrios de forma no lineal es la entropía condicional de Shannon. La entropía condicional $H_{\frac{b}{a}}$ de los casos (a y b) y (a y $\bar{b}$) dados se define por $H_{\frac{b}{a}} = -\frac{n_{a\cap b}}{n_a} log_2 \frac{n_{a\cap b}}{n_a} - \frac{n_{a\cap\bar{b}}}{n_a} log_2 \frac{n_{a\cap\bar{b}}}{n_a}$ y, de manera similar, la entropía

condicional $H_{\frac{b}{a}}$ de los casos ($\bar{a}$ y $\bar{b}$) y (a y $\bar{b}$) dado $\bar{b}$ definidos por $H_{\frac{\bar{a}}{\bar{b}}} = -\frac{n_{a\cap\bar{b}}}{n_{\bar{b}}} log_2 \frac{n_{a\cap\bar{b}}}{n_{\bar{b}}} - \frac{n_{a\cap\bar{b}}}{n_{\bar{b}}} log_2 \frac{n_{a\cap\bar{b}}}{n_{\bar{b}}}$. Aquí podemos considerar que estas entropías miden la incertidumbre promedio de los experimentos aleatorios en los que comprobamos si se realiza b (resp. $\bar{a}$) cuando a (resp. $\bar{b}$) es observado. Los complementos de 1 para estas incertidumbres $I_{\frac{b}{a}} - 1 - H_{\frac{b}{a}}$ y $I_{\frac{\bar{a}}{\bar{b}}} - 1 - H_{\frac{\bar{a}}{\bar{b}}}$ pueden interpretarse como la información promedio recopilada por la realización de estos experimentos; cuanto mayor sea esta información, más fuerte será la garantía de la calidad de la implicación y su contrapositivo. Intuitivamente, el comportamiento esperado de la nueva medida φ está determinado por tres etapas:

1. Una reacción lenta a los primeros contraejemplos (robustez al ruido).
2. Una aceleración del rechazo en la vecindad de la balanza.
3. Un rechazo creciente más allá del equilibrio que no estaba garantizado por la intensidad de implicación básica ϕ.

Más precisamente, para tener el significado esperado, nuestro modelo debe satisfacer las siguientes restricciones:

1. Integrando tanto la información relativa a $(\mathrm{a} \Rightarrow \mathrm{b})$ y como relativa a ($\bar{b}$ y $\bar{a}$) medida respectivamente por $I_{\frac{b}{a}}$ y $I_{\frac{\bar{a}}{\bar{b}}}$. Un producto $I_{\frac{b}{a}}\ I_{\frac{\bar{a}}{\bar{b}}}$ está bien adaptado para resaltar simultáneamente la calidad de estos dos valores.
2. Elevar las entropías condicionales a la potencia de un número fijo $\alpha > 1$ en las definiciones de información para reforzar el contraste entre las diferentes etapas que se detallan a continuación: $\left(\left(1 - H_{\frac{b}{a}}^{a}\right).\left(1 - H_{\frac{a}{\bar{b}}}^{a}\right)\right)^{\frac{1}{\beta}}$ con $\beta = 2\alpha$ para permanecer en la misma dimensión como φ.

La necesidad de considerar que las implicaciones han perdido su significado inclusivo cuando el número de contraejemplos es mayor a la mitad de las observaciones de a y b (lo que parece bastante natural). Más allá de estos valores, consideramos que cada uno de los términos $\left(1 - H_{\frac{b}{a}}^{a}\right)$ y $\left(1 - H_{\frac{b}{a}}^{a}\right)$ es igual a 0. Sea $f_a = \frac{n_a}{n}$ (respectivamente $f_{\bar{b}} = \frac{n_{\bar{b}}}{n}$)

la frecuencia de una (resp. $\bar{b}$) en el conjunto de transacciones y $f_{a\cap\bar{b}} = \frac{n_{a\cap\bar{b}}}{n}$ la frecuencia de contraejemplos. El ajuste propuesto de la información se define fácilmente por $\widehat{I^a}_{\frac{b}{a}} = 1 - H^a_{\frac{b}{a}} = 1 + \left(\left(1 - \frac{f_{a\cap\bar{b}}}{f_a}\right) log\left(1 - \frac{f_{a\cap\bar{b}}}{f_a}\right) + \left(\frac{f_{a\cap\bar{b}}}{f_a}\right) log\left(\frac{f_{a\cap\bar{b}}}{f_a}\right)\right)^a$ $y\ f_{a\cap\bar{b}} \in \left[0, \frac{f_a}{2}\right[$; de lo contrario $\widehat{I^a}_{\frac{b}{a}} = 0$ y $\widehat{I^a}_{\frac{\bar{a}}{\bar{b}}} = 1 - H^a_{\frac{\bar{a}}{\bar{b}}} = 1 + \left(\left(1 - \frac{f_{a\cap\bar{b}}}{f_{\bar{b}}}\right) log\left(1 - \frac{f_{a\cap\bar{b}}}{f_{\bar{b}}}\right) + \frac{f_{a\cap\bar{b}}}{f_{\bar{b}}} log\left(\frac{f_{a\cap\bar{b}}}{f_{\bar{b}}}\right)\right)^a$ $y \quad f_{a\cap\bar{b}} \in \left[0, \frac{f_{\bar{b}}}{2}\right[$; de lo contrario $\widehat{I^a}_{\frac{\bar{a}}{\bar{b}}} = 0$

Los desequilibrios se miden mediante $\mathrm{T}(a, b)$, denominado índice de inclusión, definido por $\mathrm{T}(a, b) = \left(\widehat{I^a}_{\frac{b}{a}}.\widehat{I^a}_{\frac{\bar{a}}{\bar{b}}}\right)^{\frac{1}{2a}}$ y la versión ponderada de la intensidad de la implicación, denominada intensidad de la implicación entrópica (EII), está dada por $\varphi\,(\mathrm{a} \Rightarrow \mathrm{b}) = (\phi(\mathrm{a} \Rightarrow \mathrm{b})\mathrm{T}(\mathrm{a}, \mathrm{b}))^{\frac{1}{2}}$.

9.6 Tipicalidad

La tipicalidad indica los sujetos típicos de una clase de implicaciones, es un índice porcentual que cuantifica cómo un individuo concreto se comporta en relación con la regla, cuán "típico" es. Se define como sujeto típico aquel que verifica todas las implicaciones que poseen mayor intensidad de implicación en la formación de las clases.

Formalmente, el par (a, b) tal que: $\psi(a, b) \geq \psi(i, j) \forall i \in A\ y\ \forall j \in B$ es denominado par genérico de la clase C. Llamaremos Implicación Genérica de C al número $\psi(a, b)$.

Dado el par genérico (a, b), para cada individuo x de la muestra, se define $\psi_x(a, b)$ como: $\psi_x(a, b) = 1\ si\ b(x) = 1,\ \psi_x(a, b) = 0\ si\ a(x) = 1\ y\ b(x) = 0$ o $\psi_x(a, b) = p\ si\ a(x) = 0$ y $b(x) = 0$

En caso de que la clase C tuviera g subclases (anidadas), se puede tomar cada uno de los g pares genéricos. Se define la distancia de un individuo x a la clase C: $d^2(x, C) = \frac{1}{g}\sum_{i=1}^{g} \frac{[\psi_i - \psi_{x,i}]^2}{1-\psi_i}$, donde para cada i que indica cada subclase de C, ψ_i denota la implicación genérica de esa subclase y $\psi_{x,i}$ el valor de la implicación genérica para el individuo.

La tipicalidad del individuo x se define como $\gamma_T(x,C) = 1 - \frac{d(x,C)}{máx_{y\epsilon I} d(y,C)}$. Si x es un sujeto típico verifica todas las implicaciones que poseen mayor intensidad de implicación en la formación de las clases, esto es, $\psi_{x.i} = \psi_i, i = 1, \ldots . g$. Por tanto $d(\mathrm{x}, \mathrm{C}) = 0$ y $\gamma(\mathrm{x}, \mathrm{C}) = 1$.

Por contra, cuando x es el que más en desacuerdo está con C, significa que $d(x,C) = max_{y\epsilon I} d(y,C)$ y, por tanto, $\gamma_T(\mathrm{x}, \mathrm{C}) = 0$. Llamamos grupo óptimo al grupo de individuos con las mayores tipicalidades, la tipicalidad indica los sujetos típicos de la población para las implicaciones calculadas (Gras et al., 2006).

9.7 Contribución

La contribución, indica que sujetos aportan más a una implicación, se utiliza para saber cuáles son los temas o clases de sujetos más responsables de las implicaciones calculadas. La contribución, proporciona una forma de cuantificar cómo ha contribuido un determinado individuo a formar la calidad de la regla a → b. La contribución de un individuo x a la clase C se define como $\gamma_C(x,C) = 1 - \tilde{d}(x,C)$, donde $\tilde{d}^2(x,C) = \frac{1}{g}\sum_{i=1}^{g}[1 - \psi_{x,i}]^2$ es otra distancia del individuo x a la clase C. Si x es óptimo $\gamma_C(x,C) = 1$, en cuyo caso $\psi_{x,i} = 1$ para toda regla i (Gras y Régnier, 2017).

9.8 Escenarios de análisis y experimentación

Los sistemas informáticos que automatizan el ASI permiten crear escenarios de análisis y experimentación. Esto se puede realizar manejando la ventana de variables, seleccionando o no las variables que se desean analizar, sin necesidad de modificar la base de datos para cada escenario de análisis, proveyendo un método fácil de experimentación. En la Figura 9.1 se muestra la ventana de variables que CHIC (y también Rchic) provee para facilitar la creación de escenarios de análisis y experimentación.

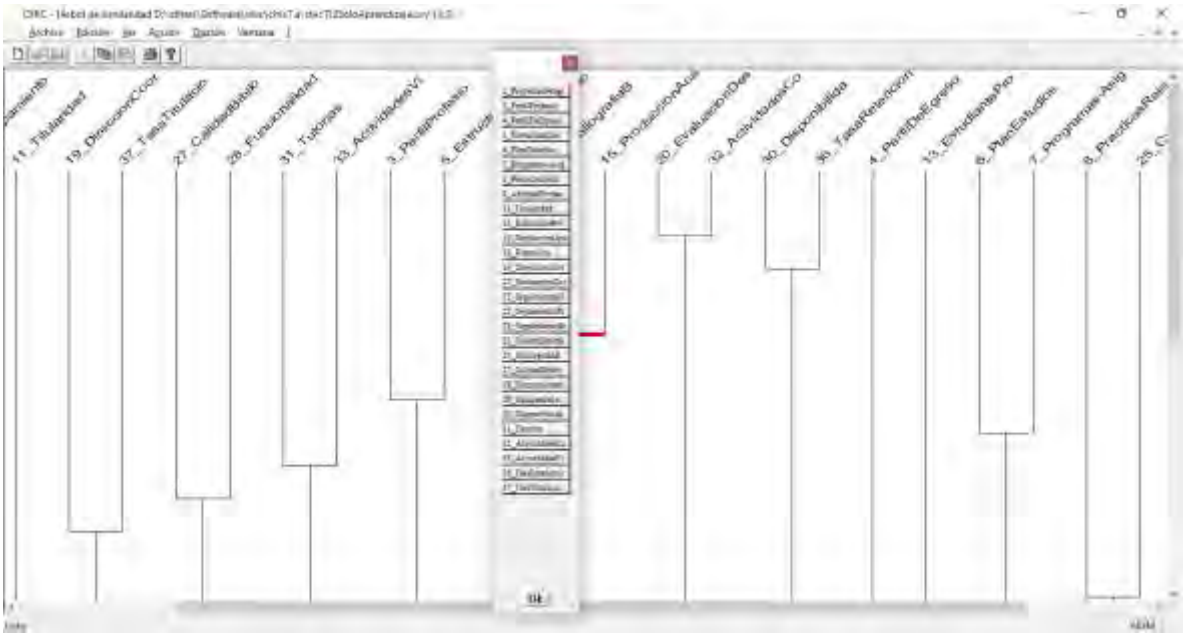


Figura 9.1.- Ejemplo de ventana de variables (ver al centro)

La Figura 9.2 muestra un primer ejemplo de escenario usando todas las variables y dejando de utilizar las variables: 4_PerfilDeEgreso, 5_EstructuraCurr, 6_PlanEstudios, 7_Programas-Asig, 8_PracticasRelac, 9_AfinidadFormac, 11_Titularidad, 13_EstudiantePro, 18_Ponencias.

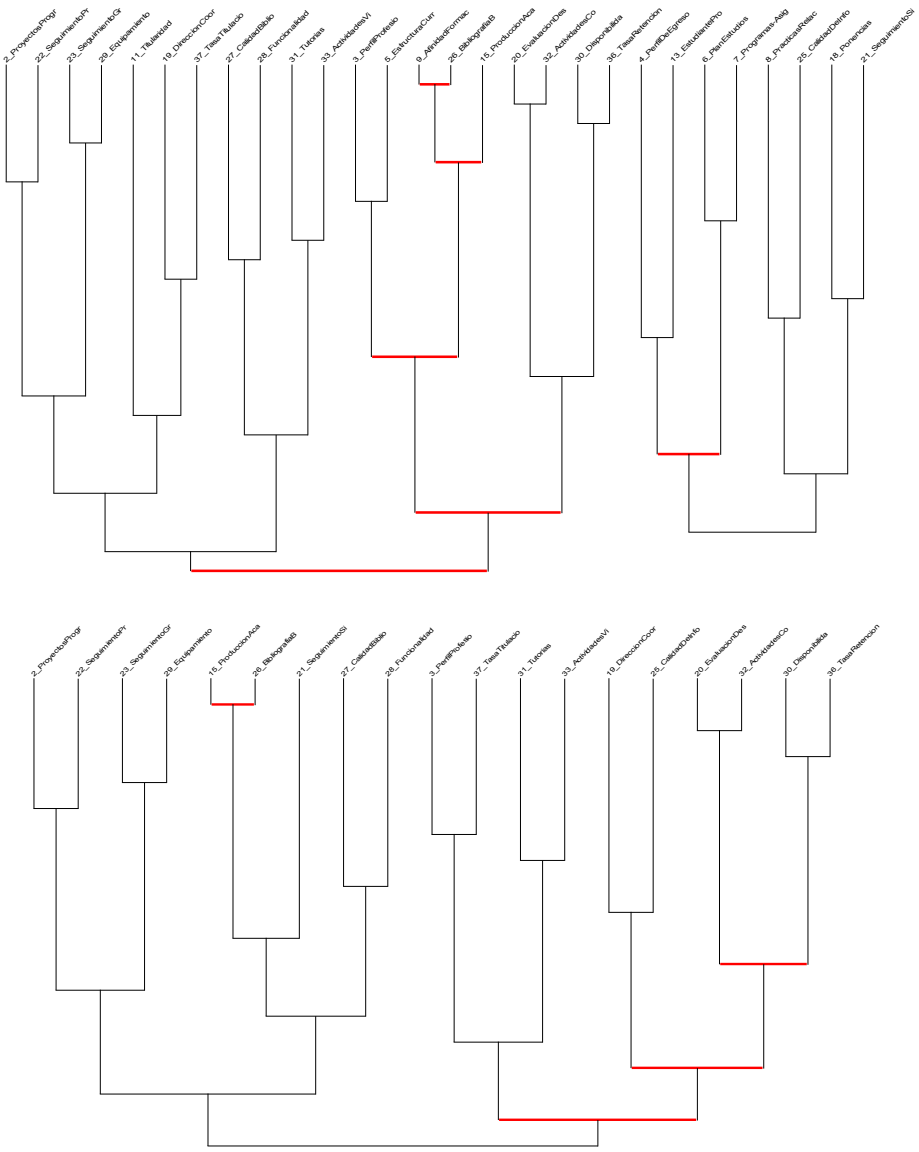


Figura 9.2.- Primer ejemplo de escenario sin las variables 4, 5, 6, 7, 8, 9, 11, 13 y 18.

Observe que en el segundo gráfico de similaridad, no se encuentran las variables: 4_PerfilDeEgreso, 5_EstructuraCurr, 6_PlanEstudios, 7_Programas-Asig, 8_PracticasRelac, 9_AfinidadFormac, 11_Titularidad, 13_EstudiantePro, 18_Ponencias y que para ello se ha desmarcado las variables antes citadas, como se muestra en la Figura 9.3.

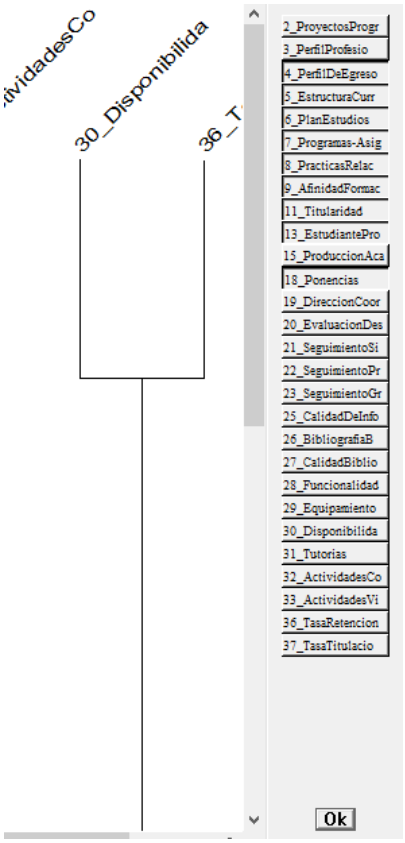


Figura 9.3.- Escenario desmarcando las variables 4, 5, 6, 7, 8, 9, 11, 13 y 18.

9.9 Visualizaciones sencillas de interpretar

9.9.1 El grafo implicativo

El grafo implicativo es un grafo orientado con vértices que representan las variables y los lados orientados que representan las cuasi-implicaciones entre variables (Figura 9.4). Generalmente los lados son unidireccionales pudiéndose mostrar otro lado con una dirección contraria si la cuasi-implicación existe entre las variables p y q así como entre

las variables q y p. No solo se muestran implicaciones entre variables, sino entre clases de variables.

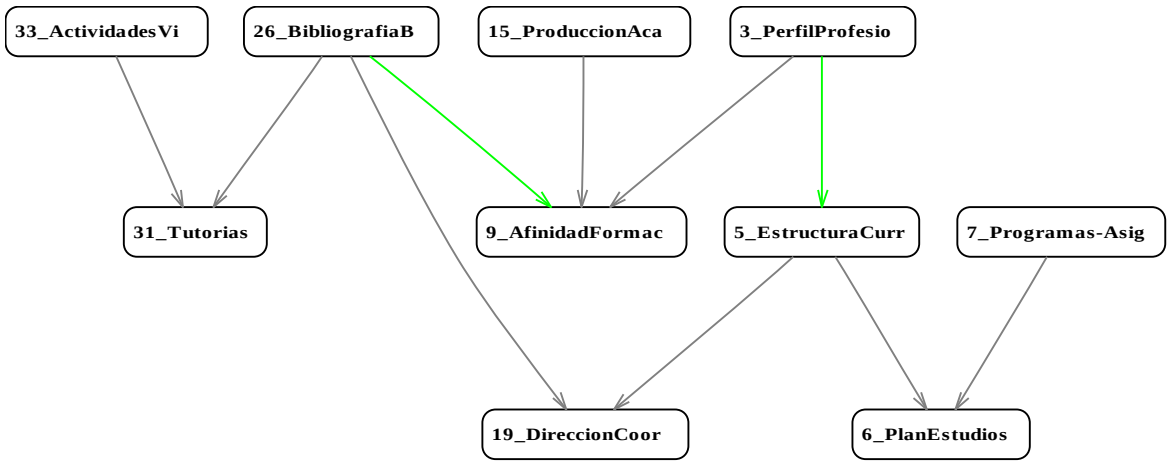


Figura 9.4.- Grafo implicativo

9.9.2 Dendrogramas simétricos

Los dendrogramas simétricos son utilizados para el estudio de la similaridad de Lerman, se muestran como gráfico de árbol que visualiza el valor de similaridad en forma ascendente (Figura 9.5), es decir mientras se sube por el árbol aumentan los valores de similaridad, es decir los valores más altos se encuentran en la parte más alta del árbol y los valores más bajos de similaridad se encuentran en la parte inferior del árbol. No solo se muestra similaridad entre variables, sino entre clases de variables, formándose así los niveles 0, 1, 2, 3,..., n.

El nivel 0 está formado por la similaridad de todas las variables 2 a 2, formando una tabla de doble entrada simétrica debido a que los valores s(p,q)=s(q,p). Es por ello por lo que se llama dendrograma simétrico.

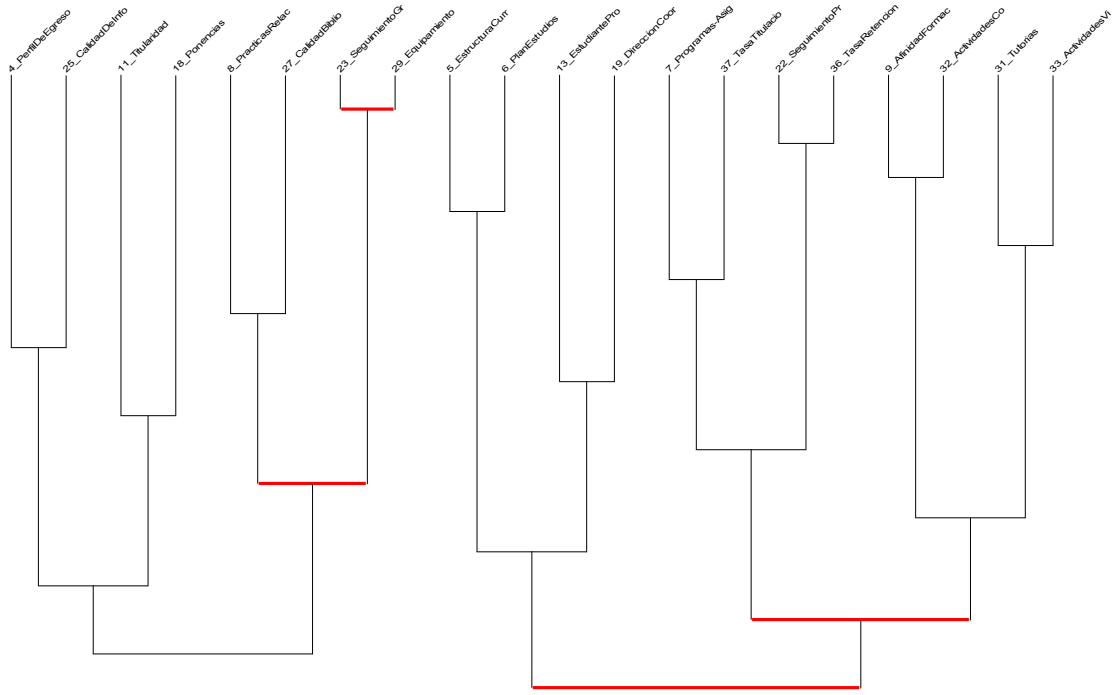


Figura 9.5.- Dendrograma simétrico

9.9.3 Dendrogramas asimétricos

Los dendrogramas asimétricos son generados como parte del estudio de cohesión, se muestran como gráfico de árbol que mide la cohesión en forma ascendente, es decir mientras más arriba del árbol nos encontramos entonces los valores de cohesión son más altos (Ver Figura 9.6). Se muestra también la cohesión entre clases de variables (no solo dos variables), formándose distintos niveles de cohesión etiquetados por 0, 1, 2, 3,..., n.

El nivel 0 está formado por la cohesión de todas las variables 2 a 2, formando una tabla de doble entrada asimétrica debido a que los valores de cohesión entre p y q son diferentes a los valores de cohesión entre q y p, es por ello por lo que se llama dendrograma asimétrico.

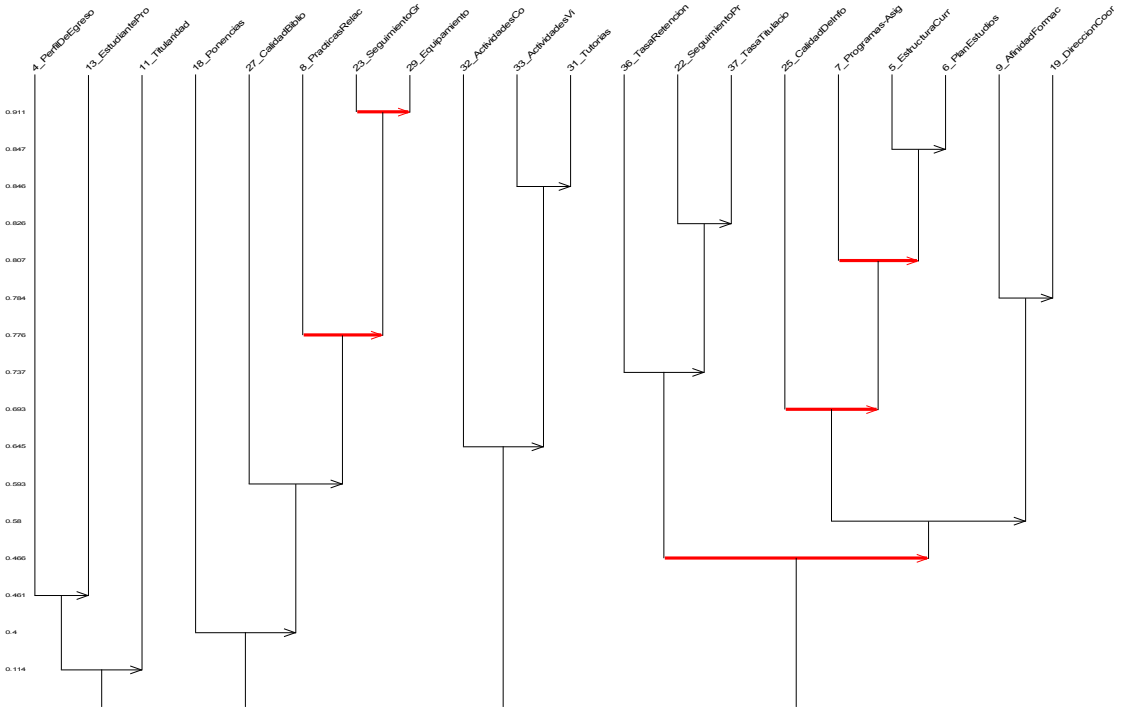


Figura 9.6.- Dendrograma asimétrico

9.9.4 El cono implicativo

Con un conjunto de datos que cruzan sujetos y variables, el Análisis Estadístico Implicativo ofrece respuestas a ciertos objetivos del investigador:

1. Generar reglas, del tipo: "si es verdadera la variable a, entonces generalmente la variable b también lo es".

2. Asignar a cada regla un valor numérico entre [0,1] que aumenta con la calidad predictiva de la regla.

3. Representar mediante un gráfico denominado implicativo, no simétrico, ponderado y sin ciclos, el conjunto de reglas de calidad al menos igual a un cierto umbral de aceptabilidad.

4. Interpretar dando significado a las líneas del gráfico implicativo, pero también a las estructuras relacionadas que lo constituyen.

Así, para un nivel dado de umbral de implicación, hay ciertas subestructuras relacionadas donde los nodos del gráfico admiten de forma aislada antecedentes ("padres") y sucesores ("hijos"). Estas subestructuras pueden ser extraídas por el propio investigador analizando el gráfico implicativo proporcionado por el software CHIC (esta opción no está en Rchic). Los antecedentes y sucesores pueden posiblemente interpretarse respectivamente como causas y consecuencias de la parte superior del gráfico desde una perspectiva causal (Figura 9.7).

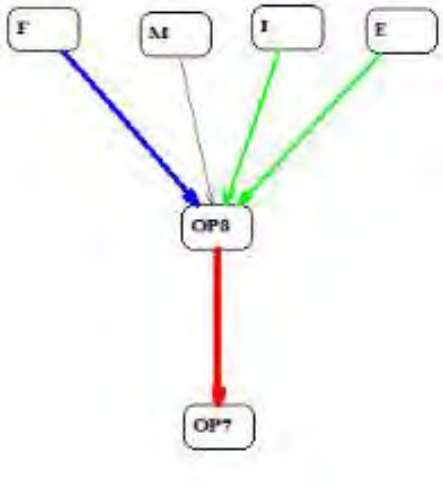


Figura 9.7.- Ejemplo de cono implicativo (Lahanier-Reuter et al., 2017)

La Figura 9.7 representa un cono implicativo que ha sido extraído de un gráfico implicativo más general con un umbral de 0,69. Los superiores son antecedentes ("padres") y los inferiores sucesores ("hijos"). La parte superior de este cono, OP8, admite como antecedentes ("padres") o posibles "causas" las variables F, M, I y E. Tiene un solo sucesor ("hijo") o consecuencia que es OP7. Este vértice juega un papel que se puede designar como “confluencia” o “variable nodal” de variables. Se eligió la palabra "cono" porque también simboliza, un embudo, donde se encuentran varios afluentes. Incluso se

puede hablar de un campo causal. Dado que este cono puede extenderse a lo largo de sus dos capas, vemos que la noción de "causa" (o "padres") y "consecuencia" (o "efecto" o "hijo") es bastante relativa. Un padre es padre solo en la medida en que tiene un hijo.

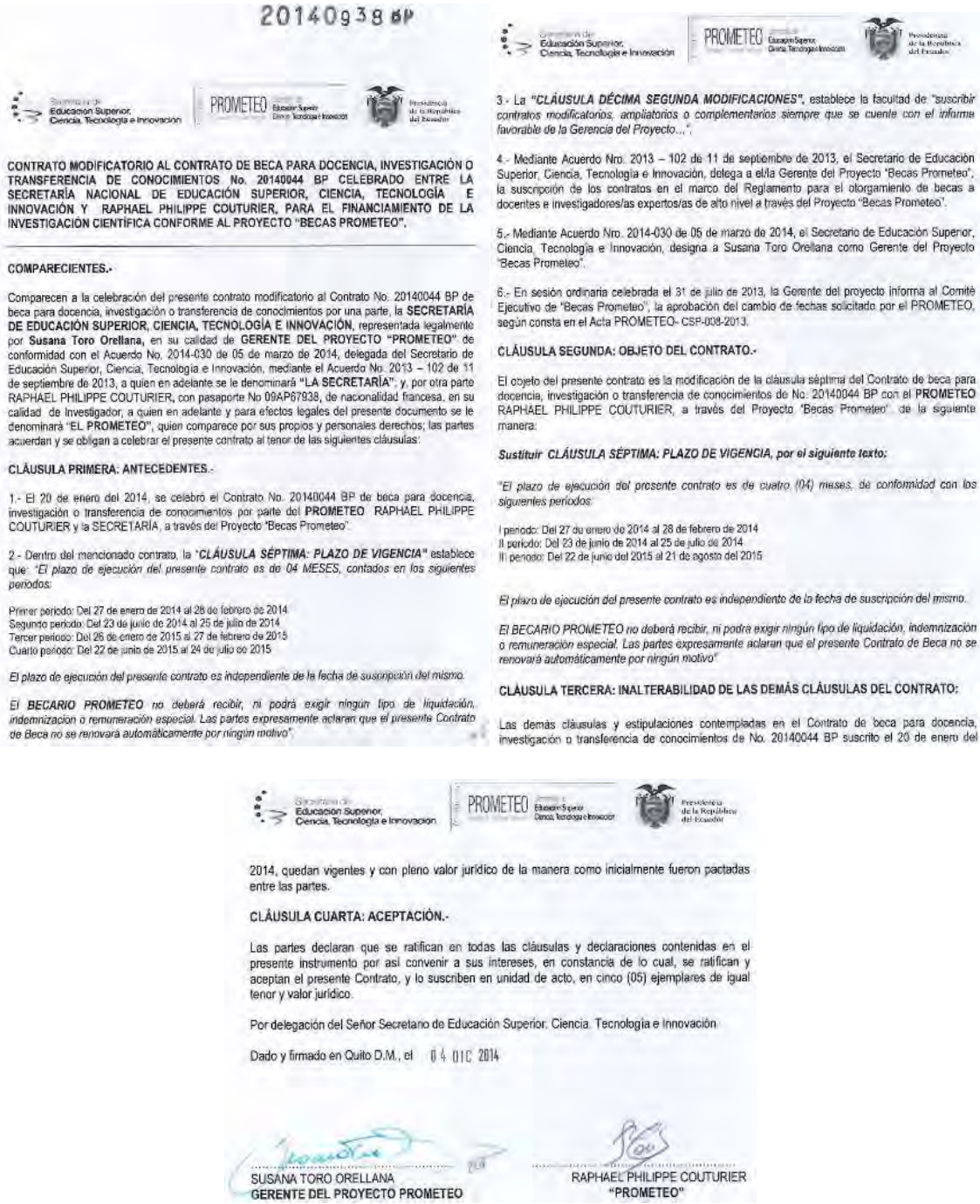
Si queremos identificar a los “padres” en la parte superior del cono como causas que no derivan de fenómenos causales comunes o estrechamente relacionados, como “llueve” y “el suelo está mojado”, debemos exigir que estas variables aguas arriba no estén vinculadas o solo estén débilmente vinculadas. Si este es el caso, podemos decir que dicha variable (por ejemplo, la parte superior del cono en este caso) depende esencialmente de tales causas separadas, al menos bajo las condiciones experimentales que son la fuente de los datos procesados. Estos datos, ya sean cuantitativos o cualitativos, para procesar con CHIC, son numéricos o binarios o decimales (Lahanier-Reuter et al., 2017).

9.10 Automatización y acceso libre a sus herramientas

En las secciones anteriores, se ha explicado sobre las bondades del software CHIC, que automatiza la mayoría de los procesos del Análisis Estadístico Implicativo, es un software muy flexible y útil. Pero este software tiene algunos inconvenientes: es propietario y esto conlleva a que no se puedan hacer las modificaciones que uno desee, el CHIC trabaja solamente en la plataforma Windows sin dar la posibilidad de utilizarlo en otros ambientes de software tal como Mac OS o Linux. Las personas interesadas en utilizar las técnicas del Análisis Estadístico Implicativo se ven limitadas debido a que al aplicarlas en forma manual conlleva a realizar demasiados cálculos largos y complejos y así aumentar las posibilidades de cometer errores y ocupar mucho tiempo. Siendo el software CHIC propietario limita su utilización a estudiantes, docentes e investigadores, en particular los investigadores al verse restringidos en su utilización no pueden contribuir con su mejora y optimización. Todo lo antes dicho motiva a la creación de un software libre multiplataforma, es por ello que se hacen esfuerzos en la Universidad Jaume 2 de España para crear una versión inicial en un software llamado ASImodel (en ambiente R) y elaborado por Xavier Valls Pla con la dirección del Profesor Pablo Grégori Huerta (Valls, 2014). También en la Universidad de Oriente en Cuba se conocen esfuerzos por automatizar al ASI. La Escuela Superior Politécnica de Chimborazo del Ecuador, utilizando el programa de becas Prometeo y gestionado por el Prof. Rubén Pazmiño, en

los años 2014-2015 invita al Profesor Raphael Couturier para que construya dicho software. Luego de un profundo análisis se decide utilizar como base el software estadístico R, que permitirá liberar el software y utilizarlo en independencia de la plataforma de hardware y software. El proyecto se logró concretar con éxito a finales del 2015, permitiendo masificar la utilización del software Rchic a nivel mundial y facilitar procesos de investigación como es el que se está desarrollando en esta tesis. Se ha podido impulsar la investigación en educación, educación matemática, estadística, matemática aplicada y matemática computacional teniendo resultados a nivel de pregrado y postgrado y abriendo la posibilidad de que nuevos docentes e investigadores se beneficien del Análisis Estadístico Implicativo. La Figura 9.8 evidencia lo anteriormente dicho con las copias del contrato del Profesor Raphael Couturier, el enlace al software Rchic terminado se muestra en la Figura 9.9 y en el Apéndice 11.6 se encuentran detalles sobre el software Rchic.

En la Figura 9.8, se evidencia la vinculación del Profesor Raphael Couturier a la Escuela Superior Politécnica de Chimborazo y específicamente a la Facultad de Ciencias.



20140938 BP

Secretaría de Educación Superior, Ciencia, Tecnología e Innovación — PROMETEO — Presidencia de la República del Ecuador

CONTRATO MODIFICATORIO AL CONTRATO DE BECA PARA DOCENCIA, INVESTIGACIÓN O TRANSFERENCIA DE CONOCIMIENTOS No. 20140044 BP CELEBRADO ENTRE LA SECRETARÍA NACIONAL DE EDUCACIÓN SUPERIOR, CIENCIA, TECNOLOGÍA E INNOVACIÓN Y RAPHAEL PHILIPPE COUTURIER, PARA EL FINANCIAMIENTO DE LA INVESTIGACIÓN CIENTÍFICA CONFORME AL PROYECTO "BECAS PROMETEO".

COMPARECIENTES.-

Comparecen a la celebración del presente contrato modificatorio al Contrato No. 20140044 BP de beca para docencia, investigación o transferencia de conocimientos por una parte, la **SECRETARÍA DE EDUCACIÓN SUPERIOR, CIENCIA, TECNOLOGÍA E INNOVACIÓN**, representada legalmente por **Susana Toro Orellana,** en su calidad de **GERENTE DEL PROYECTO "PROMETEO"** de conformidad con el Acuerdo No. 2014-030 de 05 de marzo de 2014, delegada del Secretario de Educación Superior, Ciencia, Tecnología e Innovación, mediante el Acuerdo No. 2013 – 102 de 11 de septiembre de 2013, a quien en adelante se le denominará **"LA SECRETARÍA"**, y, por otra parte RAPHAEL PHILIPPE COUTURIER, con pasaporte No 09AP67938, de nacionalidad francesa, en su calidad de Investigador, a quien en adelante y para efectos legales del presente documento se le denominará "**EL PROMETEO**", quien comparece por sus propios y personales derechos; las partes acuerdan y se obligan a celebrar el presente contrato al tenor de las siguientes cláusulas:

CLÁUSULA PRIMERA: ANTECEDENTES.-

1.- El 20 de enero del 2014, se celebró el Contrato No. 20140044 BP de beca para docencia, investigación o transferencia de conocimientos por parte del **PROMETEO** RAPHAEL PHILIPPE COUTURIER y la SECRETARÍA, a través del Proyecto "Becas Prometeo".

2.- Dentro del mencionado contrato, la "***CLÁUSULA SÉPTIMA: PLAZO DE VIGENCIA***" establece que: *"El plazo de ejecución del presente contrato es de 04 MESES, contados en los siguientes periodos:*

Primer periodo: Del 27 de enero de 2014 al 26 de febrero de 2014
Segundo periodo: Del 23 de junio de 2014 al 25 de julio de 2014
Tercer periodo: Del 26 de enero de 2015 al 27 de febrero de 2015
Cuarto periodo: Del 22 de junio de 2015 al 24 de julio de 2015

El plazo de ejecución del presente contrato es independiente de la fecha de suscripción del mismo.

*El **BECARIO PROMETEO** no deberá recibir, ni podrá exigir ningún tipo de liquidación, indemnización o remuneración especial. Las partes expresamente aclaran que el presente Contrato de Beca no se renovará automáticamente por ningún motivo".*

3.- La ***"CLÁUSULA DÉCIMA SEGUNDA MODIFICACIONES"***, establece la facultad de *"suscribir contratos modificatorios, ampliatorios o complementarios siempre que se cuente con el informe favorable de la Gerencia del Proyecto…"*.

4.- Mediante Acuerdo Nro. 2013 – 102 de 11 de septiembre de 2013, el Secretario de Educación Superior, Ciencia, Tecnología e Innovación, delega a el/la Gerente del Proyecto "Becas Prometeo", la suscripción de los contratos en el marco del Reglamento para el otorgamiento de becas a docentes e investigadores/as expertos/as de alto nivel a través del Proyecto "Becas Prometeo".

5.- Mediante Acuerdo Nro. 2014-030 de 05 de marzo de 2014, el Secretario de Educación Superior, Ciencia, Tecnología e Innovación, designa a Susana Toro Orellana como Gerente del Proyecto "Becas Prometeo".

6.- En sesión ordinaria celebrada el 31 de julio de 2013, la Gerente del proyecto informa al Comité Ejecutivo de "Becas Prometeo", la aprobación del cambio de fechas solicitado por el PROMETEO, según consta en el Acta PROMETEO- CSP-008-2013.

CLÁUSULA SEGUNDA: OBJETO DEL CONTRATO.-

El objeto del presente contrato es la modificación de la cláusula séptima del Contrato de beca para docencia, investigación o transferencia de conocimientos de No. 20140044 BP con el **PROMETEO** RAPHAEL PHILIPPE COUTURIER, a través del Proyecto "Becas Prometeo" de la siguiente manera:

Sustituir CLÁUSULA SÉPTIMA: PLAZO DE VIGENCIA, por el siguiente texto:

"El plazo de ejecución del presente contrato es de cuatro (04) meses, de conformidad con los siguientes periodos:

I periodo: Del 27 de enero de 2014 al 26 de febrero de 2014
II periodo: Del 23 de junio de 2014 al 25 de julio de 2014
III periodo: Del 22 de junio del 2015 al 21 de agosto del 2015

El plazo de ejecución del presente contrato es independiente de la fecha de suscripción del mismo.

El BECARIO PROMETEO no deberá recibir, ni podrá exigir ningún tipo de liquidación, indemnización o remuneración especial. Las partes expresamente aclaran que el presente Contrato de Beca no se renovará automáticamente por ningún motivo"

CLÁUSULA TERCERA: INALTERABILIDAD DE LAS DEMÁS CLÁUSULAS DEL CONTRATO:

Las demás cláusulas y estipulaciones contempladas en el Contrato de beca para docencia, investigación o transferencia de conocimientos de No. 20140044 BP suscrito el 20 de enero del

Secretaría de Educación Superior, Ciencia, Tecnología e Innovación — PROMETEO — Presidencia de la República del Ecuador

2014, quedan vigentes y con pleno valor jurídico de la manera como inicialmente fueron pactadas entre las partes.

CLÁUSULA CUARTA: ACEPTACIÓN.-

Las partes declaran que se ratifican en todas las cláusulas y declaraciones contenidas en el presente instrumento por así convenir a sus intereses, en constancia de lo cual, se ratifican y aceptan el presente Contrato, y lo suscriben en unidad de acto, en cinco (05) ejemplares de igual tenor y valor jurídico.

Por delegación del Señor Secretario de Educación Superior, Ciencia, Tecnología e Innovación

Dado y firmado en Quito D.M., el 04 DIC 2014

SUSANA TORO ORELLANA
GERENTE DEL PROYECTO PROMETEO

RAPHAEL PHILIPPE COUTURIER
"PROMETEO"

Figura 9.8.- Contrato del Senescyt

En la Figura 9.9, se encuentra el sitio web mediante el cual se permite el acceso público a todos aquellos investigadores que deseen utilizar el ASI mediante el paquete Rchic.

Figura 9.9.- Enlace para descargar el paquete Rchic (*Rchic*, 2016)

9.11 Técnicas clúster y de reglas de asociación incluidas en un mismo paquete

Otra de las ventajas del ASI, es que en el paquete de R libre Rchic se encuentran incluidas todas las técnicas de reglas de asociación y las técnicas clúster como similaridad y cohesión, esto facilitará la optimización en el uso de varios paquetes con sus respectivos paquetes dependientes. Las funciones del ASI se caracterizan en ser las más indispensables para ayudar al usuario a recordarlas, los comandos básicos son:

- similarityTree.
- callHierarchyTree.
- implicativeGraph.

Las opciones adicionales a las funciones principales del ASI están incluidas dentro de sus funciones:

- X: Nombre del archivo en formato csv.
- contribution.supp= FALSE/TRUE: Calcula la contribución de las variables suplementarias.
- typicality.supp= FALSE/TRUE: Calcula la tipicalidad de las variables suplementarias
- computing.mode=1,2, 3: Modo de cálculo 1=implicación clásica, 2= implicación clásica+ confianza, 3=implifiance.
- verbose = FALSE/TRUE: Se dan más detalles.

Las gráficas se generan automáticamente sin necesidad de otras opciones.

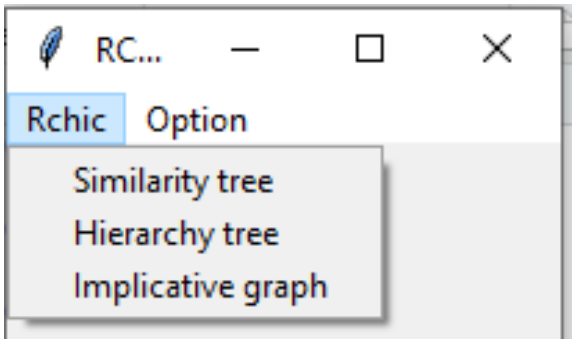


Figura 9.10.- Ambiente amigable de Rchic (Rchic, 2016)

Además de utilizar las funciones, tiene una estructura de menú que permite seleccionar en forma amigable todas las opciones antes indicadas (Figura 9.10).

9.12 Caso de estudio

Se aplican las técnicas clúster y reglas de asociación a un caso de estudio formado por datos de indicadores educativos (Apéndice J), precisando y comentando como el ASI y LA afrontan los 10 aportes presentados en este Capítulo. El caso de estudio es descriptivo y su propósito es exponer en forma articulada como el Análisis Estadístico Implicativo y las Analíticas de Aprendizaje enfrentan los aportes adicionales propuestos, resolviendo el problema del desconocimiento de la situación actual de los 10 aportes adicionales en las técnicas de análisis cluster y técnicas de reglas de asociación frente.

Los datos se tomaron de un estudio de investigación reciente sobre la selección de indicadores educativos a nivel universitario en el caso particular del Ecuador. El estudio de investigación se justifica por su importancia, pues permitirá seleccionar indicadores en los cuales se podrá aplicar técnicas de LA a nivel local y nacional, permitiendo el seguimiento y la mejora del aprendizaje en estos tiempos de pandemia. El estudio está delimitado a los indicadores que se muestran en el Apéndice.

En el caso del ASI se evidencia mediante la utilización práctica de los datos en el software CHIC y en el caso de las otras técnicas tanto clúster como reglas de asociación se trabaja con la documentación de respaldo presentada por cada uno de los paquetes y funciones (Apéndices F y G).

9.12.1 Estudio de los aportes factibles en las técnicas clúster

Se aplican las técnicas clúster a un caso de estudio formado por datos de indicadores educativos, precisando y comentando como el ASI y LA afrontan los 10 aportes presentados.

9.12.1.1 Tipo amplio de datos

El ASI en sus técnicas, acepta los dos tipos de datos universalmente considerados en estadística, en general acepta los datos de tipo numérico y los de tipo atributo ajustados al intervalo [0,1]. En nuestro caso de estudio los datos son cualitativos ordinales y se asignan números a las categorías alto, medio y bajo (1, 0,5 y 0 respectivamente). Estas categorías y sus valores respectivos se asignan realizando grupos colaborativos de especialistas en indicadores educativos.

A continuación, se presentan las características respecto a tipo amplio de datos de las cinco técnicas:

1. Técnicas clúster, función de usuario: dendro_diana, paquete de R: clúster, comando de R: diana, tipo de datos: Numéricos. Todas las variables deben ser numéricas. Y se permiten valores (NA)
2. Técnicas clúster, función de usuario: dendro_variables, paquete de R: CluMix, comando de R: dendro.variables, tipo de datos: cuantitativos, ordinales y categóricos, que se incluyen en el paquete a modo de ilustración.

3. Técnicas clúster, función de usuario: hclust_vector, paquete de R: Fastcluster, comando de R: hclust.vector, tipo de datos: realiza agrupaciones jerárquicas y aglomerativas en datos vectoriales numéricos.
4. Técnica de similaridad, función de usuario: simlrty, paquete de R: Rchic, comando de R: callSimilarityTree, tipo de datos: acepta datos de tipo binario, modal, intervalo y frecuencial.
5. Técnica de cohesión, función de usuario: hrarchy, paquete de R: Rchic, comando de R: callHierarchyTree, tipo de datos: acepta datos de tipo binario, modal, intervalo y frecuencial.

9.12.1.2 Variables suplementarias

En el ASI las variables suplementarias son descriptores que no intervienen en el cálculo. Las variables suplementarias solo se tienen en cuenta en la búsqueda de la contribución o la tipicalidad de las categorías. Para definir una variable suplementaria, se añade al nombre de la variable un espacio y una letra "s". Para crear por ejemplo la variable suplementaria carácter se lo haría por ejemplo como caracter s.

A continuación, se presentan las características respecto a variables suplementarias de las 5 técnicas:

1. Técnicas clúster, función de usuario: dendro_diana, paquete de R: clúster, comando de R: diana, variables: incluyen este tipo de variables, matriz de datos o marco de datos, o matriz de disimilitud u objeto, dependiendo el valor del argumento, pero no permiten algo parecido a las matrices suplementarias.
2. Técnicas clúster, función de usuario: dendro_variables, paquete de R: CluMix, comando de R: dendro.variables, variables: se incluye mapa de calor de datos mixtos con temas en las columnas y variables en el las filas, o no permiten algo parecido a las matrices suplementarias.
3. Técnicas clúster, función de usuario: hclust_vector, paquete de R: fastcluster, comando de R: hclust.vector, variables: no incluye variables de este tipo, pero son parte de un método distinto. Las opciones son las mismas que en el método dist : 'euclidiana' , 'máxima' , 'manhattan' , 'canberra' , 'binaria' y 'minkowski' . Se puede dar cualquier subcadena inequívoca.

4. Técnica de similaridad, función de usuario: simlrty, paquete de R: Rchic, comando de R: callSimilarityTree, variables: acepta variables suplementarias que se ingresan como var s
5. Técnica de cohesión, función de usuario: hrarchy, paquete de R: Rchic, comando de R: callHierarchyTree, variables: acepta variables suplementarias que se ingresan como var s.

9.12.1.3 Nodos significativos

Los nodos significativos son aquellos nodos indispensables para mantener la forma del dendrograma simétrico y del dendrograma asimétrico.

A continuación, se presentan las características respecto a los nodos significativos de las 5 técnicas:

1. Técnicas clúster, función de usuario: dendro_diana, paquete de R: clúster, comando de R: diana, nodos: no tiene la opción para los nodos significativos.
2. Técnicas clúster, función de usuario: dendro_variables, paquete de R: cluMix, comando de R: dendro.variables, nodos: se pueden agregar barras de colores en la parte superior e izquierda del mapa de calor para proporcionar información adicional sobre temas y/o variables, pero en la función no se puede encontrar nodos significativos.
3. Técnicas clúster, función de usuario: hclust_vector, paquete de R: fastcluster, comando de R: hclust.vector, nodos: la función no tiene nodos significativos.
4. Técnica de similaridad, función de usuario: acepta nodos significativos en los árboles de similaridad.
5. Técnica de cohesión, función de usuario: hrarchy, paquete de R: Rchic, comando de R: callHierarchyTree, función de usuario: acepta nodos significativos en los árboles de cohesión.

Aplicación de las técnicas de análisis de datos del Análisis Estadístico Implicativo en *Learning Analytics.*

La Figura 9.11, muestra los nodos significativos en líneas horizontales en rojo, para la técnica de similaridad (árbol de similaridad) del Análisis Estadístico Implicativo.

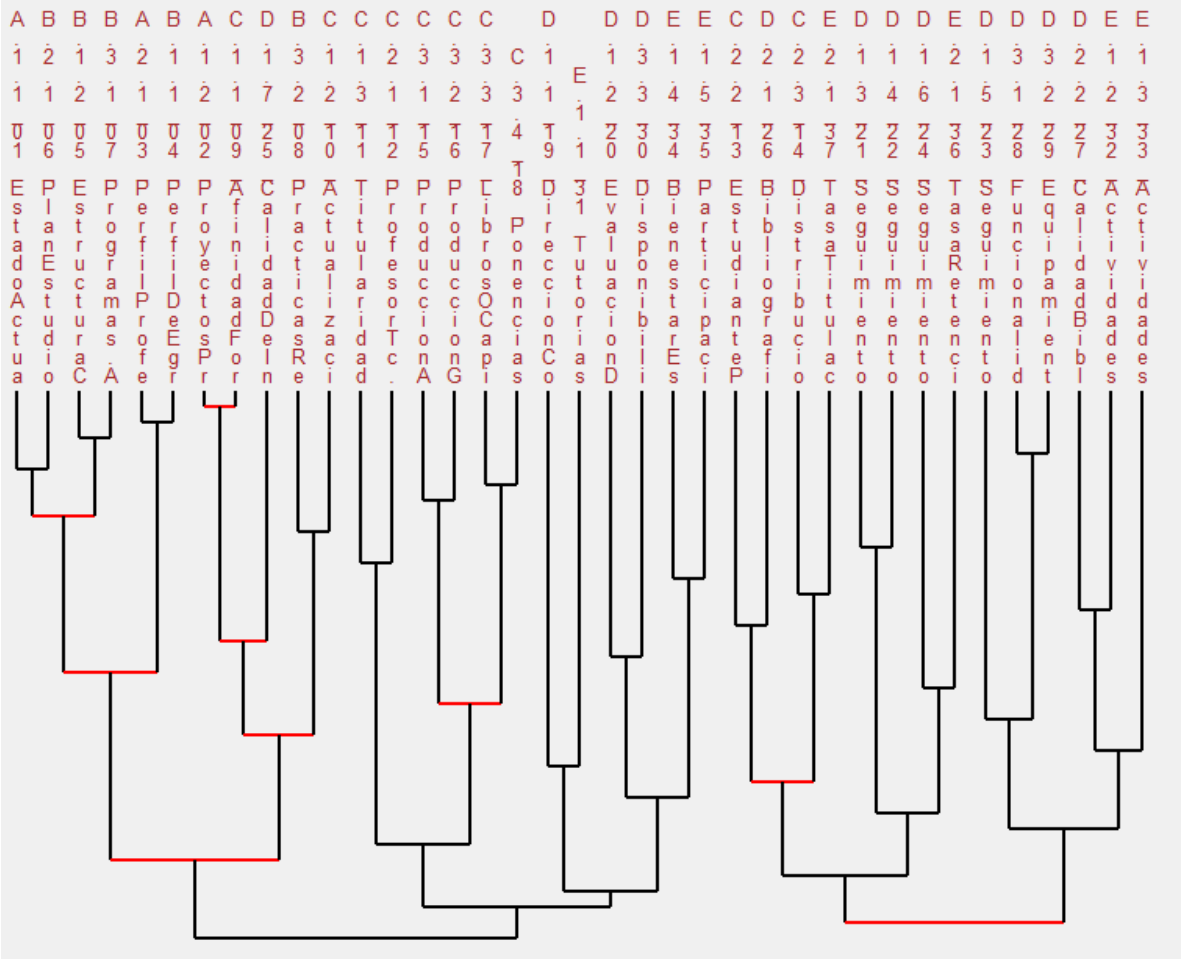


Figura 9.11.- Nodos significativos en la técnica de similaridad (árbol de similaridad)

La Figura 9.12, muestra los nodos significativos en líneas horizontales en rojo, para la técnica de cohesión (árbol de cohesión) del Análisis Estadístico Implicativo.

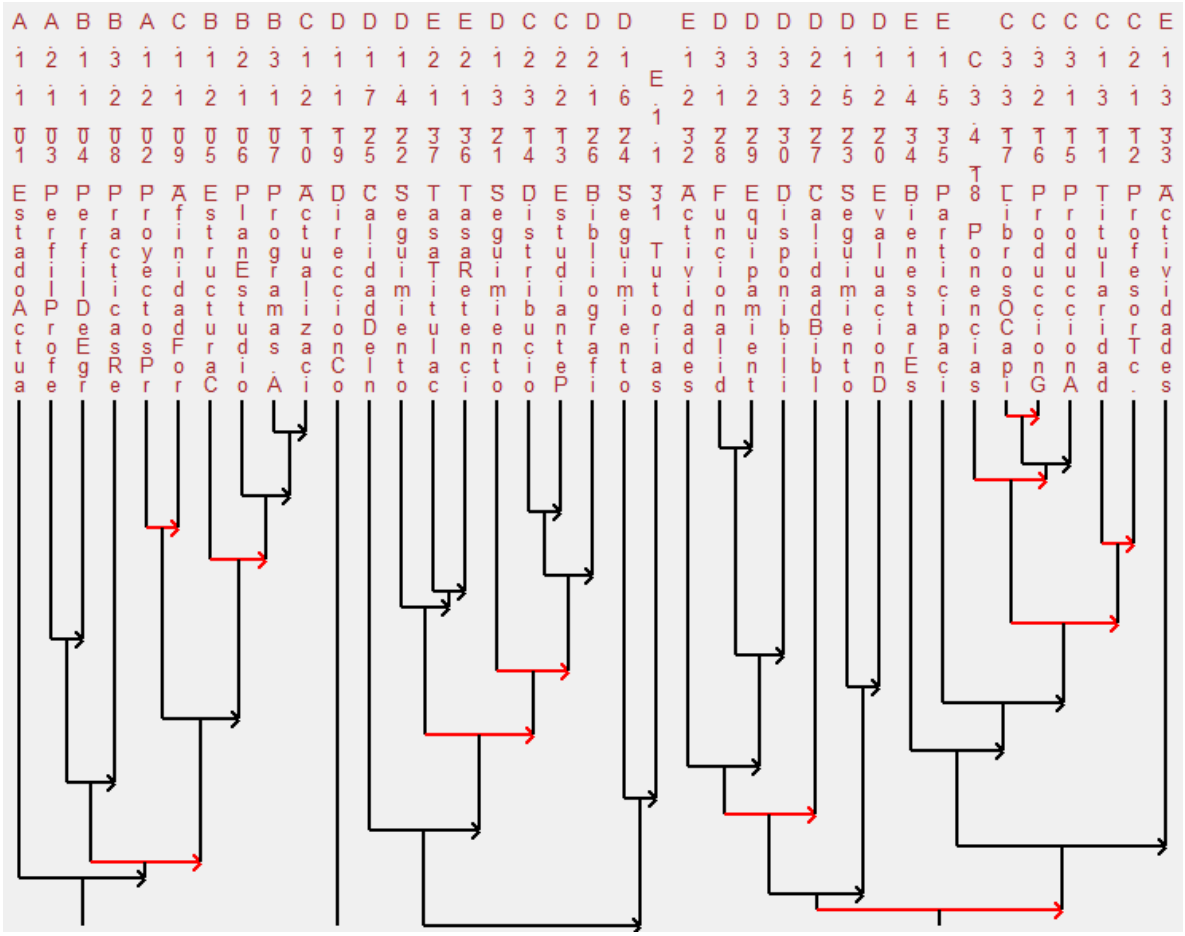


Figura 9.12.- Nodos significativos en la técnica de cohesión (árbol de cohesión)

9.12.1.4 Entropía

La entropía permite al ASI manejar de mejor manera bases de datos relativamente grandes.

A continuación, se presentan las características respecto a la entropía de las 5 técnicas:

1. Técnicas clúster, función de usuario: dendro_diana, paquete de R: clúster, comando de R: diana, cantidad de datos: no acepta trabajar con datos relativamente grandes.

2. Técnicas clúster, función de usuario: dendro_variables, paquete de R: cluMix, comando de R: dendro.variables, cantidad de datos: los datos pueden estar llenados, con un máximo de 200 variables, no acepta trabajar con datos relativamente grandes.
3. Técnicas clúster, función de usuario: hclust_vector, paquete de R: fastcluster, comando de R: hclust.vector, cantidad de datos: conjuntos grandes de datos y vectoriales, utiliza algoritmos de ahorro de memoria que permiten el procesamiento de conjuntos de datos más grandes que los de hclust, acepta trabajar con datos masivos.
4. Técnica de similaridad, función de usuario: simlrty, paquete de R: Rchic, comando de R: callSimilarityTree, cantidad de datos: no acepta trabajar con datos relativamente grandes.
5. Técnica de cohesión, función de usuario: hrarchy, paquete de R: Rchic, comando de R: callHierarchyTree, cantidad de datos: últimamente acepta la opción 3 en computing.mode = 3, que permite trabajar con datos relativamente grandes.

9.12.1.5 Tipicalidad

Determina los sujetos típicos a una estructura de árbol generado. Se utiliza cuando hay variables suplementarias.

A continuación, se presentan las características respecto a tipicalidad de las 5 técnicas:

1. Técnicas clúster, función de usuario: dendro_diana, paquete de R: clúster, comando de R: diana, opciones: tiene una función en el que se puede ver las agrupaciones que se han realizado, diana proporciona el coeficiente divisivo (see diana.object) que mide la cantidad de estructura de agrupamiento encontrada, permite una opción parecida a la tipicalidad.
2. Técnicas clúster, función de usuario: dendro_variables, paquete de R: cluMix, comando de R: dendro.variables, opciones: no permite tipicalidad.
3. Técnicas clúster, función de usuario: hclust_vector, paquete de R: fastcluster, comando de R: hclust.vector, opciones: no permite tipicalidad..

4. Técnica de similaridad, función de usuario: simlrty, paquete de R: Rchic, comando de R: callSimilarityTree, opciones: acepta la opción tipicalidad con la opción typicality.supp=TRUE.
5. Técnica de cohesión, función de usuario: hrarchy, paquete de R: Rchic, comando de R: callHierarchyTree, opciones: acepta la opción tipicalidad con la opción typicality.supp=TRUE.

9.12.1.6 Contribución

Cuantifica la contribución de los sujetos a una estructura de árbol generado. Se utiliza cuando hay variables suplementarias.

A continuación, se presentan las características respecto a contribución de las 5 técnicas:

1. Técnicas clúster, función de usuario: dendro_diana, paquete de R: clúster, comando de R: diana, opciones: mediante este paquete podemos tener contribución parcial ya que éste se encarga de ver qué variable contribuye más. El algoritmo busca primero la observación más dispar (es decir, cuál tiene la mayor disimilitud promedio con las otras observaciones del grupo seleccionado).
2. Técnicas clúster, función de usuario: dendro_variables, paquete de R: cluMix, comando de R: dendro _variables, opciones: no permite contribución.
3. Técnicas clúster, función de usuario: hclust_vector, paquete de R: fastcluster, comando de R: hclust.vector, opciones: no permite contribución.
4. Técnica de similaridad, función de usuario: simlrty, paquete de R: Rchic, comando de R: callSimilarityTree, opciones: acepta la opción contribución con la opción contribution.supp=TRUE.
5. Técnica de cohesión, función de usuario: hrarchy, paquete de R: Rchic, comando de R: callHierarchyTree, opciones: acepta la opción contribución con la opción contribution.supp=TRUE.

9.12.1.7 Escenarios de análisis y experimentación

Son escenarios propios incluidos en las funciones que permiten experimentar con las variables en estudio.

A continuación, se presentan las características respecto a escenarios de análisis y experimentación de las 5 técnicas:

1. Técnicas clúster, función de usuario: dendro_diana, paquete de R: clúster, comando de R: diana, escenarios propios: no, ya que este paquete tiene funciones genéricas directas, y el usuario no puede manipularlas, n objeto de clase "diana" que representa el agrupamiento; esta clase tiene métodos para las siguientes funciones genéricas: imprimir, resumen, trazar.
2. Técnicas clúster, función de usuario: dendro_variables, paquete de R: cluMix, comando de R: dendro_variables, escenarios propios: no acepta la opción para crear escenarios manipulando las variables.
3. Técnicas clúster, función de usuario: hclust_vector, paquete de R: fastcluster, comando de R: hclust.vector, escenarios propios: no posee escenarios propios.
4. Técnica de similaridad, función de usuario: simlrty, paquete de R: Rchic, comando de R: callSimilarityTree, escenarios propios: acepta la opción para crear escenarios manipulando las variables.
5. Técnica de cohesión, función de usuario: hrarchy, paquete de R: Rchic, comando de R: callHierarchyTree, escenarios propios: acepta la opción para crear escenarios manipulando las variables.

Aplicación de las técnicas de análisis de datos del Análisis Estadístico Implicativo en *Learning Analytics.*

La Figura 9.13, muestra los botones de selección de las variables en estudio, para generar escenarios de análisis y experimentación para similaridad o cohesión o implicación en el Análisis Estadístico Implicativo, las tres técnicas tienen la misma opción.

Figura 9.13.- Escenarios de análisis y experimentación para similaridad, cohesión e implicación en el Análisis Estadístico Implicativo

9.12.1.8 Visualizaciones sencillas de interpretar

Son visualizaciones propias incluidos en las funciones que permiten observar las salidas con las variables en estudio.

A continuación, se presentan las características respecto a visualizaciones sencillas de interpretar, de las 5 técnicas:

1. Técnicas clúster, función de usuario: dendro_diana, paquete de R: clúster, comando de R: diana, visualización propia: si admite una pantalla gráfica novedosa y parcialmente sencilla, la cual se hace uso con el comando plot.diana.
2. Técnicas clúster, función de usuario: dendro_variables, paquete de R: CluMix, comando de R: dendro.variables, visualización propia: no, ya que el usuario lo puede manipular a su conveniencia, pero utilizando otras funciones de apoyo como por ejemplo, coloreado usando el paquete dendextend y luego combinado con mix.heatmap.

3. Técnicas clúster, función de usuario: hclust_vector, paquete de R: fastcluster, comando de R: hclust.vector, visualización propia: presenta una visualización parcialmente simple, aunque raras veces pueda presentar falencias en sus resultados. En casos extremos, afectan a todo el resultado de la agrupación debido a la inherente inestabilidad de la naturaleza de los esquemas de agrupamiento.
4. Técnica de similaridad, función de usuario: simlrty, paquete de R: Rchic, comando de R: callSimilarityTree, visualización propia: utiliza como visualización el dendrograma simétrico sencillo.
5. Técnica de cohesión, función de usuario: hrarchy, paquete de R: Rchic, comando de R: callHierarchyTree, visualización propia: utiliza como visualización el dendrograma asimétrico sencillo.

9.12.1.9 Automatización y acceso libre a sus herramientas

Si las técnicas están automatizadas y se tiene acceso libre a ellas.

A continuación, se presentan las características respecto a automatización y acceso libre a sus herramientas, de las 5 técnicas:

1. Técnicas clúster, función de usuario: dendro_diana, paquete de R: clúster, comando de R: diana, licencia: libre, ya que si tenemos acceso a sus funciones.
2. Técnicas clúster, función de usuario: dendro_variables, paquete de R: cluMix, comando de R: dendro.variables, licencia: no se indica en el documento.
3. Técnicas clúster, función de usuario: hclust_vector, paquete de R: fastcluster, comando de R: hclust.vector, licencia: libre.
4. Técnica de similaridad, función de usuario: simlrty, paquete de R: Rchic, comando de R: callSimilarityTree, licencia: no se explicita la licencia, pero es de libre utilización.
5. Técnica de cohesión, función de usuario: hrarchy, paquete de R: Rchic, comando de R: callHierarchyTree, licencia: no se explicita la licencia, pero es de libre utilización.

Aplicación de las técnicas de análisis de datos del Análisis Estadístico Implicativo en *Learning Analytics.*

La Figura 9.14, muestra el acceso libre a las herramientas de Rchic, además se entiende que las herramientas están automatizadas.

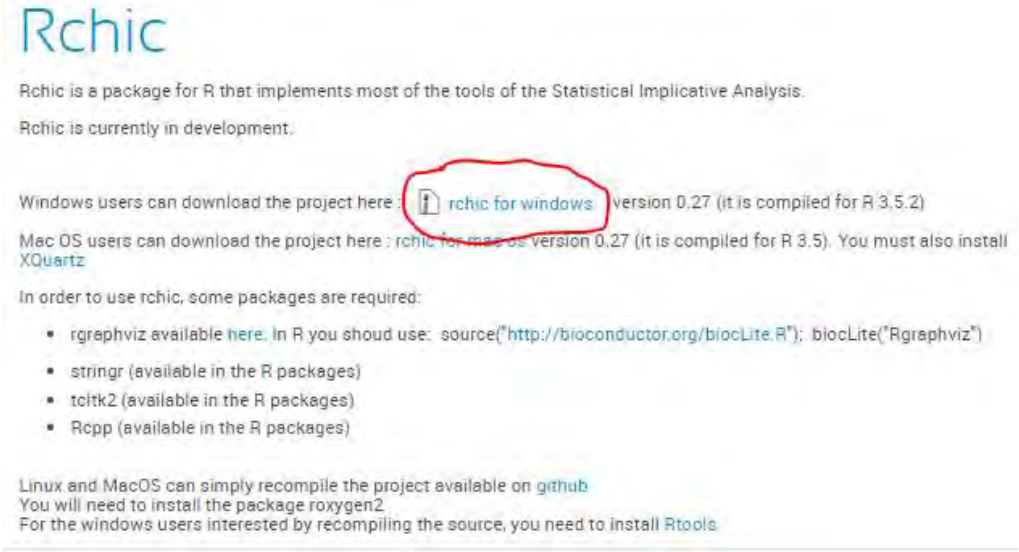


Figura 9.14.- Automatización y acceso libre a las herramientas de Rchic (*Rchic*, 2016)

9.12.1.10 Técnicas clúster y de reglas de asociación incluidas en un mismo paquete

A continuación, se presentan las características respecto a las técnicas clúster y de reglas de asociación incluidas en un mismo paquete, de las 5 técnicas:

1. Técnicas clúster, función de usuario: dendro_diana, paquete de R: clúster, comando de R: diana, incluida técnicas de reglas de asociación: la técnica clúster no integra las técnicas de reglas de asociación.
2. Técnicas clúster, función de usuario: dendro_variables, paquete de R: factoextra, comando de R: dendro.variables, incluida técnicas de reglas de asociación: la técnica clúster no integra las técnicas de reglas de asociación.
3. Técnicas clúster, función de usuario: hclust_vector, paquete de R: fastcluster, comando de R: hclust.vector, incluida técnicas de reglas de asociación: la técnica clúster no integra las técnicas de reglas de asociación.

4. Técnica de similaridad, función de usuario: simlrty, paquete de R: Rchic, comando de R: callSimilarityTree, comando de R: Se incluyen las opciones de similaridad, cohesión e implicación en el mismo paquete.
5. Técnica de cohesión, función de usuario: hrarchy, paquete de R: Rchic, comando de R: callHierarchyTree, comando de R: Se incluyen las opciones de similaridad, cohesión e implicación en el mismo paquete.

Aplicación de las técnicas de análisis de datos del Análisis Estadístico Implicativo en *Learning Analytics.*

La Figura 9.15, muestra el menú de Rchic donde se ven las técnicas clúster y de reglas de asociación integradas en un mismo paquete en el Análisis Estadístico Implicativo, a las tres técnicas se puede acceder con la misma opción.

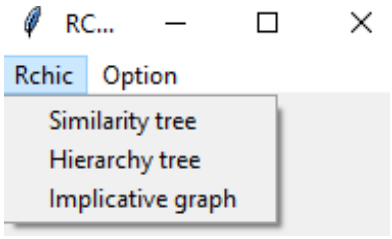


Figura 9.15.- Técnicas clúster y de reglas de asociación incluidas en un mismo paquete

9.12.2 Estudio de los aportes factibles en las técnicas de reglas de asociación

Se aplican las reglas de asociación a un caso de estudio formado por datos de indicadores educativos, precisando y comentando como el ASI y LA afrontan los 10 aportes presentados.

9.12.2.1 Tipo amplio de datos

El ASI acepta los dos tipos de datos universalmente considerados en estadística, en general acepta los datos de tipo numérico y los de tipo atributo ajustados al intervalo [0,1]. En el caso de estudio los datos son cualitativos y se asignan números a las categorías alto, medio y bajo. Estas categorías y sus valores respectivos se asignan realizando grupos colaborativos de especialistas en indicadores educativos.

A continuación, se presentan las características respecto al tipo amplio de datos de las 4 técnicas:

1. Técnicas de reglas de asociación, función de usuario: met_apriori, paquete de R: arules, comando de R: apriori, tipo de datos: casi cualquier estructura de datos. Objeto de transacciones de clase o cualquier estructura de datos que pueda ser coaccionada (por ejemplo, matriz binaria, data.frame).
2. Técnicas de reglas de asociación, función de usuario: met_eclat, paquete de R: arules, comando de R: eclat, tipo de datos: casi cualquier estructura de datos. Objeto de transacciones de clase o cualquier estructura de datos que pueda ser coaccionada (por ejemplo, matriz binaria, data.frame).
3. Técnicas de reglas de asociación, función de usuario: met_weclat, arules, comando de R: weclat, tipo de datos: casi cualquier tipo de datos.
4. Técnica de implicación, función de usuario: met_ASI, paquete de R: Rchic, tipo de datos: implicativeGraph, tipo de datos acepta datos de tipo binario, modal, intervalo y frecuencial.

9.12.2.2 Variables suplementarias

La implicación no acepta variables suplementarias.

A continuación, se presentan las características respecto a variables suplementarias de las 4 técnicas:

1. Técnicas de reglas de asociación, función de usuario: met_apriori, paquete de R: arules, comando de R: apriori, variables: no acepta variables suplementarias.
2. Técnicas de reglas de asociación, función de usuario: met_eclat, paquete de R: arules, comando de R: eclat, variables: no acepta variables suplementarias.
3. Técnicas de reglas de asociación, función de usuario: met_weclat, arules, comando de R: weclat, variables: no acepta variables suplementarias.
4. Técnica de implicación, función de usuario: met_ASI, paquete de R: Rchic, tipo de datos: implicativeGraph, variables: no acepta variables suplementarias.

9.12.2.3 Nodos significativos

La implicación no acepta nodos significativos.

A continuación, se presentan las características respecto a nodos significativos de las 4 técnicas:

1. Técnicas de reglas de asociación, función de usuario: met_apriori, paquete de R: arules, comando de R: apriori, nodos: no se aceptan nodos significativos.
2. Técnicas de reglas de asociación, función de usuario: met_eclat, paquete de R: arules, comando de R: eclat, nodos: no se aceptan nodos significativos.
3. Técnicas de reglas de asociación, función de usuario: met_weclat, arules, comando de R: weclat, nodos: no se aceptan nodos significativos.
4. Técnica de implicación, función de usuario: met_ASI, paquete de R: Rchic, tipo de datos: implicativeGraph, nodos: no se aceptan nodos significativos.

9.12.2.4 Entropía y conjuntos grandes de datos

La entropía permite al ASI manejar de mejor manera bases de datos relativamente grandes.

A continuación, se presentan las características respecto a entropía de las 4 técnicas:

1. Técnicas de reglas de asociación, función de usuario: met_apriori, paquete de R: arules, comando de R: apriori, cantidad de datos: no se indica en el documento.
2. Técnicas de reglas de asociación, función de usuario: met_eclat, paquete de R: arules, comando de R: eclat, cantidad de datos: no se indica en el documento.
3. Técnicas de reglas de asociación, función de usuario: met_weclat, arules, comando de R: weclat, cantidad de datos: no se indica en el documento.
4. Técnica de implicación, función de usuario: met_ASI, paquete de R: Rchic, cantidad de datos: últimamente acepta la opción 3 en computing.mode = 3, que permite trabajar con datos relativamente grandes.

9.12.2.5 Tipicalidad

La implicación no permite variables suplementarias, por tanto tampoco la tipicalidad.

A continuación, se presentan las características respecto a tipicalidad de las 4 técnicas:

1. Técnicas de reglas de asociación, función de usuario: met_apriori, paquete de R: arules, comando de R: apriori, opciones: no acepta tipicalidad.
2. Técnicas de reglas de asociación, función de usuario: met_eclat, paquete de R: arules, comando de R: eclat, opciones: no acepta tipicalidad.
3. Técnicas de reglas de asociación, función de usuario: met_weclat, arules, comando de R: weclat, opciones: no acepta tipicalidad.
4. Técnica de implicación, función de usuario: met_ASI, paquete de R: Rchic, tipo de datos: implicativeGraph, opciones: no acepta tipicalidad.

9.12.2.6 Contribución

La implicación no permite variables suplementarias, por tanto tampoco contribución.

A continuación, se presentan las características respecto a contribución de las 4 técnicas:

1. Técnicas de reglas de asociación, función de usuario: met_apriori, paquete de R: arules, comando de R: apriori, opciones: no acepta contribución.
2. Técnicas de reglas de asociación, función de usuario: met_eclat, paquete de R: arules, comando de R: eclat, opciones: no acepta contribución.
3. Técnicas de reglas de asociación, función de usuario: met_weclat, arules, comando de R: weclat, opciones: no acepta contribución.
4. Técnica de implicación, función de usuario: met_ASI, paquete de R: Rchic, tipo de datos: implicativeGraph, opciones: no acepta contribución.

9.12.2.7 Escenarios de análisis y experimentación

Son escenarios propios incluidos en las funciones que permiten experimentar con las variables en estudio.

A continuación, se presentan las características respecto a escenarios de análisis y experimentación de las 4 técnicas:

1. Técnicas de reglas de asociación, función de usuario: met_apriori, paquete de R: arules, comando de R: apriori, escenarios propios: el usuario solo puede hacer manipulaciones con 10 elementos ya que si no hará que se colapse el tiempo de ejecución y la memoria. Ya que al final indicará de manera directa el resultado. Devuelve un objeto de reglas de clase o conjuntos de elementos. No acepta la opción de escenarios de análisis y experimentación.

2. Técnicas de reglas de asociación, función de usuario: met_eclat, paquete de R: arules, comando de R: eclat, escenarios propios: se puede utilizar con varios fines. Se puede utilizar para realizar minería de reglas de asociación ponderada (WARM). No acepta la opción de escenarios de análisis y experimentación.
3. Técnicas de reglas de asociación, función de usuario: met_weclat, arules, comando de R: weclat, escenarios propios: el usuario si puede manejar los datos de acuerdo a lo que necesita, ya que siempre devolverá algo como apoyo. “Devuelve un objeto de clase de conjuntos de elementos. Tenga en cuenta que el soporte ponderado se devuelve en calidad como columna apoyo. No acepta la opción de escenarios de análisis y experimentación.
4. Técnica de implicación, función de usuario: met_ASI, paquete de R: Rchic, comando de R: implicativeGraph, escenarios propios: acepta la opción contribución para crear escenarios manipulando las variables y los niveles de implicación y confidence.

La Figura 9.16, muestra la barra de opciones del gráfico implicativo para poder generar los escenarios de análisis y experimentación.

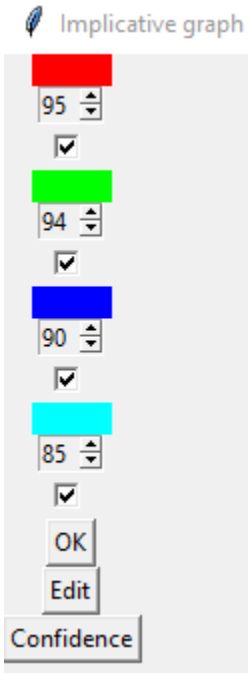


Figura 9.16.- Escenarios de análisis y experimentación para el gráfico implicativo

9.12.2.8 Visualizaciones sencillas de interpretar

Son visualizaciones propias incluidos en las funciones que permiten observar las salidas con las variables en estudio.

A continuación, se presentan las características respecto a visualizaciones sencillas de interpretar de las 4 técnicas:

1. Técnicas de reglas de asociación, función de usuario: met_apriori, paquete de R: arules, comando de R: apriori, visualización propia: si tenemos más de 10 elementos, éste se nos colapsara, de tal modo que no es tan sencillo trabajar con muchos elementos. La salida mostrará si alcanza estos límites en la línea de comprobación de subconjuntos de la salida. Parcialmente sencillo.
2. Técnicas de reglas de asociación, función de usuario: met_eclat, paquete de R: arules, comando de R: eclat, visualización propia: ya que se puede utilizar para

generar reglas mediante conjuntos. Generar reglas a partir de los conjuntos de elementos encontrados. Parcialmente sencillo.

3. Técnicas de reglas de asociación, función de usuario: met_weclat, arules, comando de R: weclat, visualización propia: ya que no es tan difícil de ver visualizaciones. Devuelve un objeto de clase de conjuntos de elementos. Parcialmente sencillo.
4. Técnica de implicación, función de usuario: met_ASI, paquete de R: Rchic, tipo de datos: implicativeGraph, no visualización propia: admite el grafo implicativo que es un gráfico muy sencillo de interpretar.

Aplicación de las técnicas de análisis de datos del Análisis Estadístico Implicativo en *Learning Analytics.*

La Figura 9.17, muestra el gráfico de implicación donde se observan las reglas de cuasi-implicación entre las 7 variables.

Figura 9.17.- Visualización de gráfico de implicación, sencillo de interpretar

9.12.2.9 Automatización y acceso libre a sus herramientas

Si las técnicas están automatizadas y se tiene acceso libre a ellas.

A continuación, se presentan las características respecto a automatización y acceso libre a sus herramientas de las 4 técnicas.

1. Técnicas de reglas de asociación, función de usuario: met_apriori, paquete de R: arules, comando de R: apriori, licencia: libre, pero no se tiene un libre acceso a sus herramientas.
2. Técnicas de reglas de asociación, función de usuario: met_eclat, paquete de R: arules, comando de R: eclat, licencia: libre ya que si podemos acceder a sus funciones.
3. Técnicas de reglas de asociación, función de usuario: met_weclat, arules, comando de R: weclat, licencia: libre, el código C puede ser interrumpido por CTRL-C. Esto es conveniente, pero produce que el código no se pueda limpiar de su memoria interna.
4. Técnica de implicación, función de usuario: met_ASI, paquete de R: Rchic, tipo de datos: implicativeGraph, licencia: libre

9.12.2.10 Técnicas clúster y de reglas de asociación incluidas en un mismo paquete

A continuación, se presentan las características respecto a técnicas clúster y de reglas de asociación incluidas en un mismo paquete, de las 4 técnicas.

1. Técnicas de reglas de asociación, función de usuario: met_apriori, paquete de R: arules, comando de R: apriori, incluida técnicas clúster: la técnica de reglas de asociación no integra las técnicas clúster.
2. Técnicas de reglas de asociación, función de usuario: met_eclat, paquete de R: arules, comando de R: eclat, incluida técnicas clúster: la técnica de reglas de asociación no integra las técnicas clúster.
3. Técnicas de reglas de asociación, función de usuario: met_weclat, arules, comando de R: weclat, incluida técnicas clúster: la técnica de reglas de asociación no integra las técnicas clúster.
4. Técnica de implicación, función de usuario: met_ASI, paquete de R: Rchic, comando de R: implicativeGraph, incluida técnicas clúster: se incluyen las opciones de similaridad, cohesión e implicación en el mismo paquete.

9.13 Conclusiones

A continuación, se indican las conclusiones del caso de estudio, considerando las técnicas cluster (Tabla 9.1), la técnica de reglas de asociación (Tabla 9.2) y las siguientes diez opciones adicionales del ASI, explicadas en este Capítulo.

El Análisis Estadístico Implicativo (ASI) tiene opciones adicionales de análisis (Couturier y Almouloud, 2009) tales como:

- TIPO DE DATOS AMPLIO: Acepta aplicarse automáticamente a diferentes tipos de datos (Zamora, Gregori, & Orús, 2009), binaria (0-1), modal (1 a 5), frecuencial (0 a 7), intervalo (rangos numéricos).
- VARIABLES SUPLEMENTARIAS: Se toma en cuenta en la búsqueda de contribución o tipicalidad.
- NODOS SIGNIFICATIVOS: Los nodos correspondientes a una clasificación compatible lo mejor posible con los valores y la calidad del agrupamiento obtenido (Zamora et al., 2009).
- ENTROPÍA Y CONJUNTOS GRANDES DE DATOS: Diferentes niveles de computación dependiendo del número de datos.
- TIPICALIDAD: El cálculo de tipicalidad indica cuál es el individuo o variable más típica o más contributiva, y también qué otras categorías pueden conducir a los mismos valores de similaridad y cohesión (Zamora et al., 2009).
- CONTRIBUCIÓN: El cálculo de contribución indica cuál es el individuo o variable más contributiva, y también qué otras categorías pueden conducir a los mismos valores de contribución (Zamora et al., 2009).
- ESCENARIOS DE ANÁLISIS Y EXPERIMENTACIÓN: Ambientes de experimentación que ayudan en la toma de decisiones.
- VISUALIZACIONES SENCILLAS DE INTERPRETAR: Herramientas visuales fáciles de interpretar, como árboles simétricos, árboles asimétricos y grafos dirigidos.
- AUTOMATIZACIÓN Y ACCESO LIBRE A SUS HERRAMIENTAS: Con el paquete Rchic (Raphael Couturier, 2016).
- TÉCNICAS CLÚSTER Y DE REGLAS DE ASOCIACIÓN INCLUIDAS EN UN MISMO PAQUETE: El paquete Rchic contiene los análisis y representaciones gráficas de similaridad, cohesión e implicación (Raphael Couturier, 2016).

La Tabla 9.1, presenta como columnas las cinco técnicas clúster y como filas los diez aportes analizados.

Tabla 9.1.- Técnicas clúster según los posibles aportes

Nº	APORTES	DETALLE	TÉCNICAS CLÚSTER				
			diana	dendro.variable.c	hclust.vector	SimilarityTree	callHierarchyTree
1	Tipo amplio de datos	Nominales	No	Si	No	No	No
		Ordinales	No	Si	No	Transformado en números y ajustados al intervalo [0,1]	Transformado en números y ajustados al intervalo [0,1]
		Razón	Si	Si	Si	Ajustados al intervalo [0,1]	Ajustados al intervalo [0,1]
		Intervalo	Si	Si	Si	Ajustados al intervalo [0,1]	Ajustados al intervalo [0,1]
2	Variables suplementarias		No	No	No	caracter s.	caracter s.
3	Nodos Significativos		No	No	No	Si	Si
4	Entropía		No	No	Si	No	`callHierarchyTree(x, contribution.supp=FALSE, typicality.supp=FALSE, computing.mode=3, verbose = FALSE)`
5	Tipicalidad		Parcial	No	No	`callSimilarityTree(x, contribution.supp=FALSE, typicality.supp=TRUE, verbose=FALSE)`	`callHierarchyTree(x, contribution.supp=FALSE, typicality.supp=TRUE, computing.mode=1, verbose = FALSE)`
6	Contribución		Parcial	No	No	`callSimilarityTree(x, contribution.supp=TRUE, typicality.supp=FALSE, verbose=FALSE)`	`callHierarchyTree(x, contribution.supp=TRUE, typicality.supp=FALSE, computing.mode=1, verbose=FALSE)`
7	Escenarios de análisis y experimentación		No	No	No	Si	Si
8	Visualizaciones sencillas de interpretar		Parcial	No	Parcial	Si	Si

9	Automatización y acceso libre a sus herramientas	Si	Si	Si	Si	Si
10	Técnicas clúster y de reglas de asociación incluidas en un mismo paquete	No	No	No	Si	Si

Los datos de la tabla 9.1, son el resultado a la pregunta si la técnica clúster puede aportar en el ámbito respectivo. Por ejemplo, la técnica callHierarchyTree puede aportar en todo y además se indica el detalle de cómo puede aportar, es decir cómo llamar a la función o cómo se debe declarar la variable suplementaria en la base de datos a estudiar. La palabra "parcialmente" indica que tiene alguna opción similar al aporte específico, pero no es exactamente igual.

La Tabla 9.2, presenta como columnas las cuatro técnicas de reglas de asociación y como filas los diez aportes analizados.

Tabla 9.2.- Técnica de reglas de asociación según los posibles aportes

Nº	APORTES	DETALLE	TÉCNICAS DE REGLAS DE ASOCIACIÓN			
			apriori	eclat	weclat	implicativeGraph
1	Tipo amplio de datos	Nominales	Si	Si	Si	No
		Ordinales	Si	Si	Si	Transformado en números y ajustados al intervalo [0,1]
		Razón	Si	Si	Si	Ajustados al intervalo [0,1]
		Intervalo	Si	Si	Si	Ajustados al intervalo [0,1]
2	Variables suplementarias		No	No	No	No
3	Nodos Significativos		No	No	No	No
4	Entropía		No se indica	No se indica	No Se indica	callSimilarityTree(x, contribution.supp=FALSE, typicality.supp=FALSE, verbose=FALSE) sm<-similarity_matrix implicativeGraph(sm, list.variables=list.variables, computing.mode = 3, complete.graph = 0)
5	Tipicalidad		No	No	No	No
6	Contribución		No	No	No	No
7	Escenarios de análisis y experimentación		No	No	No	Si

8	Visualizaciones sencillas de interpretar	Parcial	Parcial	Parcial	Si
9	Automatización y acceso libre a sus herramientas	Si	Si	Si	Si
10	Técnicas clúster y de reglas de asociación incluidas en un mismo paquete	No	No	No	Si

Los datos de la tabla 9.2, son el resultado a la pregunta si la técnica de reglas de asociación puede aportar en el ámbito respectivo. La técnica weclat, por ejemplo, admite todos los tipos de datos, no tiene un concepto parecido a los nodos significativos ni variables similares, etc., La palabra “No se indica” muestra que no se ha podido determinar si tiene o no un determinado aporte.

Las nuevas opciones podrían servir para atenuar las siguientes dificultades que se pueden presentar en LA.

Análisis inadecuados: En el artículo de la teoría a la acción (Wiley et al., 2020) se identifica que en el campo de la educación se reúnen una diversa y gran cantidad de datos de fuentes cambiantes, pero la forma en que se analizan no es adecuada, en particular en la educación superior (Siemens y Long, 2011). La obtención, recolección y análisis de datos en educación tiene desafíos significativos en varias etapas (Axelsen et al., 2020).

Acceso limitado a herramientas de análisis global: Los factores analizados por Ferguson hacen notar que hay también un problema de acceso a las herramientas de análisis, es decir, ¿Cómo hacer para que las herramientas de análisis en LA sean accesibles por un grupo diferente de usuarios y con necesidades también diferentes? Al encontrarnos en la sociedad del conocimiento global, el hombre está en búsqueda constante de nuevos métodos que simplifiquen las actividades en cualquier área o campo, el problema de acceso global, es decir ¿Cómo hacer para que las herramientas de análisis en LA sean accesibles asincrónicamente y desde lugares remotos?, es así que los problemas de LA no solo son tecnológicos, pues su constante desarrollo exige a LA tener acceso, adecuar o incluir nuevas herramientas de análisis en la resolución de problemas educativos (Ferguson, 2014).

Selección inadecuada de técnicas (y opciones) de análisis: Se nota también la existencia de un problema de selección de técnicas de análisis y opciones apropiadas utilizadas en LA de acuerdo con el tipo de datos con los que se cuente en dependencia a las necesidades, dificultades y problemas educativos a tratar.

Lentitud de cálculo: Dentro de la etapa de procesamiento de datos, las técnicas de análisis son necesarias en LA y si las empleadas no fueran eficientes (por ejemplo en cuanto a complejidad algorítmica) se crea el problema de lentitud de cálculo o que ciertos problemas sean computacionalmente irresolubles (Laxmi et al., 2020).

Acceso a técnicas de análisis con ingresos de tipo amplio: Uno de los desafíos del LA, es utilizar herramientas de análisis aplicables a un amplio tipo de datos generados por el contexto educativo a estudiar (Ferguson, 2014).

Con la experiencia del caso de estudio y el trabajo con el ASI, se realiza un acercamiento a los posibles aportes con las opciones adicionales del ASI (Figura 9.18), asociados a las dificultades.

Posibles dificultades de LA	Análisis inadecuados	Variables suplementarias
		Nodos significativos
		Tipicalidad
		Contribución
	Acceso limitado a herramientas de análisis global	Automatización y acceso libre a sus herramientas
		Técnicas clúster y de reglas de asociación incluidas en un mismo paquete
		Visualizaciones sencillas de interpretar
	Selección inadecuada de técnicas (y opciones) de análisis	Escenarios de análisis y experimentación.
		Técnicas clúster y de reglas de asociación incluidas en un mismo paquete
		Tipo de datos amplio
	Lentitud de cálculo	Entropía y conjuntos grandes de datos
	Acceso a técnicas de análisis con ingresos de tipo amplio	Tipo de datos amplio

Figura 9.18.- Aportes factibles del ASI a LA

C a p í t u l o 10mo | CONCLUSIONES

En este capítulo se destacan las principales conclusiones extraídas y resumidas de cada uno de los capítulos anteriores

10 Capítulo.- Conclusiones

En este Capítulo, primero se relacionan las conclusiones con los objetivos, preguntas de investigación, hipótesis y problema planteados en la sección introductoria y propuestos inicialmente en el plan de investigación (R. Pazmiño-Maji, 2014a). Luego se presentan las principales conclusiones agrupadas en tres tipos de aportes:

- Aportes del ASI a LA desde la definición, fuentes de datos y etapas de LA, desarrollados en el Capítulo 5.- Aportes del ASI a LA (Sección 10.3).
- Aportes a LA desde las técnicas de análisis del ASI, desarrollados en el Capítulo 6.- Aporte con Técnicas ASI a LA (Sección 10.3).
- Aportes desde la complejidad algorítmica de LA y ASI, desarrollados en los Capítulos 7 y 8 (Sección 10.4).

Finalmente se presentan los resultados asociados (Sección 10.5) y las futuras investigaciones (Sección 10.6) derivadas de este trabajo de investigación.

10.1 Relación con objetivos, preguntas de investigación, hipótesis y problema

Con el propósito de profundizar en el aporte del ASI a LA, se utilizaron varias revisiones sistemáticas por aproximadamente 11 años y la teoría de conjuntos, se demostró que el ASI y LA al menos son comunes en el campo educativo y en tres métodos de análisis: minería de relaciones, descubrimiento de estructura y estadísticas, según las clasificaciones de Baker e Inventado y Papamitsiou y Economides. Se profundizó en la comparación de la complejidad algorítmica entre las técnicas clúster y de reglas de asociación entre LA y ASI, debido a que LA frecuentemente necesita el análisis de grandes cantidades y nuevos tipos de datos surgidos de fuentes diversas tales como tuits, páginas web, redes sociales, emails, foros, chats, etc., para ello se utilizó un diseño pre-experimental del tipo un solo grupo aleatorio de la forma RGXO1. Esta sección tiene el propósito de verificar el cumplimiento y relación entre los objetivos específicos, objetivo general, preguntas de investigación, hipótesis, problema y tema de este trabajo de tesis.

10.1.1 Sobre los objetivos específicos

Se propusieron cuatro objetivos específicos en el Capítulo 1.- Introducción, Sección 1.5 Objetivos específicos.

- El primer objetivo específico fue “Describir las técnicas de análisis de datos utilizadas en el Análisis Estadístico Implicativo”, que se cumplió en el Capítulo 3.- Aportes a LA desde las técnicas de análisis del ASI, específicamente en la Sección 3.7 Técnicas de análisis.
- El segundo objetivo específico fue “Seleccionar las técnicas de análisis de datos del Análisis Estadístico Implicativo aplicables en *Learning Analytics*”, la selección de técnicas de análisis ASI aplicables en *Learning Analytics* se estudió en el Capítulo 6.- Aporte con Técnicas ASI a LA, su aproximación se realizó en el Capítulo 4.- Aproximación a los Elementos Comunes de LA y ASI y se utilizan en la Sección 10.3 Aportes a LA desde las técnicas de análisis del ASI.
- El tercer objetivo específico fue “Realizar un análisis comparativo entre las técnicas de análisis de datos utilizadas en el Análisis Estadístico Implicativo con sus similares en *Learning Analytics”*, el análisis comparativo detallado de la complejidad temporal y complejidad espacial se realizaron en el Capítulo 7 y en el Capítulo 8 y se resume sus resultados en el Capítulo 10 específicamente en las Sección 10.4 Aportes a LA desde las técnicas de análisis del ASI.
- El cuarto objetivo específico fue “Determinar las ventajas de aplicar las técnicas del Análisis Estadístico Implicativo en el marco de *Learning Analytics*”,
    - Primera ventaja, en aplicar las técnicas clúster ASI en LA es desde la complejidad algorítmica espacial (memoria), las técnicas clúster del ASI (Simlrty y hrarchy) tienen ventaja frente a la técnica clúster de LA (dendro_variables) y comparten la mínima ocupación de memoria que las otras técnicas (dendro_diana y hclust_vector) de LA, Capítulo 7.- Complejidad algorítmica entre técnicas clúster de LA y ASI, Sección 7.5 Conclusiones.
    - Segunda ventaja, en aplicar las técnicas clúster ASI en LA es desde la complejidad algorítmica temporal (tiempo), las técnicas clúster del ASI (Simlrty y hrarchy) tienen ventaja frente a la técnica clúster de LA (dendro_variables) de LA, Capítulo 7.- Complejidad algorítmica entre técnicas clúster de LA y ASI, Sección 7.5 Conclusiones.
    - Tercera ventaja, finalmente las técnicas clúster del ASI tienen ventajas en diez opciones adicionales como: tipo de datos amplio, variables

suplementarias, nodos significativos, entropía y conjuntos grandes de datos, tipicalidad, contribución, escenarios de análisis y experimentación, visualizaciones sencillas de interpretar, automatización y acceso libre a sus herramientas, técnicas clúster y de reglas de asociación incluidas en un mismo paquete, Capítulo 9.- Aportes factibles desde las opciones adicionales del ASI, Sección 9.13 Conclusiones, Tabla 9.1 y Tabla 9.2.

- Cuarta ventaja, las técnicas de reglas de asociación del ASI tienen ventajas en seis opciones adicionales como: tipo de datos amplio, entropía y conjuntos grandes de datos, escenarios de análisis y experimentación, visualizaciones sencillas de interpretar, automatización y acceso libre a sus herramientas, reglas de asociación y técnicas clúster incluidas en un mismo paquete, Capítulo 9.- Aportes factibles desde las opciones adicionales del ASI, Sección 9.13 Conclusiones, Tabla 9.1 y Tabla 9.2.

10.1.2 Sobre el objetivo general

El objetivo general planteado fue "Caracterizar el aporte de las técnicas del Análisis Estadístico Implicativo en las Analíticas de Aprendizaje". El aporte de las técnicas del ASI y LA se caracterizaron en las tres Secciones 10.2, 10.3 y 10.4 de las conclusiones:

- En la Sección 10.2 Aportes del ASI a LA desde la definición de LA, parte de la definición de LA dada en LAK 2011, los ingresos, los procesos y las salidas de LA, se analizó detalladamente en el Capítulo 5.- Aportes del ASI a LA.
- En la Sección 10.3 Aportes a LA desde las técnicas de análisis del ASI, se basa en las clasificaciones para técnicas de análisis propuestas por Baker e Inventado (Baker y Inventado, 2014) y Papamitsiou y Economides (Papamitsiou y Economides, 2014) y se analizó detalladamente en el Capítulo 6.- Aporte con Técnicas ASI a LA.
- En la Sección 10.4 Aportes desde la complejidad algorítmica de LA y ASI, es resultado de la comparación desde la complejidad algorítmica entre las técnicas de análisis comunes entre LA y ASI, como son técnicas clúster y técnicas de reglas de asociación y se analizó detalladamente en el Capítulo 7.- Complejidad algorítmica entre técnicas Clúster de LA y ASI y Capítulo 8.- Complejidad algorítmica entre reglas de asociación de LA y ASI.

10.1.3 Sobre las preguntas de investigación

Pregunta de Investigación 1: ¿El árbol de similaridad es una técnica dentro del Análisis Estadístico Implicativo que se puede utilizar en *Learning Analytics*?

Se observó que el árbol de similaridad se ha utilizado en LA, con un porcentaje de 37,5% en los años 2011-2016 y con un porcentaje de 61,9% en los años 2016-2021, dando un promedio de 49,70% en los 11 años de aplicación de la revisión sistemática de literatura desde la clasificación de Baker e Inventado (Baker y Inventado, 2014) o la clasificación de Papamitsiou y Economides (Papamitsiou y Economides, 2014), Capítulo 5.- Aportes del ASI a LA.

Pregunta de Investigación 2: ¿Cuáles son las ventajas del árbol de similaridad frente a otras técnicas de análisis similares utilizadas en *Learning Analytics*?

- Primera ventaja, la utilización del árbol de similaridad está dada por el uso de 10 opciones adicionales definidas en el Capítulo 9.- tipo de datos amplio, variables suplementarias, nodos significativos, entropía y conjuntos grandes de datos, tipicalidad, contribución, escenarios de análisis y experimentación, visualizaciones sencillas de interpretar, automatización y acceso libre a sus herramientas, técnicas clúster y de reglas de asociación incluidas en un mismo paquete.
- Segunda ventaja, en aplicar la técnica clúster (árbol de similaridad) del ASI en LA es desde la complejidad algorítmica espacial (memoria), tiene ventaja frente a la técnica clúster de LA (dendro_variables) y comparten la mínima ocupación de memoria que las otras técnicas (dendro_diana y hclust_vector) de LA, Capítulo 7.- Complejidad algorítmica entre técnicas clúster de LA y ASI, Sección 7.5 Conclusiones.
- Tercera ventaja, en aplicar las técnicas clúster (árbol de similaridad) ASI en LA es desde la complejidad algorítmica temporal (tiempo), tienen ventaja frente a la técnica clúster de LA (dendro_variables) de LA, Capítulo 7.- Complejidad algorítmica entre técnicas clúster de LA y ASI, Sección 7.5 Conclusiones.

Pregunta de Investigación 3: ¿El grafo implicativo es una técnica dentro del Análisis Estadístico Implicativo que se puede utilizar en *Learning Analytics*?

Se observó que el grafo implicativo se ha utilizado en LA, con un porcentaje de 95,8% en los años 2011-2016 y con un porcentaje de 66,7% en los años 2016-2021, dando un promedio de 81,25% en los 11 años de la aplicación de la revisión sistemática de literatura desde la clasificación de Baker e Inventado (Baker y Inventado, 2014) o la clasificación de Papamitsiou y Economides (Papamitsiou y Economides, 2014), Capítulo 5.- Aportes del ASI a LA.

Pregunta de Investigación 4: ¿Cuáles son sus ventajas del grafo implicativo frente a otras técnicas de análisis similares utilizadas en *Learning Analytics*?

- Las ventajas en la utilización del grafo implicativo están dadas por el uso de las opciones adicionales definidas en el Capítulo 9.- Aportes factibles desde las opciones adicionales del ASI: uso tipo de datos amplio, entropía y conjuntos grandes de datos, escenarios de análisis y experimentación, visualizaciones sencillas de interpretar, automatización y acceso libre a sus herramientas, técnicas clúster y de reglas de asociación incluidas en un mismo paquete.
- La técnica de reglas de asociación del ASI no tiene ninguna ventaja desde la complejidad algorítmica frente a las otras técnicas, pero esto no significa que el orden de complejidad de las cuatro técnicas sea diferente (met_apriori, met_eclat, met_weclat y met_ASI), es decir pueden estar en el mismo orden de complejidad algorítmica, Capítulo 8.- Complejidad Algorítmica entre Reglas de Asociación de LA y ASI.

Pregunta de Investigación 5: ¿El árbol cohesitivo es una técnica dentro del Análisis Estadístico Implicativo que se puede utilizar en *Learning Analytics*?

Se observó que el árbol cohesivo se ha utilizado en LA, con un porcentaje de 37,5% en los años 2011-2016 y con un porcentaje de 61,9% en los años 2016-2021, dando un promedio de 49,70% en los 11 años de la aplicación de la revisión sistemática de literatura desde la clasificación de Baker e Inventado (Baker y Inventado, 2014) o la clasificación de Papamitsiou y Economides (Papamitsiou y Economides, 2014), Capítulo 5.- Aportes del ASI a LA.

Pregunta de Investigación 6: ¿Cuáles son las ventajas del árbol cohesivo frente a técnicas de análisis similares utilizadas en *Learning Analytics*?

- Primera ventaja, en la utilización del árbol cohesivo está dada por el uso de 10 opciones adicionales definidas en el Capítulo 9.- Tipo de datos amplio, variables suplementarias, nodos significativos, entropía y conjuntos grandes de datos, tipicalidad, contribución, escenarios de análisis y experimentación, visualizaciones sencillas de interpretar, automatización y acceso libre a sus herramientas, técnicas clúster y de reglas de asociación incluidas en un mismo paquete.
- Segunda ventaja, en aplicar la técnica clúster (árbol cohesivo) del ASI en LA es desde la complejidad algorítmica espacial (memoria), tiene ventaja frente a la técnica clúster de LA (dendro_variables) y comparten la mínima ocupación de memoria que las otras técnicas (dendro_diana y hclust_vector) de LA, ver Capítulo 7.- Complejidad algorítmica entre técnicas clúster de LA y ASI, Sección 7.5 Conclusiones.
- Tercera ventaja, en aplicar las técnicas clúster (árbol cohesivo) ASI en LA es desde la complejidad algorítmica temporal (tiempo), tiene ventaja frente a la técnica clúster de LA (dendro_variables) de LA, Capítulo 7.- Complejidad algorítmica entre técnicas clúster de LA y ASI, Sección 7.5 Conclusiones.

10.1.4 Sobre la hipótesis

La hipótesis del trabajo de esta tesis es: El Análisis Estadístico Implicativo aporta a las Analíticas de Aprendizaje.

La respuesta es afirmativa, es decir se cumple la hipótesis para los 521 artículos científicos considerados en las 2 revisiones sistemáticas de literatura realizadas en los últimos 11 años de estudio 2011-2021. No solo se ha determinado la existencia de los aportes (Capítulo 4.- Aproximación a los elementos comunes de LA y ASI), se los ha categorizado (Secciones 10.2, 10.3 y 10.4) y también se los ha cuantificado (Capítulos 5, 6, 7 y 8.

10.1.5 Sobre el problema

¿Existen elementos comunes entre el ASI y LA, se puede determinar el aporte del ASI a LA?, el problema lo dividimos en dos partes para presentar de mejor forma su solución:

- ¿Existen elementos comunes entre el ASI y LA? se demostró utilizando las revisiones sistemáticas de literatura y la teoría de conjuntos que el ASI y LA al

menos son comunes en el campo educativo y en tres métodos de análisis: minería de relaciones, descubrimiento de estructura y estadísticas, según las clasificaciones de Baker e Inventado y Papamitsiou y Economides.

- ¿Se puede determinar el aporte del ASI a LA? la respuesta es sí, primeramente, desde los aportes del ASI a LA basados en la definición de LA dada en LAK, las fuentes de datos de LA y las etapas de LA definidas por Campbell (Capítulo 5.- Aportes del ASI a LA), luego los aportes desde la técnicas clúster y de reglas de asociación de ASI (Capítulo 6.- Aporte con técnicas ASI a LA) y finalmente los aportes desde la complejidad algorítmica de LA y ASI (Capítulo 7.- Complejidad algorítmica entre técnicas clúster de LA y ASI y Capítulo 8.- Complejidad algorítmica entre reglas de asociación de LA y ASI).
- Además, en el Capítulo 9.- Aportes factibles desde las opciones adicionales del ASI, se describen detalladamente las opciones de las técnicas ASI con las que ha estado aportando y serían factibles de continuar aportando en LA.

Para aquellas personas e investigadores que accedan a este trabajo de tesis y lo lean con interés, se habrá resuelto el problema, pues se cambiaría la situación actual del objeto que es el desconocimiento de los aportes del ASI a LA y se habrá convertido en la situación deseable que es el conocimiento de los aportes del ASI a LA.

10.1.6 Limitaciones y problemas

- El tiempo utilizado fue muy extenso debido a que se necesitó primeramente la programación de la versión libre de Rchic para lograr la validez interna, luego la adecuación de los tres ambientes hardware, la programación y ejecución en los tres ambientes de software y finalmente su análisis.
- Tanto el Plan de Investigación como el desarrollo de la tesis tienen 4 objetivos específicos. En el Plan de Investigación se tiene el tercer objetivo específico como: Aplicar las técnicas de análisis de datos del Análisis Estadístico Implicativo en las Analíticas de Aprendizaje, éste se cumple en el Capítulo 9.- Aportes factibles desde las opciones adicionales del ASI, Sección 9.12 caso de estudio. En la tesis se desarrolló además un objetivo más que es determinar las ventajas de aplicar las técnicas del Análisis Estadístico Implicativo en el marco de *Learning Analytics*.

- En el Plan de Investigación las preguntas de investigación se refieren a las características de la cuasi-implicación, cohesión y reducción, las que se encuentran en el Capítulo 3.- El Análisis Estadístico Implicativo (ASI), Sección 3.7 Técnicas de análisis y en el Capítulo 9.- Aportes factibles desde las opciones adicionales del ASI.
- No se midió la complejidad computacional de la categoría estadística (Papamitsiou y Economides, 2014), la cual se refiere a valores numéricos tales como la media, desviación estándar, coeficiente de correlación de Pearson, etc., cuya complejidad algorítmica al máximo es de n y se puede reducir en la complejidad espacial a un orden constante.
- Por el gran tamaño de la población (cálculo en el Apéndice C), fue necesario tomar una muestra.
- Algunos de los paquetes analizados ya no se actualizan a las nuevas versiones de R, por ejemplo, la librería *CluMix*.

10.2 Aportes del ASI a LA desde la definición de LA

Los aportes del ASI a LA se basan en la definición de LA dada en (*LAK 2011 : 1st International Conference Learning Analytics and Knowledge*, 2011), las fuentes de datos de LA (aprendizaje de los estudiantes, demografía, percepción y procesos de la institución educativa) y las etapas de LA definidas por Campbell (captura, información, predicción, actuación y refinamiento).

- El ASI aporta a la definición de LA en forma global en aproximadamente un 60% de los artículos científicos del ASI analizados. El ASI estudia sobre todo a los estudiantes (91,7% y 71,4%), pero en los últimos 5 años se observa un crecimiento en el estudio del contexto de aprendizaje (8,3% y 66,6%). El ASI últimamente contribuye significativamente a LA en el análisis e informes del aprendizaje y su contexto (20,8% y 47,6%), por su facilidad de interpretación mediante árboles de similaridad, árboles de cohesión y gráfico implicativo. El ASI ayudaba más en la comprensión (83,3% y 61,9%), pero actualmente aumentan los artículos científicos que optimizan los aprendizajes (16,7% y 19,0%).

- Sobre el aporte del ASI a las fuentes de datos de LA, como resultado se tiene que el ASI continúa contribuyendo y creciendo en los datos del aprendizaje (4,2% y 19,0%), pero también ha crecido su contribución en los datos de los procesos de la institución educativa, la demografía y la percepción (12,5% y 23,8%).
- En cuanto el aporte del ASI a las etapas en LA definidas por Campbell, en la revisión sistemática del 2016 se observa que el ASI contribuye en la captura (90,4%), información (57,1%) y predicción (9,5%), pero contribuye poco en la etapa de actuación (4,2%) y hasta el año 2016 no se detectó la contribución en la etapa de refinamiento (0%), pero en la revisión sistemática del 2021 se detectó que el ASI aporta al LA en el actuar (33,3%), con un incremento en refinar (19%).

10.3 Aportes a LA desde las técnicas de análisis del ASI

A partir de la clasificación de las técnicas de análisis utilizadas en LA de Baker e Inventado y la clasificación de Papamitsiou, se determinó aquellas en las que más aporta el ASI (Tabla 10.1).

Tabla 10.1.- Comparativo sobre métodos de análisis en ASI que aportan en LA según Baker e Inventado y Papamitsiou de las revisiones sistemáticas del 2011 y del 2021 (Baker y Inventado, 2014; Papamitsiou y Economides, 2014; Pazmiño-Maji et al., 2016)

MÉTODOS DE ANÁLISIS EN LAS LA SEGÚN BAKER E INVENTADO Y PAPAMITSIOU		2011-jun2016	2016-jun2021	11 años
		FREQ_A[8]	FREQ_D	PROMEDIO
1. Minería de relaciones	Minería de reglas de asociación	95,8%	66,7%	81,25%
	Minería de datos causales	4,2%	4,7%	4,45%
2. Descubrimiento de estructura	Clúster	37,5%	61,9%	49,7%
3. Estadísticas		20,8%	33,3%	27,05%
Descubrimiento con modelos, Minería de texto y Visualización *				

El ASI aporta con las siguientes técnicas a LA:

1. Minería de relaciones (85,7%), que en caso del ASI están dadas por los análisis de cuasi-implicación.

[8] _A representa los resultados de la revisión sistemática de literatura entre los años 2011 hasta junio del 2016.
_D representa los resultados de la revisión sistemática de literatura entre los años 2016 hasta junio del 2021.

2. El método clúster (49,7%), que en caso del ASI están dadas por el árbol similaridad de Israel Lerman y podría aportar también por su forma de estructurar los datos el árbol de cohesión.

3. Las estadísticas (27,05%) que en caso del ASI están dadas básicamente por la frecuencia, la media, la desviación estándar, la frecuencia de parejas de variables, el coeficiente de correlación de Pearson y los índices de similaridad, cohesión e implicación.

Los aportes anteriores son resultado de las revisiones de literatura del año 2016 y la más reciente del año 2021, en total se han considerado los últimos 11 años (desde el año 2011 hasta el año 2021).

En la Tabla 10.1 en la última fila se dejan indicadas tres técnicas de LA, de las cuales se tienen pocas evidencias de su aporte mediante el ASI, pero vale la pena tomarlas en cuenta debido a que pueden ser tendencia en los próximos años, de la técnica descubrimiento con modelos se evidencia su aporte con el artículo científico (Sagaró del Campo y Zamora Matamoros, 2019), de la técnica minería de texto se evidencia su aporte con el artículo científico (David et al., 2008) y de la técnica de visualización: los árboles de similaridad, cohesión y el gráfico implicativo podrían aportar debido a su facilidad de utilización e interpretación.

Es importante notar que el ASI tiene una técnica llamada reducción que permite formar grupos y determinar los representantes de grupo que podría ser una alternativa al clúster no jerárquico, pero debido a que no se encuentra implementado en Rchic (pero sí en el software CHIC) no ha sido posible profundizar en su estudio.

10.4 Aportes desde la complejidad algorítmica de LA y ASI

Se presentan los resultados del análisis comparativo entre las técnicas clúster de LA y ASI y las técnicas de asociación de LA y ASI. El análisis comparativo se realizó desde el punto de vista de la complejidad espacial y complejidad temporal, es decir desde el espacio de memoria ocupado por los algoritmos al momento de su ejecución y el tiempo utilizado por los algoritmos para cumplir la tarea asignada. La complejidad algorítmica se compara en forma experimental y los resultados se basan en un pre-experimento previamente diseñado. Primeramente, se realiza el análisis estadístico comparativo de las técnicas

clúster analizadas en el Capítulo 7.- Complejidad algorítmica entre técnicas clúster de LA y ASI y luego el análisis estadístico comparativo de las técnicas de reglas de asociación analizadas en el Capítulo 8.- Complejidad algorítmica entre reglas de asociación de LA y ASI.

10.4.1 Complejidad algorítmica entre técnicas clúster de LA y ASI

Se compararon la complejidad espacial y temporal de los métodos clúster del ASI:

- Hrarchy, correspondiente a la técnica cohesiva con la función callHierarchyTree perteneciente al paquete Rchic.
- Simlrty, correspondiente a la técnica de similaridad con la función callSimilarityTree perteneciente al paquete Rchic.

y las técnicas clúster de LA:

- hclust_vector, correspondiente a la función hclust.vector perteneciente al paquete fastcluster.
- dendro_variables, correspondiente a la función dendro.variables perteneciente al paquete cluMix.
- dendro_diana, correspondiente a la función diana, perteneciente al paquete clúster.

El estudio descriptivo en cuanto a la complejidad espacial, indica que la técnica que usa en promedio menor memoria es hclust_vector con un valor igual a 223,0, con un mínimo de memoria igual a 103,0 y un máximo de 381,0, con una mediana de 247,0. Respecto a la complejidad temporal, se tuvo como resultado que la técnica que usa en promedio menor tiempo es hclust_vector con un valor promedio de 0,242727, con un mínimo de memoria igual a 0,007087 y un máximo de 24,250590, con una mediana de 0,050403.

Respecto a los supuestos que permiten seleccionar una prueba de hipótesis paramétrica o no paramétrica se obtuvo:

- Sobre la normalidad, tanto los datos de memoria como de tiempo no son normales para ninguna de las pruebas utilizadas a ninguno de los niveles de significancia α=0,01, α=0,05 y tampoco α=0,1, y tampoco son posibles de normalizar, ya que los p-values obtenidos son bastante pequeños (< 2,22e-16 para ambos métodos).

- Sobre la igualdad de varianzas, se comprobó que no existía homocedasticidad en los datos de memoria, pero sí en los datos de tiempo.
- Las técnicas clúster y la memoria no son independientes
- Las técnicas clúster y el tiempo no son independientes.

Al no cumplirse todos los supuestos de normalidad, homocedasticidad e independencia se optó por una prueba no paramétrica.

- Las pruebas de hipótesis no paramétricas aplicadas en la comparación de la complejidad espacial (memoria) entre las 5 técnicas clúster indican que a un nivel de significancia de α=0,05 y utilizando las pruebas de Kruskal Wallis (p-value < 2,2e-16) y ANOVA no paramétrico (p-value=0,000) se rechaza la hipótesis nula con alta significancia, por lo tanto, existe al menos un par de poblaciones clúster diferentes respecto a la memoria. Aplicando las pruebas posteriores de Mann Whitney Wilcoxon y Tukey, se obtuvieron 2 grupos de homogeneidad, el primero con menor ocupación de memoria conformado por dendro_diana, hclust_vector, hrarchy y simlrty y el segundo formado solo por la técnica dendro.variables.
- La comparación de la complejidad temporal (tiempo) entre las 5 técnicas clúster indican que a un nivel de significancia de α=0,05 y utilizando las pruebas de Kruskal Wallis (p-value < 2,2e-16) y ANOVA no paramétrico (p-value=0) se rechaza la hipótesis nula con alta significancia, por lo tanto, existe al menos un par de poblaciones clúster diferentes respecto al tiempo. Aplicando las pruebas posteriores de Mann Whitney Wilcoxon y Tukey, se obtuvieron 5 grupos de homogeneidad. Cada técnica clúster es diferente, la que ocupa menor tiempo es hclust_vector y la que ocupa mayor tiempo es dendro_variables.

La Tabla 10.2 resume los resultados tanto de complejidad temporal, así como complejidad espacial.

Tabla 10.2 Resultados de complejidad algorítmica para técnicas clúster

ORDEN	TÉCNICA CLÚSTER
1	hclust_vector
2	dendro_diana
3	hrarchy
4	Simlrty
5	dendro_variables

Es importante notar que el hecho de que las cinco técnicas clúster (dendro_diana, dendro_variables, hclust_vector, hrarchy y simlrty) se encuentren en grupos de homogeneidad diferentes no significa que el orden de complejidad de las cinco técnicas sea diferente, es decir pueden estar en el mismo orden de complejidad algorítmica.

10.4.2 Complejidad algorítmica entre reglas de asociación de LA y ASI

Se compararon la complejidad espacial y temporal de las técnicas de reglas de asociación del ASI (met_ASI) correspondiente a la técnica implicativa con la función implicativeGraph perteneciente al paquete Rchic y las técnicas de reglas de asociación de LA:

- met_apriori, correspondiente a la función met_apriori perteneciente al paquete arules.
- met_eclat, correspondiente a la función met_eclat perteneciente al paquete arules.
- met_weclat, correspondiente a la función met_weclat perteneciente al paquete arules.

El estudio descriptivo en cuanto a la complejidad espacial, indica que la técnica que usa menor memoria es met_weclat con un valor igual a 156,9, con un mínimo de memoria igual a 136,9 y un máximo de 238,2, con una mediana de 156,6. En cuanto a la complejidad temporal la técnica que usa en promedio menor tiempo es met_eclat con un valor igual a 0,98292, con un mínimo de tiempo igual a 0,03404 y un máximo de 115,99705, con una mediana de 0,20032.

Respecto a los supuestos que permiten seleccionar una prueba paramétrica o no paramétrica se obtuvo que:

- Sobre la normalidad, ni los datos de espacio ni de tiempo son estadísticamente normales por ninguna de las pruebas ni de los niveles del sesgo α=0,01, α=0,05 y tampoco α=0,1, tampoco se pudo transformar a normales aplicando los métodos de transformación.
- Se comprobó que existía heterocedasticidad tanto para los datos de memoria, así como para los datos de tiempo.
- Sobre la independencia, las técnicas de reglas de asociación y la memoria no son independientes al igual que las técnicas de reglas de asociación y el tiempo.

Al no cumplirse todos los supuestos de normalidad, homocedasticidad e independencia se optó por utilizar pruebas no paramétricas para las pruebas de hipótesis.

- Las pruebas no paramétricas seleccionadas indican que a un nivel de significancia de α=0,05 y utilizando la prueba de Kruskal Wallis (p-value=6,391e-11) y ANOVA no paramétrico (p-value=6,8834e-14) se rechaza la hipótesis nula con alta significancia, es decir existe al menos un par de técnicas de reglas de asociación que son diferentes en cuanto a la memoria. Para determinar los pares diferentes se utilizó las pruebas posteriores de Mann Whitney Wilcoxon y Tukey, se obtuvieron dos grupos de homogeneidad, el grupo de menor ocupación de memoria formado por las tres técnicas de reglas de asociación (met_apriori, met_eclat y met_weclat) y el otro grupo formado por el método met_ASI.
- Las pruebas de hipótesis aplicadas para la comparación de la complejidad temporal (tiempo) entre las 4 técnicas de asociación indican que a un nivel de significancia de α=0,05 y utilizando las pruebas de Kruskal Wallis (p-value < 2,2e-16) y ANOVA no paramétrico (p-value=0), se rechaza la hipótesis nula con alta significancia, por lo tanto, existe al menos un par de poblaciones de reglas de asociación diferentes respecto al tiempo. Aplicando las pruebas posteriores de Mann Whitney Wilcoxon, y Tukey, se obtuvieron 4 grupos de homogeneidad, es decir cada método está en un grupo de homogeneidad diferente, el más pequeño es met_apriori y el más grande es met_ASI.

La Tabla 10.3 resume los resultados tanto de complejidad temporal, así como complejidad espacial.

Tabla 10.3 Resultados de complejidad algorítmica para técnicas de reglas de asociación

ORDEN	TÉCNICA DE REGLAS DE ASOCIACION
1	met_eclat
2	met_apriori
3	met_weclat
4	met_ASI

Es importante resaltar que el hecho de que las cuatro técnicas clúster se encuentren en grupos de homogeneidad diferentes no significa que el orden de complejidad de las cuatro técnicas (met_apriori, met_eclat, met_weclat y met_ASI) sea diferente, es decir pueden estar en el mismo orden de complejidad algorítmica.

10.5 Resultados asociados

Los resultados asociados a esta investigación son: La gestión por parte del doctorando para que el profesor Rafael Couturier participe por dos años como investigador Prometeo en la ESPOCH (*Inicio - Escuela Superior Politécnica de Chimborazo*, 2016) y pueda elaborar el programa Rchic, sin el cual no se podrían elaborar trabajos similares al de esta tesis. Crear en el año 2017 (RESOLUCION ADMINISTRATIVA 128.2017/RESOLUCION CONSEJO POLITECNICO 301.CP.2018) y ser el coordinador del grupo de investigación multidisciplinario Ciencia de Datos – CIDED (*Cided*, 2019) cuyo propósito es el de motivar, gestionar y realizar investigación en Ciencia de Datos y sus aplicaciones. Actualmente se han promovido 4 tesis de postgrado (1 finalizada y 3 en elaboración) y 2 de pregrado (en elaboración). Se realiza investigación mediante 1 proyecto de investigación financiado con $81107,2 (ya finalizado), 1 proyecto de investigación actual financiado con $30000 y un proyecto aprobado para iniciarse en el año 2022 con un presupuesto de $75000. La existencia del grupo de investigación y la gestión para la adquisición de equipos de cómputo de alto rendimiento posibilitan a la ESPOCH el iniciar con el estudio para la implementación de LA y ASI. Se ha elaborado el proyecto de una Maestría en Ciencia de Datos y Estadística para promover las aplicaciones en esta área, está planificada para ejecutarse en el año 2022. En cuanto a publicaciones se tiene:

Tesis finalizada y artículo publicado:

- NARANJO SERRANO, M. M., Y PAZMIÑO MAJI, R. A. (2018). ESTUDIO COMPARATIVO DEL ANÁLISIS ESTADÍSTICO IMPLICATIVO Y EL LEARNING ANALYTICS EN RELACIÓN AL USO DE LAS TÉCNICAS DE EXPLORACIÓN DE DATOS EDUCATIVOS (Naranjo Serrano y Pazmiño Maji, 2018b).
- NARANJO, M., PAZMIÑO-MAJI, R., CONDE, M., Y PEÑALVO, F. (2018). LA&SIA CLUSTER METHODS: COMPUTATIONAL COMPARISON (Naranjo et al., 2018).

Como publicaciones asociadas se tienen:

- Pazmiño-Maji, R., Conde, M. Á., y García-Peñalvo, F. (2021). Learning analytics in Ecuador: A systematic review supported by statistical implicative analysis. Universal Access in the Information Society, 1-18 (R. Pazmiño-Maji et al., 2021).
- Pazmiño-Maji, R. A., Conde González, M. Á., y García Peñalvo, F. J. (2019). Learning Analytics in Ecuador: Analysis based in a Mapping Review. Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality (R. A. Pazmiño-Maji, Conde González, et al., 2019).
- Pazmiño-Maji, R., Conde González, M. Á., y García Peñalvo, F. J. (2019). Las analíticas de aprendizaje en el Ecuador: Un análisis inicial basado en el mapeo sistemático de los trabajos de graduación. Explorador Digital, 3(3.1), 224-245. https://doi.org/10.33262/exploradordigital.v3i3.1.885 (R. A. Pazmiño Maji et al., 2019).
- Pazmiño-Maji, R., García-Peñalvo, F., Conde-González, M., y Solis Benavides, C. (2019). La Investigación De Pregrado En La Escuela Superior Politécnica De Chimborazo: Mapeo Sistemático Y Analíticas (R. Pazmiño-Maji, García-Peñalvo, et al., 2019).
- Pazmiño-Maji, R., Bonilla, M., Baquero, J., y Miguez, R. (2018). Software estadístico chic: Descubriendo sus potencialidades mediante el análisis de percepción sexual universitaria. Ciencia digital, 2, 17 (R. Pazmiño-Maji et al., 2018).
- Barragán-Pazmiño, B. M., y Pazmiño-Maji, R. (2018). Literatura Científica Sobre Análisis Estadístico Implicativo: Un Mapeo Sistemático De La Década Que Transcurre. Ciencia Digital, 2, 16 (Barragán-Pazmiño y Pazmiño-Maji, 2018).
- Pazmiño-Maji, R. A., García-Peñalvo, F. J., y Conde-González, M. A. (2017b). Comparing Hierarchical Trees in Statistical Implicative Analysis & Hierarchical Cluster in Learning Analytics. 1-7 (R. Pazmiño-Maji et al., 2017b).
- Pazmiño-Maji, R. A., García-Peñalvo, F. J., y Conde-González, M. A. (2017c). Statistical Implicative Analysis Approximation to KDD and Data Mining: A

SYSTEMATIC AND MAPPING REVIEW IN KNOWLEDGE DISCOVERY DATABASE FRAMEWORK. DBKDA 2017, 79 (R. Pazmiño-Maji et al., 2017c).

- PAZMIÑO MAJI, R., GARCÍA PEÑALVO, F. J., Y CONDE GONZÁLEZ, M. Á. (2017). IS IT POSSIBLE TO APPLY STATISTICAL IMPLICATIVE ANALYSIS IN HIERARCHICAL CLUSTER ANALYSIS? FIRSTS ISSUES AND ANSWERS (R. Pazmiño Maji et al., 2017).
- PAZMIÑO-MAJI, R., GARCÍA-PEÑALVO, F. J., Y CONDE-GONZÁLEZ, M. A. (2017A). ASSOCIATION RULES WITH SIA IN B-LEARNING COURSES: A MAPPING REVIEW (R. Pazmiño-Maji et al., 2017a).
- PAZMIÑO-MAJI, R., GARCÍA-PEÑALVO, F. J., Y CONDE-GONZÁLEZ, M. A. (2016). APPROXIMATION OF STATISTICAL IMPLICATIVE ANALYSIS TO LEARNING ANALYTICS: A SYSTEMATIC REVIEW. PROCEEDINGS OF THE FOURTH INTERNATIONAL CONFERENCE ON TECHNOLOGICAL ECOSYSTEMS FOR ENHANCING MULTICULTURALITY, 355-376 (Pazmiño-Maji et al., 2016).
- COUTURIER, R., PAZMIÑO-MAJI, R., GARCÍA-PEÑALVO, F., Y CONDE-GONZÁLEZ, M. (2015). STATISTICAL IMPLICATIVE ANALYSIS FOR EDUCATIONAL DATA SETS: 2 ANALYSIS WITH RCHIC (Coutrier et al., 2015)

Lo presentado, es el resultado de lo aprendido en el Programa de Doctorado Formación en la Sociedad del Conocimiento y consecuencia del acompañamiento y ayuda del Dr. D. Francisco Peñalvo y del Dr. D. Miguel Conde, para ellos mi eterno agradecimiento.

10.6 Futuras investigaciones

En esta sección se destacan las futuras investigaciones y los nuevos temas de investigación originados en esta tesis.

Sobre la aproximación y Aportes del ASI a LA

- En los estudios realizados, el ASI no contribuye en la definición de LA en la parte del contexto de aprendizaje. ¿Cuáles son las razones? ¿Cuáles son las estrategias para contribuir en la parte del contexto de aprendizaje?
- ¿Por qué el ASI contribuye poco en la optimización del aprendizaje desde la definición de LA? ¿Cómo incrementar la contribución?
- ¿Por qué el ASI no aporta en la definición de LA en demografía y percepción? ¿Cómo incrementar la contribución?

- ¿Cómo incrementar la contribución del ASI con el paso de actuar?
- ¿El ASI puede contribuir en el paso refinar de las etapas de las LA? ¿Cómo incrementar esta contribución?
- ¿Existen otros elementos comunes en ASI y LA?

De los resultados obtenidos, el ASI puede contribuir en mayor forma en las reglas de asociación, pero surgen las siguientes preguntas:

- ¿Cuál es la veracidad de las reglas de asociación utilizando el ASI?
- ¿Existen diferencias en la veracidad entre las reglas de asociación en el ASI y el uso de las técnicas tradicionales en las LA?

La segunda mayor forma en que el ASI puede contribuir a LA es en las técnicas clúster, pero también surge la siguiente pregunta:

- ¿Existen diferencias en la veracidad entre el clúster jerárquico en el ASI y el uso de las técnicas tradicionales en las LA?

Sobre la complejidad algorítmica

- ¿Es posible obtener una función de orden de complejidad espacio (tiempo) que permita proyectar su valor a partir de la dimensión de la base de datos de ingreso?
- ¿Es posible aproximarnos a los órdenes de complejidad tanto en espacio como en tiempo en las técnicas analizadas?

Nos permitimos además realizar las siguientes recomendaciones:

- Probar las diferentes opciones de Rchic que tienen que ver con la opción modo de computación en Rchic (computing.mode = 1, 2, 3).
- Determinar el impacto de utilización de las herramientas de ASI en LA.
- Actualizar las investigaciones realizadas a las nuevas funciones clúster y de reglas de asociación y al hardware y software de uso más frecuente en LA.
- Realizar investigaciones sobre el comportamiento de la complejidad algorítmica en el caso de datos de tipo cualitativo.
- Realizar investigaciones sobre el comportamiento de la complejidad algorítmica en el caso de datos de tipo intervalo.
- Actualizar la revisión sistemática del ASI realizada por primera vez en el año 2018.

- Demostrar que las opciones adicionales del ASI, contribuyen a atenuar las dificultades que podría presentar LA.
- Programar CHIC en Python (PyCHIC), para utilizarlo de mejor manera con GPU y CPU y que guarde mayor compatibilidad con el software científico.
- Caracterizar las necesidades de LA que pueden ser resueltas por el ASI.
- Adecuar la teoría ASI para poder aportar aún más en LA
- Adecuar los programas informáticos del ASI para poder aportar aún más en LA

Finalmente, deseo expresar mi deseo de que este trabajo se constituya en un primer acercamiento entre el Análisis Estadístico Implicativo y *Learning Analytics* y sea el promotor de muchos trabajos colaborativos.

C a p í t u l o 11ᵛᵒ | APÉNDICES

Se adjuntan detalles de información adicional que mejora, expande y facilita la comprensión de la investigación realizada en esta tesis

11 Apéndices

Se muestra en forma detallada las referencias bibliográficas utilizadas en la revisión sistemática de literatura elaborada en el año 2016.

11.1 Apéndice A.- Aportes del ASI a LA: 2011 - junio 2016

Nº	ARTÍCULO CIENTÍFICO	REFERENCIA
[1]	Kindergartners' perspective taking abilities. In Proceedings of the Seventh Congress of the European Society for Research in Mathematics Education	(Aaten et al., 2016)
[2]	Implicative Statistical Analysis and PrincipalComponents Analysis in Recording Students' Attitudesto Electronics and Electrical Construction Subjects.	(Anastasiadou et al., 2011)
[3]	The Beliefs of Electrical and Computer Engineering Students' Regarding Computer Programming	(Anastasiadou y Karakos, 2011)
[9]	Drawings of the hand and numerical skills in children of preschool age. Canadian Journal of Behavioural Science-Revue Canadienne Des Sciences Du Comportement	(Bonneton-Botte et al., 2015)
[10]	eaching fractions through situations: A fundamental experiment	(Margolinas, 2014)
[14]	Use of *Statistical implicative analysis* in Complement of Item Analysis	(Couturier y Pazmiño-Maji, 2016)
[16]	An application of multiple behavior SIA for analyzing data from student exams	(Delacroix y Boubekki, 2014)
[17]	Students' mathematical work on absolute value: focusing on conceptions, errors and obstacles	(Elia et al., 2016)
[18]	Investigating the quality of mental models deployed by undergraduate engineering students in creating explanations: The case of thermally activated phenomena	(Fazio et al., 2013)
[19]	Quantitative and qualitative analysis of the mental models deployed by undergraduate students in explaining thermally activated phenomena	(Fazio et al., 2017)
[28]	Apport de la Combinaison de la Méthode d'Analyse Statistique Implicative (ASI) avec la Théorie de Réponse aux Items (IRT)	(Khaled y Couturier, 2015)
[31]	Analysis of University Entrance Test from mathematics	(Kohanova, 2012)
[32]	Flexible use and understanding of place value via traditional and digital tools	(Kortenkamp y Ladel, 2014)
[35]	Motivations et compétences interculturelles pour la mobilité académique France-Brésil : le cas des étudiants de l'Université Lumière Lyon 2	(Santos et al., 2014)
[36]	Upper-secondary students' strategies for solving combinatorial problems	(Melusova y Vidermanova, 2015)
[38]	Positions numeration in any base for future Elementary school teachers in France and Greece: one discusión via registers and praxis	(Nikolantonakis y Vivier, 2013)
[40]	Lifelong Learning Policy Agenda in the European Union: A bi-level analysis	(Panitsides y Anastasiadou, 2015)
[42]	The Attitudes of Students to the Geometry and Their Concepts about Square	(Pavlovicova y Zahorska, 2015)
[44]	*Statistical implicative analysis* for educational data sets: 2 analysis with Rchic	(Couturier et al., 2015)
[45]	Competences professionnelles et linguistiques de professionnels de santé dans l'espace frontalier UruguayenBresilienProfessional	(Pérez-Caraballo et al., 2014)

	and linguistic competence of health care providers in the Uruguay-Brazil borderegion: SIA'S contribution.	
[47]	Open-inquiry driven overcoming of epistemological difficulties in engineering undergraduates: A case study in the context of thermal science	(Pizzolato et al., 2014)
[54]	Kindergartners' Performance in Two Types of Imaginary Perspective-Taking	(van den Heuvel-Panhuizen et al., 2015)
[56]	Misconceptions in Pre-service Primary Education Teachers about Quadrilaterals	(Žilková, 2015)
[57]	(Mis)conceptions about geometric shapes in pre-service primary teachers	(Žilková, 2015)

11.2 Apéndice B.- Aportes del ASI a LA: 2016 - junio 2021

Se muestra en forma detallada las referencias bibliográficas utilizadas en la revisión sistemática de literatura elaborada en el año 2021.

Nº	ARTÍCULO CIENTÍFICO	REFERENCIA
[1]	A Novel Cohesitive Implicative Classification Based on MGK6 and Application on Diagnostic on Informatics Literacy of Students of Higher Education in Madagascar	(Rakotomalala et al., 2019)
[2]	Analysis of the feelings of the population's opinion in social media: a look at education	(Kwecko et al., 2020)
[3]	Approximation of *Statistical implicative analysis* to *Learning Analytics*: A systematic review	(Pazmiño-Maji et al., 2016)
[4]	Association rules with SIA in B-learning Courses: A mapping review	(R. Pazmiño-Maji et al., 2017a)
[5]	Association-based recommender system using statistical implicative cohesion measure	(Phan et al., 2016)
[6]	Comparison of multivariate pattering methods in group/cluster identification regarding the science of educational research: Implicative Statistical Analysis vs. Lâ€™ Analysee Factorielle des Correspondances	(Anastasiadou, 2019)
[7]	Fourth International Congress on Information and Communication Technology	(Yang et al., 2015)
[8]	Investigation of Selected Aspects of Fraction Understanding	(Pavlovičová y Vargová, 2020)
[9]	Is it possible to apply *Statistical implicative analysis* in hierarchical cluster Analysis? Firsts issues and answers	(R. Pazmiño Maji et al., 2017)
[10]	*Learning Analytics* in Ecuador: a systematic review supported by *Statistical implicative analysis*	(R. Pazmiño-Maji et al., 2021)
[11]	*Learning Analytics* in Ecuador: An Initial Analysis Based in a Mapping Review	(R. Pazmiño-Maji, Naranjo-Ordoñez, et al., 2019)
[12]	*Learning Analytics*: Expanding the Frontier	(Conde y Hernández-García, 2017)
[13]	*Learning Analytics*: Needs and Opportunities	(Hernández-García y Conde, 2016)
[14]	Mapping kindergartners' quantitative competence	(Van den Heuvel-Panhuizen y Elia, 2020)
[15]	Mathematical Content on STEM Activities	(Lasa et al., 2020)
[16]	New methods and technologies for enhancing usability and accessibility of educational data	(Fonseca et al., 2021)
[17]	On hierarchical classification implicative and cohesive mgk-based: Application on analysis of the computing curricula and students abilities according the anglo-saxon model	(Rakotomalala y Totohasina, 2020)
[18]	Probability distribution of the classical implication intensity seen as a random variable in *Statistical implicative analysis*	(Marín Martínez, 2017)
[19]	Student acceptance of online assessment with e-authentication in the UK	(Okada et al., 2017)
[20]	Synergies among students' thinking modes and representation types in linear algebra: employing *Statistical implicative analysis*	(Turgut, 2018)
[21]	Use of *Statistical implicative analysis* in Complement of Item Analysis	(Couturier y Pazmiño-Maji, 2016)

11.3 Apéndice C.- Tamaño de la población (colectivo de estudio)

El tamaño de la población considera primero el número de elementos y se tienen las 100000 bases de datos formadas hasta por un máximo de 1000 sujetos y 100 variables, dado que los elementos de las bases de datos son aleatorios binarios, el tamaño de la población se lograría multiplicando por $2(2^{100000}-1)$, es decir $200000(2^{100000}-1)$, considerando que hasta llegar a las 100 variables se pueden generar base de datos del mismo número de elementos pero transpuestas (6=6x1=1x6, menos la base de datos 1x1), la población final aumentará en un factor $2(2^{100}-1)$ y la población hasta allí será de $200000[(2^{100000}-1)+2(2^{100}-1)-1]$, debido a que no considera los casos por ejemplo 20=10x2=2x10=5x4=4x5, se obtiene una población mayor que $200000[(2^{100000}-1)+2(2^{100}-1)-1]$, es decir mayor que el siguiente número generado utilizando el software WOLFRAM MATHEMATICA (*Wolfram Mathematica*, 2015).

199800418602876901588806552866006718196085827810836338354305854772629166492851469665497466266489930080632878889111170986003759932153
123531258169427084949857503977792597473421864927008547462249585316005570624821774741712105744567803291373820537013518470358293941057
152893936030496646909510865005855730416139155419434822044640859527024106615559937958502332397415435715519110434401626405904092358984
585185912478419315957471163173350509915946262896124985205236758826101611653720630702683574792456699817727155161242092132727442611755
906446899440216169727390828036717027197160712071480437458163111331612143729225379456795892436845351586992777867144951775239182753135
248222500414174097409303587927974202184007278694912361812032267557971205937271971160495228978660941044457202627541919167146389717969
928091447677503414044846652668737888465947637554663065738884358722506038157378072073265663230054522798683056081423438298478066837498
707889117926025843945128354344670870895031047586217845363648049055115041894092843718877257312654884626641694844431029866300054355001
284576524236450986992011149146699293569665383619037919115383490193464488354808656809117642182758107512935442799532435705301143397096
691249750366447665006372910109442287398683359633563405102456259561303896125908106783093149598825949983806970150886728290112633147920
133867648546328680791602425605219684244950284156694244496628206081372074392803237114833128789445069298904994006290197801863245379054
888574109528509445063350290423644629107767486164652844012660502750662587303286834504125123106235894772382858095228913098542568363503
670626541059509907411228764791464558793460602121549136969548556643906984559676728723275294859390918133447382482726518642466712862717
889304382037642476594818158327720464709019187775334728064591559878023043088960072744301382311822239920030611782154588420644608485240
713869864321058551392517168916447091892905538462163946116125606530334728986875234648195066846665794605659183471385460265728466235192
104609903433540663274190445139049208042867753103952880330562960446966637621951188439209529587770403970820346979718970220109384932344
682862706198768118465379071730777739488540172140572710041711240590987049601015930431298393665302134882019356459039083232354350859950
400197746147557524213717815419388232208760572479008906475791837415205785207869765220154977534570583621293697828778729812956918242322
243866014158010741180843760257131118807398141776065933743223311922464663996621846450165732360643760878895145973524193871639568 7718559
385002466538703893864154486710547311324964475756677761499985536632668806372089272374075795686260656876469408218886131829438566823819
503704784246553487699811231273768658780788840052350619537012102658742028981727922832411121070946711398534018827505436582848144685358
751301395311349518682026204506856601608181591746590884271026146041003431968484615209384194658145802832127079217611184047147537712957
044801855542229782689848339912143435725968730679563617389482135022270704742308073198621779394971317601775723949868715858492 4081035
344920245012368080239325797453476061409967224359489693582014956927123893296584494722682302711343585835639361121074529682822566957164
825182439092023688248186995659266340840050608373233899246374717213049708204444237390884475765783794241610291502827239296107394463291
411299969590753143490962571948121546783175506647104312188718385503987020284444939260340274348386750098385907265902022305859036725656
383836433033529118930316560979685122334963007356105357573254339992985938987540915897522932562218599640414740266606489020107707571023
776069482963973302291586453698019860004734723371105883468841198507439304899958509663186874126879407436192229406461483739700701094445
780543497006667366566005622658216833863009147798663678691865859898555920306195122374178378590568981484865695340124863423432454635332
135922039356044091291780317990494094820023162219272674626587767137378988175186683538187756127971692946011778563519976889549722601263
061375201401696745350555795547136600855578055442113676660429405594571906726642211281285278194491598993723258160392082835078715377531
759848570998243034758485406864972968284949136777790837864829019750115188060264993950833939106605937604386097483270021958400724204775
365503527399619552299592721934086962804826136715375980999487319259299141091904947076400072754064978996420666266582721246303397088208
306341083878564694467976969071043464073761766242018878828698765644070993005630615021741972093624496059476512624899786639305924047452
171730181006159866173040024633438303655314841913790262723681908242429475726220857954357228963166339316975338991096525050099224540894
292445924054936472581960775493875397471788425088358471059677495966090050781957746693946520619508831294961094746546553449730551806 99
906727082539078009177099773671498557292305040816009802295717845781708867079939895617349432270395716771429128431663423860082359788815
805366927151006797761734502557671545952529984276548731479858546044775851538484655709744025945107721439366075649661274517996169692770
076567125168078346237453887629291073033801250600464351826861695104318029505982984305938887247338216664673875359862764185517400484924
766624364734304735441968343754077203446170448960863526672055194663224025244966461706585779723091184428547570148219576444894590253271
444511343395588195346830860345785366652701549024203357382426689313614795947454229238385998762363576550828435858537675805708618198848
825210238916984758199326591005277314022297682845323259396201473054218570091589417230161881091555957286030097999172683294010564411255
720177280514188648850808086846280407624149715075998032132931041973961581178694640486101271814727643042561200083655058609704958558
084714551971484191092647276618648565014230376035512674796230475239893725265411012701997025086677511892030818017240285872513474766633
861646577086540029749532702376617703475505376390527203306918011123215354269072353109019488499581521278121866000568339296951880540933
389369731872728508572048328967330434784517305694848990472660461062282689866467964467310322286293826380034097645367305183279944782 52
532322804114159934547670591949582509779228385745225195151234031852916470823038443545078393020686895873607381140076261131157326220229
526263791431126730374555159838177257815309899040389498442977028341585047047885876034022989704780116887166594975385598831727692817545
318034982098664777069308599585078011226231245764822943843162744202405347992972456632208605745374796806702842405990332216923863293761
518890539304971401411089043050949869008697058359750243719472943809230308271651642780803442365914046550745477957558701380808987710 71
753010071143117457464031937701226622909542031513987508821949867482319823982299254536034360779018156060823688001511709371295393133 91
686512545666548328360891814556893601027215483085768254249127067672509381378728618041364335009196386434890267258277079663091212209193
852090175754006083691583069565834515256212654442160716521218091449212384084751607262944003174981507232675704869245975398357756173429
078576931448344470097755336077389069491776638151947105856014184829427413929405906140101416618282498554280955238691801463041246726 84

5225627414900832504094689919483135776400769089354877790075838468918834249102046347799006069684387417616665941821635312202141738631604
41390120192856704093294666722326953328212622494130347605021198818533817968093326597227297709742461319807131544655335392114374 1145536
28789865118742736058751949208232151283199838804533588461362971446722727185807353682960716656186255013602223143230125523113214316473 2
24537088536660549451698589751704179581701925670471055956982951127488636967986949266600661944994025616831801938910380751699891 5007589
30038332019723005589212261589453797015699220607769692070846784350899011752314260689400831646160451573386601024253663692019020407086 3
48647566584353731952152550824843785616277745760351626218592040301492663959122976292666825349792513767568702356955185321154425 4685386
56764769423492167564419879293224616687904339153162130847543963799147680860631861946430119802742436799525170110870919032680110298161 1
31254660950725057853890040452326261804841590012517862735626010444281485929512388107564364905666194043108421859277386010920023 8543566
05523127011431470811345305048351850872743726943672584024324913324187284149211001684898694579661239012155141057509690555361322436716 1
32260582926577864481402087775070015703994318396780169309193992394276799446991727499613164878769230098368096971638387121334251 9360371
49755639122208670476841746835486770371471326202585514818561173680023609709988298957473765898737573274405365214397415312572873 5075514
19120699436794811131010538850436708602697821570469035591039515032969423091856932007509116970941989474987593231682080828479 75 15266704
03591037289713264403197112683868573337825044306892697582436319245489050744628438369477541193319884362550807227320877077658403 6204097
01835435582970512052485059604984618459124354125540055318576316947989608510135461806840086983265827177289254830636937 0349251603618026
28954717274973056442548901323767335747090074279071126520699556419984833118223204194874982864721575758662030104834094875647107 0124112
34035144350774123502385839431320726056604687639169893188656920965863921030249734247209251307807130346645713516421875082445348447694 0
93329467240585649668130275628950735495343764440196779364039568433448030982506720872875694959541267315810836267046021609119917 0947593
71729417875583318680447591074090547698870882211967775939482286102138802542131257015074079646617735639736596342830370436542987226221 9
27968043824896646847802785107623451908306418870005909615280583965531483028085913333906354608006717403007406994849795796217878 9060539
53756463115876317857993737532735207158110645589645515318209625670439491448044695139829300481273460985666572303750098259746915861749 9
98976097362501605809212892447139125535929797829739848403892917042710331419774236756580874348751250565212281069223974790669355 0018732
51493531276919259043744525554946960982467930388562707450137321564153367725130974558076040973556199983508761631579641650511132469967 8
66434982987729932568233779749330010829496529199945504006740169085185088602380798082463505543987535599695102558896025827684068 6463097
76275865049774344198762391444326296203340549754758323661937874697440337889806599317865023993008219307349237829723198963264081783861 1
54477260792623716426682674220192778227673193791829430741850147996923364092852894581577953051187010273093956729206367641239121 1570351
23009945323635298060609964277069477392424469252228086071201934085094024634720899449246574905150302397543603171485658778051301651976 5
50990221730849408436674528046156091363302841035614836392192803026923521588725539224456252237221825533629761001901927778065755 4216753
02103800152256116947938517537475874613329502775884434709388042315351115137940337468208684893051045137948659432305485116221006990091 4
37863504894140820615521660731073428360777447205897745611181505422231181589513853807957039203879580623536140713603889872213701 2811370
38581290097371071256513574451469088293131082375633435459701225748089241781437004217036050105849180719628235045440641105285195503968 8
21484984358484078160029212451998844219434352237493691605347449602731207733819942142694511719446434055108110170164180837975069 6584440
08357996950061039074358124003018666046047763613038364811101637344329423404615059845304456067640808226773250671630085868230287961879 9
72832731267847241347748518685426889402485405444454395146406388978815712711023279238231971815990798167360258937621543191876169 8162225
03876032829732500282190573361829657006247877921995318351954630865594347891525120730047175863119852341704630148495699628513129386016 2
10123952794791183467090945835053479197580234954898435491543816338979087580629156305334779378821209176702890052261291274475375 2622599
29153398951534681347167074437409870354640429558807945133065163463318404399505885648865556204215064321160208864242134416547752201556 6
64853139249531242126253950983092448687956122507998627787831564017120022342639546886260825996431253970197791445556319049011280 8710699
59711744469043983556839128213244244098078035773475995810541104604824005561729452565035050184665110757475558486983185236547230248518 44
85451745399688028265135909280951484902252205714788386959432766374276741445564845238768042021792542563371464575528420597417791 1142967
95486994836219699267267820795556508450358800044286971772415064245293227228975037470284988939151673489570056026386067803899477432326 2
27601728186817058595459482925672284402241146054854619133317699769930268590376575874032294999009370370233703419516293373988487 3462800
73984762464967841237257326074804702667815543814910448497031497452272150096041921953135724050464712546051114108773458378511143344 7937
43983339303669473974880591884790441726962682968059677878971186654513254636388275470903783088769921932902571122953545129865033 8476440
06459466966663452612400118384134882273826487440807156197352881789344734669099807810567364822844023773898530649062798647528201 1279934
88547387213173702487873474310699392717940941412493486129363023051161544734258717383093354358561349206394887320471329623 8 223506484795
84543206427865381800862867145023871183571574877863821673444439499188770435365634411074867975878744841396845172004053404409004 6829688
64462836319674266446934676371991440430425998133710637187530376245631288699685795608088948506346374396499760438112249513670106243873 3
68182227226903480760800082562949054907375060245482452585098862016917736165254338625782828474933946431376613925192595648573656 9595163
96764680690445417514690787794253820349670359942641845823492870910047853258770911485261680946491978748615673751728175828624638456614 6
09950724963417159699741100628714737320907008199359370779605156955820838462105010289393770497452688730738907842744389431996313 2489086
61125596565554699169333514037939801970220328982095987205845923290265990448542809816373896161636289530882952141669886359842695386411 5
50987455162293598955639614089819661775604836709056191152536356443192821424433019990393923864737533919848673878037351660955902 6971299
34993286013957096241722026058052965280804633242036517208345469003941677140445530377180786247254490464013071967815208255796015256441 5
82492326445426103813356059288546457526920120578426632188114792007047216298975711569639847208067418487841791570201907332491076 0942368
39668137179897868053036265483903498250385137459649755965876703420448779773669511124436429732151037256841016882481710176816998810453 0
97071553357728889361008602571976939831003385242346099100536730996681275886683169594781333946543635134222246890425402276575466 3414072
97687121385070025715491274332025855692388271512983016538646020127118175983662960359213329787213729751138309401721205407819280181029 2
18721026074561969979952945868411942152208731413340739273462173233546128722666368328664208146958421551377407508386482203528976 1433466
08352536069056255893458488905838323764190671891334316901882952603074065176218509843241204847996767879279141728536631115475847085660 9
54338107701443879854533661134454885827505360474356351816874539561416199380385390005184842040970387907610310317326532836460904 2587420
93680437613330330462877194996163242840296102710263730906029742596499877248544308690744783604430252231709723950797966221776392717751 1
31587186621195791984106480879369181724386464030464515337939018830796787442617494495465888986107515155388725360656632110120706 4393700
02414390383342852126728112358012437343209277368498134495145711528143262795649278377387579270495416464391880800901659187707303035285 0
20258671822956675991317003532351570859933674834495677677291784639840082392781850708452148398824363900392503196274152934170044 9584032
98348113589988049935186006259914199484053191570780021337851729437678052116835164792994219045889459748366547904584495184313887391364 0
74955462099028257636862420634953201014657226091396283753477411671179361137624125803344779415206079099416546836968138074430359 7245331
63859110841511391308279953628470054981189325185541972089177587351124287419297411489330639629545784354581875599669905539990290122314 4
08788257374243077513685136073244642739016083831474938019834096079771894520972517368899524724637448170679000409978591927237478 8748178
37777137807382257975664651785202662310521802263823925073058769287986931332980542217045668940369536783147247790479465562316798710356 0
44152354901511858276762937053154618304581005018290482268013791322147522640969091291087220477546175188616817399565422630851026 9441178
78678906027096270488353508262770025540048312523872339299419888313867083333181871519569551892937907847032459447346807301611772372419 7
76247354197785889539308720649337748859480260730019540583207072853005689236092256274648012069895779415446524729855031090027463 5945552
30544472017193197473387682935370060332207033976781991141318663426439148006502092607208392590552633945787981890182559090723255186973 6
36911660337268561306410057458210053813287452341323556878622309474299518192232288987624213492074502905232543230425870086282248 3116996
95148183756916820354650061106414475323867940508976670870138227155515531169102273429269258446228995619092810158053511708321908 5211155
66812934712378797006320606093375045159416775986597538365973645941334712493295879306730909088536772808645300592057762597114484 3963744
16080414592154992257772333546577677241219955850833644313427407334782666644173041743808946097186034933587156347661664812919170 0363545
34895380753046357480285273004344186205238526093084225416189539364784414767594964114069586393050278442740399755087862 3572024763978494
14095259175003032599825785866138050687150976502909896938656645796405913013806616940965035488262332929549501446955723036110686 9337160
81974264546906163673589128728860310492548316201462138646161469803181034017507308339201997823139196909405853258748098787355370571446 0
79509382362042938534438667693126107347693075649940099012418807013212489168825255804768485570561103198506376525572543063979134 3307967
70857818623154244877236239449971288467815117625149718435374299442683666634579429955389964234454675339756722425576930578698941353990 2
93571671910148243136750065875644704757330206533846899775381239588233149872836695829647967976004135299934578429966972848937475 4845728
77126128850446922421279904836292332222783172816179480827891759698073365390976395617532043919929693542291396318735135809536224094111 4
81204702474998024517134715930214026503124475687509989494473767642121291938960134632064762702385070296644433522884822893907223 656775
22231591822620724841483021652644096749080808293980959294146226856736600316494701425844662239578026359403303528066073299213499037257 3
13672065112706411234773205498654835599048487799761162645241362595871231642405472911080828857411714263670477903484942478101517 8387069
81140536605625621409501206363418548869613009686844268304403159253552787285396337702971090413795604072514992087168361471046184820529 5
88872847902949711707960518734034411177535970619942086580809768935083403837824551491504804243489780653362321295675970009279728 3872915

8416421053211500784239062291518524280166132060353624035712065730044034282762328431197340527506231110964391677781078708918436117454448
674997984840830856333336436927497266750454581559884179210905297277250328234360982258069027562800430137999335241614653027608555849218
615245268939883183864462851300599928062687671281755333790755197149468062409541837731927111598892873764119624840809556058876078649194
444150834791791933840681043233616751778432720971695798593278062146998913340556612209662617257453613220306014910433397877147675653819
500521573824901541933947114899421277567581672171266495480228532091775954256866701786835042416862670983116118904350075334562736379794
021515900699518622942277023550581694167652938618873276757029775974473775430695406669131405080099280799755214919025522918733893910177
555926368001864599474735410370251464664602542684218741373619005287244510217457542832712034325525340803379936325744266528445367557106
557756446204090795245451660298052194218623899913719016691970185176067260391037836924259818298403116765972151363814747519171466965234
275894071850079442526286323843191429021634880682296958087732025684892867257147848846335348220728351155539837734398210229138040058115
433983549954967189796202313011145070024850391918629668808342324444110091793432735089560857869767710005504229950524928422440215497246
669524923651531943568079608732353391211340998735787007417863492351205641493581162216125606977862735375462664694423570820765784261611
607386057078207694090560366881166265054882598567114654080564418039031201088995292718641301675217289479793854187622410085025338616696
799999137204327122978803348851281569387003054604914988714444100188073616206422737978928861240833197567208136226624342670099210098883
309457678527572157367714630307327468326251322452064719554849478806495869574815407054508730649106168765017313877286600419142414382456
533784752185470489351532959635952456634184963742863789488998566581021741337325278029255092003733671672441248911076577342872428315472
764897051413293068460040067721743713667741905651284637846746116549356502058104784289282351814001657257498312944259244462002996855478
211705254241914052668502781385510824998516537975118260969781181673300947731380119614530253652876196001615374337794995340114944565703
308618709388474025938024504000835519505561986827982989113513108315825780282836143619416360213372520006694935669842327086614835331123
957308258886631592217134678144219416304953887204164621547546385118711838387230199844837037395250236439431968987928458464451738791664
116990752003467764858336265140063969818856103570575656851200430984664700658837456852022192468641229710951580776056984562787942619583
611248954069885990611718188300624748004268908632140990776182357411420913724232421393100460577732341206291604457686898137588097814881
343291151303211927949763294891508188017838221678902524679809067295450834826626616423075179661547808050994795032416516757967699924017
494992446628727091071994209457927925258293510845634712493316084240145508025479645393036051922606686595857306601273083513520067393764
878527825614262460775201520799145713166667816771850684986743061409149429654871020748342940480137348382837487964608984055440193645649
492816055581362094832057652268816065132150604603440639621415937861039082904786553235678116578878642922977896433918371589970045277166
374089465805995809463875076070878405714127013449684900178413794259906164274946705115859651316069395590099481066590124411691739535647
530947161853415077834733048031550517804047097664549613579226569336437319457395711853456536453129151559173411421954567932102030327743
272306438364073054597951736155728860897932573407863769738438636783658543451090075354661537126173152013306852445090701152580617254378
747286548961671178564402047499030693305257150084459019908920288576037181228867593582854348265565739519965199876381689549395504411357
967907571403092183153900122573792216510387469732251364500211715047501009010072832556088133538102611296143820282645701269653314536570
238661188674347477992473094310696331832601736770418109821203891935164474038230271622841430424413470438975541390432170634349551610851
082760779577474321164835438085048021127413235859927916174134341677459004761144770952209901234653791758964806391389980011891874396980
082091000688300281323090539892181888827828781437326180184993836249376261762524847175744097121792164680753548750189935369021161191599
543780132107098067199235684902421922897541263749338205847132555423304070452940323325929761676179957753755386164417130601363078817134
960826989389837414099853204201329116396657722613560429207397863489998782144624608101766849707675934844176941207706698878562950393715
825843488838347727974783118178317186215008988399388517138410077825551001851200584511320564914661083287409659307030653915146239720663
077235342541446459253402449612246443909777836960802987155142626046177950026971472007142686140932563143574400431574119666057128037028
073483370471584046128812883188971731324938109559281731449252236681264260502524617938994272846691708387022567462643144642031553757682
959735047288834225969447095262854317005868694928682544604514715841941668798209967880395799597407036755720215573344522044833689973727
977484551612625451629487933534388088885801425698285725037341701847169065541828821991507153196243570475136978688593407373723624598560
668669133929869215418071844636081456996552819117477089829108329612421972381808986635225779322706508261246279839429246655369744483682
381776821096385417578978791813635571591778371757628521984835355077247648120049036422980703202489857878811569881108643471795590077355
848972295388974693437271676345932541466828080552620596431246062843154867585892332987487231209180721205971624491521705054011401237552
191654919499348106317643253919657022860126417954609462669223442858916108143577524460229251170210418615235429865572070825254397610456
537282141100987287411116338916616813019208946437365337164338075697360645609597280434126313073749785809353580317294187589575247460276
440527475776456551283274517453418432048681545585971313596504698928567523258374351139869093874495194892887491722471555376852946938487
318191171063251431296159380681053430095215094853539200765116280687094330416691977290768651878163087858378781797922181670439850381070
607967181630325784951420867095728499099614827774907720462266591827812720133890788748773815191782425775160156081658239504141905562666
711004655535235667827635882102188259993608602077989501024864646628622413271974532372201694345093938174532749137169332923241864468553
613692034370334048553422518261601191661073163600719824288953705730699707940277521616340418369457267027375674905025858725887462755955
883363262619673776672491849765400850665506203480748037163103853262951760649006326996578881818829188714196786629081558752770677696481
139291673238131607246241195200742553296619902517186204285804746121040017668314994488100577268050000574941104673770821371867780793247
598275529714315441874993621114365555628593184885033334062202716412506330378138729253272267535673540031498191228899256564113816406600
262425797460157104125618300717606681434924670222856761931013036365423556519282946845517716063249472965070177585525981496264853849234
283390421370976365507232255739104328247665264923280692055421061044342548338196634288516855087186503874167358410497403724254171490630
363941595952483126447392550606072500201148030063704835427634338339181331009609885349473933121036390554594808815649748327061741496068
942526151009060375374690232625200051435314165882268943181166537189280459509781237371394310586506673583858090183626779780790624522281
552082926827902405063981374413809013086347215987048695821640793909561238529065092881062491727240520323139157766528398206910394835842
807921236349189346644227342261033516899414441612764986185706965154941655098797212170169727730998258244407837907586646436286899969173
611145358311070504242278154467780759588549704186190092831992783921065920605854785493278124683801697644396000815087805014368603369617
323140180441063165477148345358046203871321284873216681604055727845273049067910094124882179155820985978167941954316137140730557458656
935439045497504026437132885684782330040115726281665332131289380470977309048739908563461087669630650503126232068778828812251993245299
197575380074248248126101454260987221215550387035333687019581902249407554570540903633071125712433135239587659216745041695184063622287
688080102292279806525467242923013357037601586551269728896074539265569516369069370191307307614476154735870524842910661776673594541483
856777453709699547746133400151039315228631678822040849852462664855405416434200942743043358977449374823785196112478209742779500684067
533466899781201042234556893410610426548506643794699577766158872521982511394827016788681235014257528904440791149103000388859924654786
341829758102976256431657891598994972468669057309205498902669545346907255993694132218754153760967591357209997746545745463175703910104
338939731435187457519690299881747334713993806444252381467442639699457474296478441127223165534385873750359019146295506954595601477409
019384084028112832993409176981149335873643110438021900559320471804012591268532433951123325239134428068796072438364863544842209696478
032442428664927890236643814740725527251138950707452388722539932906844591780134739040297407877300301741488294704404415408221401766560
614630900821714435248972006491490592848916031017750711628902578363473726922331342221003022194295210887230758404150922301470583358707
196074924663160708050145871708442699004793100096369125126555898608838525658827260144483015162889553779111392320514828477185488314809
544352869498973263663652964201675714981604443982442893882348005417123181902097716919347846652046036874806574056171395646223649207852
842580494583850704535751357957371292011410902178907483812987431177885066594069739653867732691496156021574385082595698143105533668107
248144579500360328349733904532789377955996594940688296932104659550638285332769478136210833423361322625258376695296012793095246903850
811608005154559017181236346520003767133551524992701690245328211151032902689405241669709615523474038472528127696627436841749544461889
359573224432993796846455784975252103357973729514780245207221358493234567800325346881198064480530726871006031059665290397317626078851
732354699745604637850885560322266367085522065160047878852157175138989702386014317078596863279351510627049684425496297280545946143109
142188468649437139410472713445731030325313447784961613185588979061574163072133334826291965215976225575093126260155714869683792216189
982388720751221506527188282869537649893850615755813865505817918198901305441439588393654836816396611349953761724729216386912472754049
356301462669323958462398782343919587056662349936410011545998771329539685216310378817335526417109663196723777008966798227030102658194
079801778381262667516081897739117909318252962090692428605914667519857979186593288893738012973490633237576425943882923838282659573467
505674059903814123598483562804987125700130725475432727111394400064939922613305649033903345463236627877351175825565298675160800221184
227445416782470369194182149613397053921618183637459570031481290890965289421572667731982263776202754926379689934919780230806685186306
954801642214696369430425067490145430762620965932259581629553265395827891407147810748455395592676231367924464193404952507670682717731
7262949900000

11.4 Apéndice D.- Manual de estadísticas utilizadas

Las estadísticas describen el orden en que aparecen en la tesis, tienen el propósito de facilitar la comprensión de las medidas estadísticas en particular y de esta tesis en general. Describen las medidas utilizadas especialmente en los capítulos 7 y 8, pero siendo útiles en otras partes de la tesis.

11.4.1 Estudio descriptivo

Se encarga de detallar ciertos rasgos de la población de estudio, tiene como objetivo describir las características sin preocuparse del por qué, lo que se busca es especificar ciertas características de las variables de interés.

El estudio descriptivo es muy útil para mostrar con precisión las dimensiones o formas de un fenómeno, suceso, hecho, contexto o situación. Dentro de este estudio el investigador debe ser capaz de definir o describir qué se va a medir y sobre qué se va a recolectar los datos (Torre, 2004).

11.4.1.1 Mínimo (Min)

El mínimo es el valor menor o igual que todos los valores de un grupo de datos, al momento de ordenar nuestros valores en orden ascendente se observa al valor mínimo como el primero en nuestra lista de datos, el valor mínimo puede repetirse dentro del conjunto de datos, es decir, se puede tener más de un valor mínimo.

El valor mínimo se utiliza para identificar valores atípicos dentro del conjunto de datos, los posibles errores en la entrada de datos, también es útil para observar en forma sencilla la dispersión de datos restándolo del valor máximo (Salazar Pinto et al., 2017).

11.4.1.2 Primer cuartil (1st Qu)

Los cuartiles son aquellos valores que dividen en cuatro partes al conjunto de datos, tienen aproximadamente el mismo número de observaciones es decir que del 100% de nuestros datos se van a dividir en 25%, 50%, 75% y 100%.

El primer cuartil o también llamado el cuartil inferior es el valor del 25% del conjunto de datos, los cuartiles pueden ayudarnos a evaluar la dispersión y la tendencia central. Para calcular el primer cuartil se debe verificar si nuestros datos son agrupados o no, en el caso de los datos no agrupados se verifica primero si el conjunto de los datos es par o impar, en caso de ser par se aplica la fórmula $\frac{n}{4}$ y en caso de ser impar se aplica $\frac{n+1}{4}$. En

forma similar se procede para el cálculo de los otros 2 cuartiles, la mediana y el cuartil superior (Moore, 2005)

11.4.1.3 Mediana (Median)

La mediana es aquel valor que luego de ordenar el conjunto de datos de forma creciente o ascendente ocupa el valor central. La mediana junto con la media y la varianza son estadísticos muy expresivos en una distribución, la mediana al contrario que la media siempre se va a ubicar en el centro de nuestros datos, en cambio la media puede estar desplazada hacia un lado o al otro por influencia de valores extremos. Para calcular la mediana es importante que nuestros datos estén ordenados de mayor a menor (o de menor a mayor) es decir que tengan un orden. Para el cálculo de la posición de la mediana debemos tener en cuenta si el número total de datos u observaciones es par o impar, si las observaciones son en cantidad par se utiliza la fórmula $\frac{n}{2}$ y si la cantidad son impares se utiliza la fórmula $\frac{n+1}{2}$. Las medianas de la muestra se representa por M_e y la mediana poblacional se representa por $\tilde{\mu}$ (Acevedo, 2006).

11.4.1.4 Media (Mean)

La media también se conoce como el valor promedio de un conjunto de datos numéricos, se puede calcular como la suma del conjunto de datos y dividido para la cantidad total de los mismos. Para calcular la media se verifica si los datos son agrupados o no lo son, en caso de ser agrupados se aplica la fórmula $\bar{x} = \sum \frac{x_i f_i}{n}$ y en caso de ser datos no agrupados simplemente sumamos cada uno de los datos y lo dividimos para el total. La media de la muestra se representa por $\bar{x}$ y la media poblacional se representa por μ (Triola, 2004).

11.4.1.5 Tercer cuartil (3st Qu)

El tercer cuartil es el valor menor que el 25% y a su vez mayor que tres cuartas partes de los datos que están en estudio, es decir mayor que el 75% de las observaciones. Para calcular el tercer cuartil se analiza si nuestros datos son agrupados o no, en caso de ser datos no agrupados debemos analizar si nuestros datos son en cantidad par o impar, en caso de ser cantidad par tenemos la fórmula $\frac{3n}{4}$ y en caso de que datos sean de cantidad impar tenemos la fórmula $\frac{3(n+1)}{4}$ (Berenson et al., 2006).

11.4.1.6 Máximo (Max)

El valor máximo es el valor más grande de una muestra a estudiar, al igual que el valor mínimo el valor máximo se ocupa para identificar posibles datos atípicos dentro de la muestra, también se ocupa en la evaluación de la dispersión ya que se compara el valor mínimo con el valor máximo. Para calcular el valor máximo de forma manual debemos ordenar nuestros datos de manera ascendente (o descendente), así el valor máximo se encontrará al inicio de nuestros datos de ser ordenandos descendentes y de lo contrario se encontrará el valor máximo al final de los datos ordenados (E. R. González, 1968).

11.4.1.7 Desviación estándar (sd)

Es la medida que ofrece información sobre la dispersión media de un conjunto de datos, la desviación estándar siempre va a ser mayor o igual que cero, esto quiere decir que la desviación estándar no puede tener un valor negativo. La desviación estándar indica que tan dispersos se encuentran los datos con respecto a la media, mientras mayor sea la desviación estándar mayor será la dispersión. La desviación estándar de la población se representa con el símbolo griego sigma σ y la desviación estándar de la muestra con la letra latina s. Para calcular la desviación estándar se calcula primero la varianza utilizando la fórmula $Var = \frac{\sum f(x_i - \bar{x})^2}{n}$, una vez calculada la varianza se calcula la desviación estándar con la fórmula $S = \sqrt{Var}$, obteniendo así el valor de la desviación estándar (Mode, 1990).

11.4.1.8 Intervalo intercuartil (IQR)

El rango (o intervalo) intercuartil es una medida de dispersión, que representa la diferencia o la distancia que existe entre el primer y el tercer cuartil, generalmente este intervalo se lo utiliza en el gráfico BoxPlot, el símbolo para representarlo es RIC, RQ o IQR. El intervalo intercuartil permite eliminar los datos que se encuentran extremadamente alejados, es altamente recomendable utilizar este intervalo cuando la medida de tendencia central utilizada es la mediana. Para calcular la diferencia entre los cuartiles uno y el cuartil tres se utiliza la fórmula $IQR = Q_3 - Q_1$, para recordar esta medida hay que pensar en intercuartílico que significa entre cuartiles mientras que al rango se entiende como distancia entre puntos, así se entiende al rango intercuartílico como la distancia que existe entre cuartiles (Madrigal, 1996).

11.4.1.9 Asimetría (Skewness)

El skewness es una medida que permite visualizar la simetría de la distribución de una variable respecto a la media. Existen tres tipos de asimetría empezando por la asimetría negativa que hace que la cola de distribución se alargue para los valores inferiores a la media, la simetría quiere decir que existen el mismo número de elementos a la izquierda y a la derecha de la media. Esta distribución se adapta a la forma de la campana de Gauss o la distribución normal, la asimetría positiva hace que la cola de la distribución se alargue a la derecha para valores superiores a la media. La asimetría se calcula por la fórmula de Karl Pearson $A_S = \frac{3(\bar{x} - Me)}{S}$ (Fernández et al., 2002).

11.4.1.10 Curtosis (Kurtosis)

La curtosis (apuntamiento) es una medida (de forma) que indica la cantidad de datos que están cercanos a la media. Existen varios tipos de curtosis: leptocúrtica, que indica que existe una gran concentración de los valores en torno a su media, mesocúrtica que indica una concentración normal de los valores en torno a su media, platicúrtica que indica una baja concentración de los valores en torno a su media. Para datos sin agrupar se utiliza la fórmula $g_2 = \frac{1}{N}\frac{\sum(x_i - \bar{x})^4}{\sigma^4}$, para datos agrupados se utiliza la fórmula $g_2 = \frac{1}{N}\frac{\sum f_i(x_i - \bar{x})^4}{\sigma^4}$ (R. B. Gutiérrez y Pere, 2010).

11.4.1.11 Gráfico de violín

El gráfico de violín es el resultado de la combinación de un diagrama BoxPlot y un diagrama de densidad girado y colocado a cada lado, para de esta manera mostrar la forma cómo se distribuyen los datos. Posee una línea negra gruesa en el centro que representa el intervalo intercuartil, también tiene una barra negra fina que se extiende desde ella, la cual representa el 95% de los intervalos de confianza y el punto blanco es la mediana. Este diagrama se utiliza para visualizar la distribución de los datos y su densidad de probabilidad (Iglesias Pedrejón, 2018).

11.4.1.12 Gráfico BoxPlot

Los gráficos o diagramas BoxPlot también conocidos como diagramas de caja y alambre son muy útiles para representar visualmente a un conjunto de datos numéricos, las líneas que se extienden en paralelo a los dos extremos de las cajas se las conoce como bigotes, el cual se utiliza para indicar la variabilidad fuera de los cuartiles superiores e inferiores

(Lozano y Fuentes, 2012). Dentro de estos diagramas se puede visualizar los datos atípicos, ya que suelen representarse como puntos individuales que se encuentran en línea con los bigotes, estos diagramas se los representa horizontal o verticalmente.

11.4.2 Normalidad

En estadística la normalidad también es conocida como estudio de la semejanza con la distribución de Gauss, la distribución normal se caracteriza por su simetría alrededor de una media que siempre coincide con la mediana y que es mesocúrtica. La distribución normal (Gaussiana) representa la forma en que se distribuyen los valores numéricos de las variables continuas en procesos naturales, tales como datos de notas, estatura, peso, etc., para determinar la normalidad se aplica las pruebas de Kolmogorov-Smirnov o Shapiro Wilk, que tienen como objetivo verificar si una muestra aleatoria tiene una distribución normal planteando sus hipótesis:

H_o: La muestra se aproxima (~) a una distribución normal (ℕ (μ, σ^2)
H_1: La muestra no se aproxima a una distribución normal (ℕ (μ, σ^2)

La distribución normal es significativa ya que posee ciertas propiedades importantes como que admite una única moda que coincide con su media y su mediana, la curva normal es asintótica al eje de las abscisas, el área bajo la curva normal es uno (considerando los límites hacia más y menos infinito), es simétrica con respecto a su media, según esto para este tipo de variables existe una probabilidad de un 0,05 de observar un dato mayor que la media y un 0,05 de observar un dato menor. Las pruebas de normalidad tienen como objetivo principal analizar cuánto difiere la distribución de los datos observados respecto a lo esperado si procediesen de una distribución normal con la misma media y desviación típica (Douglas y Marchal, 2018).

11.4.2.1 Gráfico de cuartiles (QQ)

La gráfica de probabilidad normal es una técnica gráfica para evaluar si un conjunto de datos está o no distribuido aproximadamente normalmente. Los datos se trazan contra una distribución normal teórica, de tal manera que los puntos deben formar una línea recta aproximada. Las salidas de esta línea recta indican desviaciones de la normalidad (Almenara Barrios et al., 2004).

11.4.2.2 Prueba de normalidad de Anderson-Darling

La prueba de normalidad de Anderson Darling es utilizada para probar si una muestra viene de una distribución normal. Esta prueba es una modificación de la prueba de Kolmogorov Smirnov donde se les da más importancia a las colas de la distribución que la prueba de Kolmogorov, en estadística esta prueba es una prueba no paramétrica sobre si los datos siguen una distribución normal. La fórmula para el estadístico determina si los datos vienen de una distribución con función acumulativa F, $A^2 = -N - \frac{1}{N}\sum(2i - 1)(lnF(Y_I) + \ln(1 - F(Y_{N+1})))$ (Razali y Wah, 2011).

11.4.2.3 Prueba de normalidad de Lilliefors (Kolmogorov-Smirnov)

Es una prueba no paramétrica que ayuda a determinar la bondad de ajuste de 2 distribuciones de probabilidad entre sí, es decir, ayuda a verificar la normalidad de una distribución. La prueba de Lilliefors es mejor con respecto a la prueba de Kolmogorov-Smirnov en la cual se basa, además se utiliza cuando la hipótesis nula no especifica el valor esperado y la varianza de la distribución. Aunque sea mejor que la prueba de Kolmogorov no es útil en la parte práctica, ya que como la gran mayoría de las veces desconocemos cuál es la media y cuál es la desviación estándar de la población, se deben estimar; esto genera que sea muy conservadora aceptando desde ya la hipótesis nula en la mayoría de las ocasiones (Gonzalez et al., 1977).

11.4.2.4 Prueba de normalidad de Cramer-von Mises

Esta prueba al igual que la prueba de Kolmogorov-Smirnov también se basa en una comparación de la función de distribución empírica denotada por F_n y la función de distribución planteada en la hipótesis nula. En general este estimador acepta o rechaza la hipótesis nula.

H_o: Los datos provienen de una distribución normal.

H_1: Los datos no provienen de una distribución normal.

Mediante la comparación de la función de distribución empírica de los datos, denotada por F_n y la función de distribución teórica F_0 (Laio, 2004).

11.4.2.5 Pearson chi-square normality test

Es una prueba no paramétrica que ayuda a medir la discrepancia entre una distribución de frecuencias observadas y esperadas. Una de sus características generales es que toma valores entre 0 e infinito y no tiene valores negativos ya que es la suma de los valores elevados al cuadrado. Para hacer uso de este contraste hay que disponer de los datos en una tabla de frecuencia, para cada valor o intervalo de valores se indica la frecuencia absoluta observada o esperada. Este estadístico tiene una distribución chi cuadrado con k- 1 grados de libertad si n es suficientemente grande, es decir, si todas las frecuencias esperadas son mayores que 5. Una de las limitaciones es que la muestra debe ser lo suficientemente grande ya que si tenemos menos del 20% de las celdas de la tabla de contingencia esta presentará valores esperados <= 5, la cual no se recomienda aplicar una prueba x^2. Si existe concordancia perfecta entre las frecuencias observadas y esperadas el estadístico tomará un valor igual a cero; pero si sucede lo contrario cuando exista una gran discrepancia entre estas frecuencias, el estadístico tomará un valor grande, y, en consecuencia, se rechazará la hipótesis nula (Gaboardi y Rogers, 2018).

11.4.2.6 Prueba de normalidad de Shapiro-Wilk

La prueba de Shapiro-Wilk sirve para contrastar la normalidad de un conjunto de datos, se plantea como:

H_o: La muestra proviene de una población normalmente distribuida.

H_1: La muestra no proviene de una población normalmente distribuida.

Se dice que se rechaza la hipótesis nula de normalidad si el estadístico W es menor que el valor crítico proporcionado por la tabla para el tamaño muestral y el nivel de significación dado (Darling, 1957).

11.4.2.7 Prueba de normalidad de Shapiro-Francia

La prueba de Shapiro-Francia es una modificación de la prueba Shapiro-Wilk y se basa en una aproximación del estimador σ bajo el supuesto de normalidad, la prueba de Shapiro Francia tiene como estadístico W'. El estadístico W' tiene la ventaja de obtener mejores aproximaciones con menor costo computacional, además el comportamiento asintótico del estadístico W'permite trabajar apropiadamente con tamaños de muestra grandes (Montilla y Kromrey, 2010) .

11.4.3 Normalización

Es el proceso que permite buscar que una muestra se considere que proviene de una población normalmente distribuida. En este trabajo de tesis se utilizó la función Power transformation de R, sobre todo considerando su actualidad y facilidad de uso.

11.4.3.1 Normalización usando Power Transformation

Power transformation, es una familia de funciones aplicadas para crear una transformación monótona de datos utilizando funciones de potencia, también es conocido como una técnica de transformación de datos que se utiliza para estabilizar la varianza, hacer que los datos tengan una distribución más normal o en el caso que tengamos un conjunto de datos no normales esta función permite transformar a datos normales. Utiliza el camino de máxima verosimilitud de Box y Cox para escoger una transformación para normalidad, tiene como opciones bcPower, bcnPower, yjPower (Draper y Cox, 1969).

11.4.3.2 bcPower

bcPower se encuentra dentro de la familia de funciones de PowerTransform como una de las opciones que se encarga de transformar los elementos de un vector o columnas de una matriz usando Box-Cox, la función bcPower calcula la transformación de potencia escalada de x=U+gamma, donde gamma lo establece el usuario, por lo que U+gamma es estrictamente positivo para que estas transformaciones se puedan realizar (Fox y Weisberg, 2011).

11.4.3.3 yjPower

yjPower, también es parte de la familia de funciones de powerTransform, es uno de los argumentos. Se ocupa de datos positivos, es decir, este argumento no admite datos negativos, en el caso de utilizar datos negativos se utiliza el parámetro 2-lambda (Fox y Weisberg, 2011).

11.4.4 Homocedasticidad

Un grupo de muestras de datos se dice homocedásticos, si las varianzas de las muestras en análisis son estadísticamente similares. Las hipótesis de homocedasticidad para dos variables son $H_0: \sigma_1^2 = \sigma_2^2$, y $H_1: \sigma_1^2 \neq \sigma_2^2$. Una de las utilidades más importantes del contraste de homocedasticidad es que si se quiere por ejemplo probar la hipótesis de la diferencia entre las medias de 2 poblaciones normales, podemos encontrarnos con 2

situaciones: las varianzas son iguales, en este caso es el más favorable pues utilizamos la distribución t de student, caso contrario podría utilizar un test no paramétrico como Wilcoxon o Mann Whitney (Montilla y Kromrey, 2010).

11.4.4.1 Prueba de Levene

Esta prueba se utiliza para evaluar la igualdad de las varianzas para dos o más grupos, algunos procedimientos estadísticos comunes asumen que las varianzas de las poblaciones de las que se extrae en diferentes muestras son iguales, es por eso que la prueba de Levene evalúa este supuesto. Se pone a prueba la hipótesis nula de que las varianzas poblacionales, son iguales, si el p-valor resultante de la prueba es inferior a un cierto nivel de significación 0,05 es poco probable que las varianzas sean iguales, por lo tanto, se rechaza la hipótesis nula de igualdad de varianzas, y se concluye que hay una diferencia entre las varianzas de la población. La prueba de Levene se utiliza a menudo antes de una comparación de medias para decidir si se utiliza una prueba paramétrica o una no paramétrica. El estadístico de prueba W se define con la fórmula: $w = \frac{(N-k)}{(K-1)} \frac{\sum_{i=1}^{k} N_i (z_i - z_{...})^2}{\sum_{i=1}^{k} \sum_{j=1}^{Ni} (z_{ij} - z_i)^2}$, donde, w es el resultado de la prueba, K es el número de grupos, N es el número total de casos en todos los grupos, N_i es el número de casos en el grupo i-ésimo (Nordstokke y Zumbo, 2010).

11.4.4.2 Prueba de Bartlett

La prueba de Bartlett sirve para probar si k muestras provienen de poblaciones con la misma varianza. A las varianzas iguales a través de 2 muestras se llama homocedasticidad u homogeneidad de varianzas. Algunas pruebas estadísticas como el análisis de varianza ANOVA suponen que las varianzas son iguales en todos los grupos o muestras, la prueba de Bartlett se utiliza para verificar esta suposición. Un datoy importante sobre la prueba de Bartlett es que es sensible a las desviaciones de la normalidad, es decir, si las muestras provienen de distribuciones no normales, entonces la prueba de Bartlett puede servir simplemente para probar la normalidad. La prueba de Levene es una alternativa para la prueba de Bartlett ya que es menos sensible a las desviaciones de la normalidad. La prueba de Bartlett sirve para probar la hipótesis.

Ho: Todas las k varianzas de la población son iguales

H1: Al menos 2 varianzas de la población son diferentes.

Luego de tener planteadas las hipótesis, se calcularel estadístico de prueba con la fórmula: $x^2 = \frac{(N-k)\,ln(S_P^2) - \sum_{i=1}^{k}(n_i-1)\,ln(S_i^2)}{1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i-1}\right)-\frac{1}{N-K}\right)}$, donde s_p^2, es la estimación combinada de la varianza. El estadístico de prueba tiene aproximadamente una distribución χ_{k-1}^2. Así la hipótesis nula se rechazará si $\chi^2 > \chi_{k-1}^2$. La prueba del Barnett es una modificación de la correspondiente prueba de razón verosimilitud diseñada para hacer que la aproximación de la distribución χ_{k-1}^2sea mejor (Correa et al., 2006).

11.4.5 Independencia

Cuando los valores de una variable no se ven afectados por los valores que toma otra variable se dice que son independientes, esto es cuando la distribución condicional no se ve afectada por la condición y coinciden en todos los casos con las frecuencias relativas marginales. Entonces, las variables son independientes cuando para todos los pares de valores se cumple que la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales $\frac{n_{i,j}}{N} = \frac{n_{i,}}{N} * \frac{n_{,j}}{N}$. La independencia de sucesos es algo muy importante para estadística y es una condición necesaria para la aplicación de multitud de teoremas (Montgomery et al., 1996).

11.4.6 Pruebas de hipótesis

Una prueba de hipótesis es una regla que especifica si se puede aceptar o rechazar una afirmación acerca de una población dependiendo de la evidencia proporcionada por una muestra de datos. Una prueba de hipótesis tiene dos afirmaciones opuestas sobre una población la hipótesis nula H_0, y la hipótesis alternativa H_1. Por lo general la hipótesis nula es un enunciado que indica que no hay efecto o que no hay diferencia, la hipótesis alternativa es la que se desea poder concluir que es verdadero de acuerdo con la evidencia proporcionada por los datos de la muestra. Con base en los datos de la muestra se determina si se rechaza la hipótesis nula, se debe utilizar el valor de p para tomar esta decisión, si el valor de p es menor que el nivel de significancia es decir menor que α, entonces se puede rechazar la hipótesis nula (Inzunsa Cazares y Jiménez Ramírez, 2013).

11.4.6.1 Pruebas paramétricas y no paramétricas

Las pruebas no paramétricas son aquellas que a pesar de basarse en determinadas suposiciones no parten de la base de que los datos analizados adopten normalidad, homocedasticidad e independencia. Además, se puede decir que es una técnica estadística que no presupone ninguna distribución de probabilidad teórica de la distribución de los datos, por ello también se conoce como una distribución libre. En la mayor parte de ellas los resultados estadísticos se derivan únicamente a partir de procedimientos de ordenación por lo que su base lógica es fácil de comprender. Cuando se trabaja con muestras pequeñas (menor a 10) en las que se desconoce si es válida la normalidad de datos, conviene utilizar pruebas no paramétricas al menos para corroborar los resultados obtenidos a partir de la utilización de la teoría basada en la normal, en estos casos se emplea como parámetro de centralización la mediana. Las pruebas no paramétricas no requieren asumir normalidad de la población y en su mayoría se basan en el ordenamiento de los datos, además la población tiene que ser continua. También son técnicas estadísticas que no presuponen ningún modelo probabilístico teórico además son menos potentes que las técnicas paramétricas, aunque tienen la ventaja de que se pueden aplicar más fácilmente. La Tabla 11.1 resume las pruebas no paramétricas y paramétricas.

Tabla 11.1.- Pruebas no paramétricas y su equivalente paramétrico (Castor et al., 2013)

TIPO DE ANÁLISIS	PRUEBA NO PARAMÉTRICA	PRUEBA PARAMÉTRICA EQUIVALENTE
Comparación de dos muestras relacionadas	Prueba de rangos con signo de Wilcoxon	Prueba t para muestras dependientes
Comparación de dos muestras independientes	Prueba U de Mann-Whitney	Prueba t para muestras independientes
Comparación de más de dos muestras relacionadas	Prueba de Friedman	Análisis de varianza de Medidas repetidas (ANOVA)
Comparación de más de dos muestras independientes	Prueba H de Kruskal-Wallis	ANOVA de una vía

11.4.6.2 Prueba de hipótesis de 5 pasos

Para las pruebas de hipótesis existen 3 valores que puede tomar α (alfa nivel de riesgo o nivel de error), éstos pueden ser 0,01, 0,05 y 0,1, es decir en un estudio siempre se tratará de trabajar con el menor riesgo posible, con un α de 0,01, el estudio será mucho más significativo con grado de confiabilidad del 99%, esto quiere decir que el nivel de riesgo tan solo será del 1%. El siguiente valor menor que toma α es 0,05, ya que muchas veces los estudios no salen tan factibles con un α de 0,01 de tal modo que reducimos a

α=0,05 es decir ya no será del 99% ahora será al 95% de confiabilidad y tendrá un nivel de riesgo del 5%, si con el 0,05 en estudio aún no es factible cambiamos al tercero y último valor, donde α toma el valor de 0,1 ahora el estudio será solo al 90% de confiabilidad y tendrá un nivel del riesgo del 10%, si con este valor el estudio aún no es factible habría que buscar otros métodos ya que α no puede tomar valores menores que éstos, ya que si se realiza un estudio con valores muy altos de α, el estudio no serviría y habría un nivel de riesgo muy grande. La prueba de hipótesis es un procedimiento basado en la evidencia encontrada en una muestra y el uso de la teoría de probabilidad para determinar si la hipótesis es una afirmación razonable también en la población. Las pruebas de hipótesis generalmente se realizan utilizando 5 pasos (que se utilizan también en esta tesis) los cuales son:

1. Establecer la hipótesis nula y la alternativa.
2. Seleccionar el nivel de significancia.
3. Identificar el estadístico de prueba.
4. Formular la regla de decisión.
5. Tomar la decisión.

PASO 1: Establecemos la hipótesis nula (Ho) y la hipótesis alternativa (H19). La Hipótesis nula, que es un enunciado relativo al valor de un parámetro que se formula con el fin de probarlo mediante evidencia numérica. La hipótesis alternativa, que es la negación del enunciado de la hipótesis nula.

PASO 2: Seleccionamos el nivel de significancia Alfa $\alpha = 0{,}05$ es el nivel de significancia (error o riesgo) es decir el complemento del nivel de confianza (95%) denotado en porcentajes.

PASO 3: Identificamos el estadístico de prueba, que puede estar relacionado con el estadístico Z, estadístico t, estadístico χ^2, Estadístico F, etc. y además calculamos el estadístico de prueba.

PASO 4: Formular la regla de decisión, si se utiliza el p-valor es generalmente de la forma si el p-valor es menor que 0,05 entonces se rechaza la Hipótesis Nula, caso contrario no se la rechaza.

PASO 5: Tomar la decisión, comparando α y el estadístico de la prueba. Si el p-valor es mayor a α, se rechaza la hipótesis nula pero si sucede lo contrario rechazamos la hipótesis alternativa (Fallas, 2012).

11.4.6.3 Kruskal Wallis H-Test

Esta prueba es un método no paramétrico para comparar las medianas de n muestras, es equivalente a la prueba ANOVA con los datos reemplazados por categorías. Las hipótesis por demostrar son:

H_o: $\tilde{\mu}_1 = \tilde{\mu}_2 = \cdots = \tilde{\mu}_n$

H_1: $\tilde{\mu}_i \neq \tilde{\mu}_j \ \ para\ al\ menos\ un\ par\ de\ (i,j)$

Además, es una extensión de la prueba de U de Mann-Whitney para n grupos. El estadístico viene dado por la fórmula: $K = (N-1)\dfrac{\sum_{i=1}^{G} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{G}\sum_{j=1}^{n_i}\left(r_{i_j} - \bar{r}\right)^2}$ donde n_i es el número de observaciones en el grupo *i*, r_{i_j} es el rango entre todas las observaciones del grupo *j* en el grupo *i*, *N* es el número total de observaciones entre todos los grupos. $\overline{r_i} = \dfrac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$, $r = \dfrac{(N+1)}{2}$,es el promedio de r_{i_j}. Notamos que el denominador de la expresión para *K* es exactamente: $\frac{(N-1)N(N+1)}{12}$. Luego K= $\frac{12}{N(N+1)}\sum_{i=1}^{G} n_i(\bar{r}_i - \bar{r})^2$. Se puede realizar una corrección para los valores repetidos (ligaduras) dividiendo *K* por $1 - \dfrac{\sum_{i=t}^{G}(t_i^3 - t_i)}{N^3 - N}$, donde *G* es el número de los grupos diferentes repetidos, y t_i es el numero de observaciones repetidas dentro de cada grupo i que tiene observaciones repetidas para un denominado valor. Esta corrección hace cambiar a *K* muy poco al menos que exista un gran número de observaciones repetidas. Y, por último, el p-valor es aproximado por $P_r(\chi^2_{G-1} \geq k)$ (Soto, 2013).

11.4.6.4 Mann Whitney Wilcoxon U-test

Esta prueba se usa para comparar las medianas de 2 muestras independientes, la interpretación del valor p permite encontrar evidencia a favor o en contra de la igualdad de las medianas. Esta prueba es robusta para evitar el error de Tipo 1 y tiene mayor poder

estadístico cuando se compara con otras pruebas, específicamente cuando la de distribución de las muestras tienen sesgo. Esta prueba es semejante a la t de student para muestras independientes, en términos de aceptar o rechazar la hipótesis nula, cuántos los datos originales son reemplazados por su rango. Para aplicar la prueba de Mann Whitney Wilcoxon:

1. Primero se combinan los datos de 2 muestras en una sola, y se ordenan de menor a mayor;
2. Sean n1 el número de datos de la muestra 1 y n2 el número de datos de la muestra 2.
3. Luego se obtienen los rangos que ocupan los datos de la muestra única, al dato más pequeño se le asigna 1; al que sigue en magnitud, el número 2 y así sucesivamente. En el caso de que 2 o más datos tengan el mismo valor en el rango, en este proceso se utilizará el promedio de todos los rangos asignados.
4. En esta estructura de datos se debe hacer una anotación del origen de cada dato, es decir, si el dato pertenece a la muestra 1 o a la muestra 2, posteriormente se separan los rangos de la muestra única, determinando a que muestra pertenecen.
5. De la muestra 1 se obtienen los rangos.
6. Con este procedimiento se obtiene el estadístico U, el cual está representado por la fórmula $U = S + 0{,}5 \times n_1 \times (n_1 + 1)$, donde *S* es la suma de los rangos de la muestra n1, este valor de U se encuentra en las tablas apropiadas para esta prueba y así se obtiene el valor de p (Turcios, 2015).

11.4.6.5 ANOVA no paramétrico

La estimación basada en rangos para modelos estadísticos es una alternativa no paramétrica robusta a los procedimientos de estimación clásicos como los mínimos cuadrados. Los métodos de rangos se han desarrollado para modelos que van desde modelos lineales a modelos lineales mixtos, series de tiempo y modelos no lineales. Las ventajas de los métodos de rangos sobre los métodos tradicionales como la máxima verosimilitud o los mínimos cuadrados, es que requieren menos suposiciones, son robustos a valores atípicos y son altamente eficientes en una amplia gama de

distribuciones. La función Rfit utiliza la sintaxis de modelo lineal estándar e incluye funciones de uso común para procedimientos de inferencia y diagnóstico, para esto se utiliza el paquete de R, Rfif el cual es la estimación e inferencia basadas en rangos para los modelos lineales, una de las funciones del paquete es el ANOVA no paramétrico, el análisis de varianza unidireccional basado en rangos utiliza el comando oneway.rfit, que realiza un análisis de varianza robusto para un diseño de un factor, el análisis se basa en las estimaciones de rango, el comando está compuesto de la siguiente manera: oneway.rfit(y, g, scores = Rfit::wscores, p.adjust = "none"), actualmente proporciona la opción de Tukey para métodos de ajuste de intervalo de confianza, así como para la comparación entre pares (Rfif,2020).

11.5 Apéndice E.- Programas

Se describe la instalación de los programas base para realizar la comparación de los métodos clúster y reglas de asociación realizados.

11.5.1 Programa estadístico R

R es un lenguaje y un entorno para gráficos y computación estadística. Es un proyecto GNU que es similar al lenguaje y entorno S que fue desarrollado en Bell Laboratories (antes AT&T, ahora Lucent Technologies) por John Chambers y sus colegas. R se considera como una implementación diferente de S. Hay algunas diferencias importantes, pero gran parte del código escrito para S se ejecuta inalterado en R. R proporciona una amplia variedad de técnicas estadísticas (modelos lineales y no lineales, pruebas estadísticas clásicas, análisis de series de tiempo, clasificación, agrupamiento) y técnicas gráficas, y es altamente extensible. El lenguaje S es a menudo el vehículo de elección para la investigación en metodología estadística, y R proporciona una ruta de código abierto para participar en esa actividad. Uno de los puntos fuertes de R es la facilidad con la que se pueden producir gráficos con calidad de publicación bien diseñados, incluidos símbolos matemáticos y fórmulas cuando sea necesario. Se ha tenido mucho cuidado con los valores predeterminados para las opciones de diseño menores en los gráficos, pero el usuario conserva el control total. R está disponible como software libre bajo los términos de la Free Software Foundation‘s Licencia Pública General de GNU en forma de código fuente. Se compila y se ejecuta en una amplia variedad de plataformas UNIX y sistemas similares (incluidos FreeBSD y Linux), Windows y MacOS.

Para la instalación de R primero se descarga el programa de la página web de R en donde está la versión más actualizada del software al momento de utilizarlo (Figura 11.1).

Figura 11.1.- Página web de R (*R*, 2021)

Para el trabajo de investigación se necesitó el uso del paquete Rchic, mismo que funciona con la versión de R 3.5.2 o menor, por lo cual se procedió con su respectiva instalación.

Una vez descargada la versión 3.5.2 de R (Figura 11.2), se procedió a instalar dicho archivo haciendo doble clic sobre él (Figura 11.3).

Figura 11.2.- Versión 3.52 de R (*Download R-3.5.2 for Windows. The R-project for statistical computing.*, 2021)

Al iniciar la instalación el programa solicita que se seleccione el idioma a utilizar como se puede ver en la Figura 11.3.

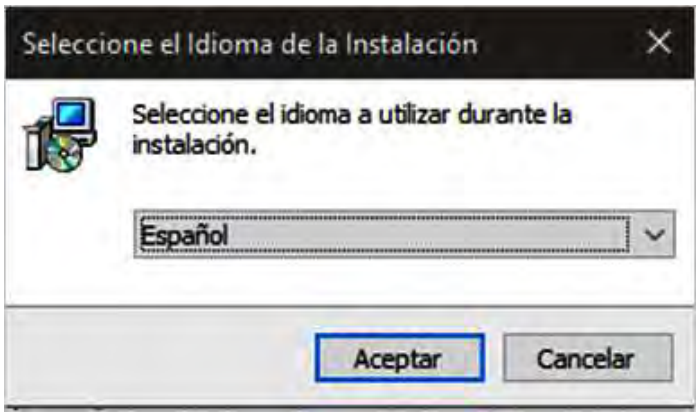


Figura 11.3.- Iniciando la Instalación

A continuación, se observará una ventana en donde se escoge la ubicación de la carpeta en donde se instalará el programa R (Figura 11.4).

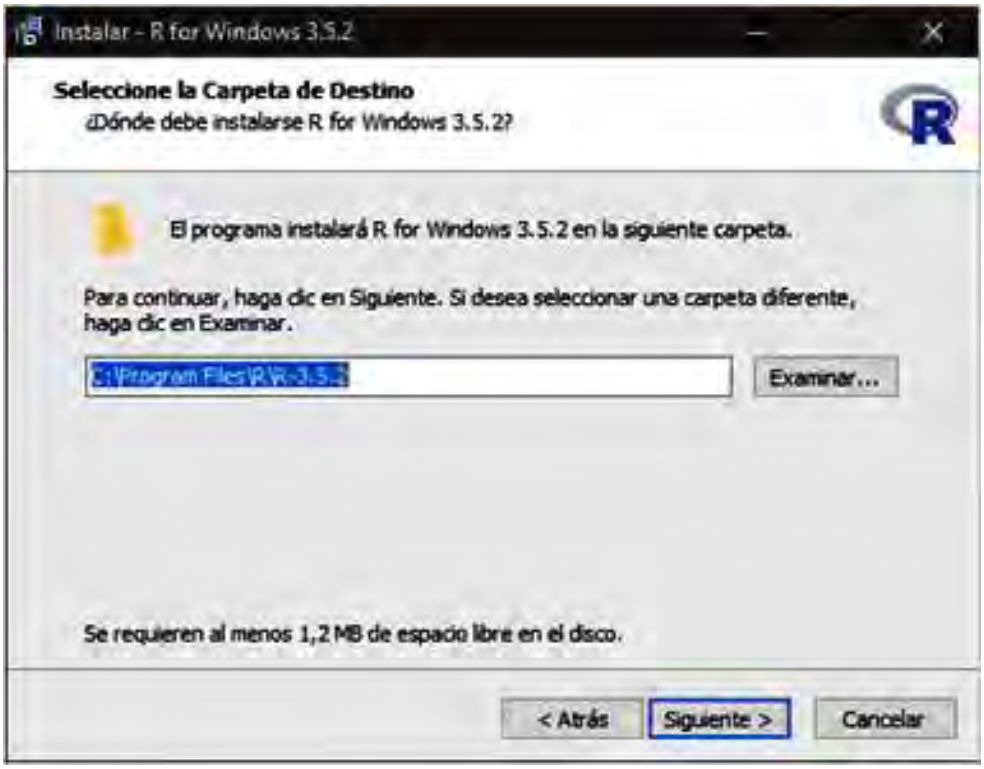


Figura 11.4.- Ubicación del programa luego de instalarlo

Haciendo clic en el botón siguiente hasta que se visualice la ventana presentada en la Figura 11.5, se observará el proceso de instalación, una vez completado hay que hacer clic en finalizar.

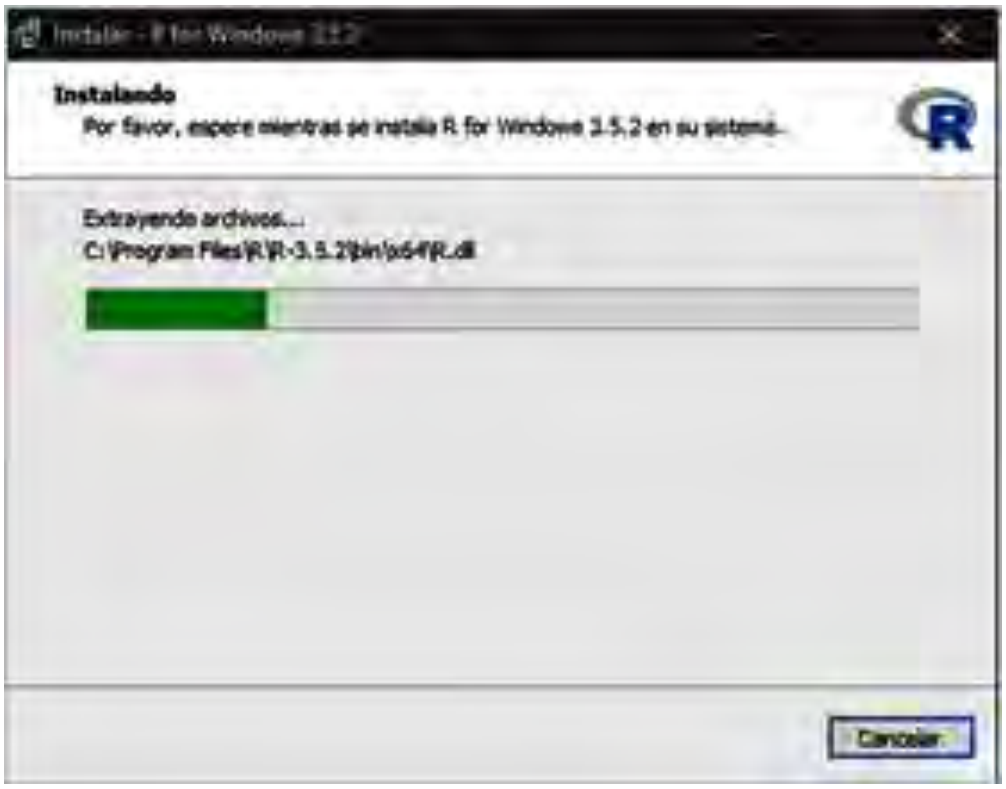


Figura 11.5.- Instalación de R

R es un conjunto integrado de funciones utilizadas para la manipulación, el análisis estadístico y visualización gráfica de datos. Además, permite (*The Comprehensive R Archive Network*, 2015):

- Instalación eficaz de manejo y almacenamiento de datos.
- Conjunto de operadores y funciones para cálculos con números, vectores, matrices, etc.
- Una colección amplia, coherente e integrada de herramientas estadísticas para el análisis de datos.

- Facilidades gráficas para la visualización en pantalla o impresión en papel.
- Conjunto de paquetes especializados en diferentes áreas estadísticas.
- Un lenguaje de programación orientado a objetos simple y efectivo que incluye condicionales, bucles, funciones recursivas definidas por el usuario e instalaciones de entrada y salida.

Los programas R y S, están diseñados en torno a un verdadero lenguaje informático y permite a los usuarios agregar funciones adicionales mediante la definición de nuevas funciones. Gran parte del sistema está escrito en el dialecto R de S, lo que facilita a los usuarios seguir las elecciones algorítmicas realizadas. Para tareas de computación intensiva, el código C, C ++ y Fortran se puede vincular y llamar en tiempo de ejecución. Los usuarios avanzados pueden escribir código C para manipular objetos R directamente. Muchos usuarios piensan en R como un sistema de estadísticas. Preferimos pensar en él como un entorno en el que se implementan técnicas estadísticas. R se puede ampliar (fácilmente) mediante *paquetes*. Hay unos ocho paquetes que se suministran con la distribución R y muchos más están disponibles a través de la familia de sitios de Internet CRAN que cubren una amplia gama de estadísticas modernas (Ihaka y Gentleman, 1996; Team, 2013).

R tiene su propio formato de documentación similar a LaTeX, que se utiliza para proporcionar documentación completa, tanto en línea en varios formatos como en papel (*R*, 2021).

11.5.2 El entorno de desarrollo integrado RStudio

RStudio es un entorno de desarrollo integrado (IDE) para R (Allaire, 2012). Incluye una consola, un editor de resaltado de sintaxis que admite la ejecución directa de código, así como herramientas para el trazado, el historial, la depuración y la gestión del espacio de trabajo.

RStudio está disponible en ediciones comerciales y de código abierto y se ejecuta en el escritorio (Windows, Mac y Linux) o en un navegador conectado a RStudio Server o RStudio Server Pro (Debian / Ubuntu, Red Hat / CentOS y SUSE Linux) (*RStudio*, 2020).

Una vez que está instalado R, se descarga RStudio y también se lo instala (Arifin, 2019). De acuerdo con el sistema operativo en el cual deseamos instalar RStudio, se descargará el fichero ejecutable (Figura 11.6), la versión que se muestra de ejemplo es la versión 1.3.1093.

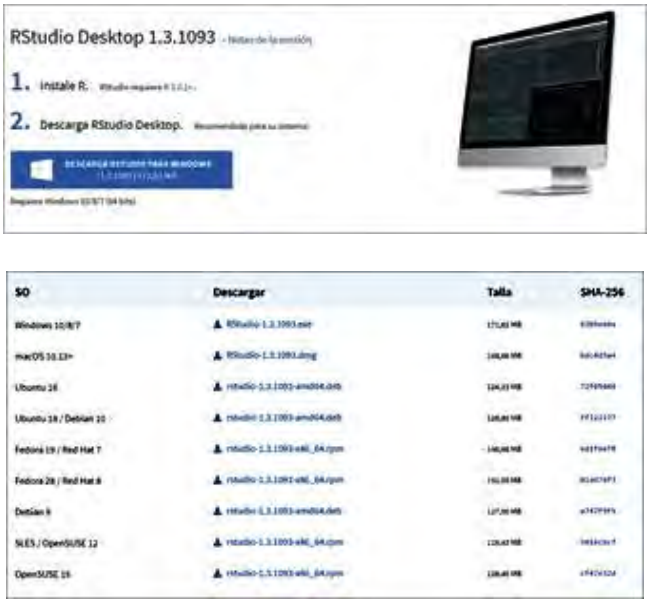


Figura 11.6.- Descargar RStudio

Finalizada la descarga, se selecciona la opción abrir, preguntará si realizaremos cambios, hacemos clic en la opción sí, e inmediatamente se desplegará la ventana de Instalación de RStudio (Figura 11.7).

Figura 11.7.- Asistente Instalación RStudio

Se puede presionar en el botón siguiente hasta que se muestre la ventana que indica que se está instalando RStudio, esperar hasta que la instalación se complete y dar clic en el botón terminar (Figura 11.8).

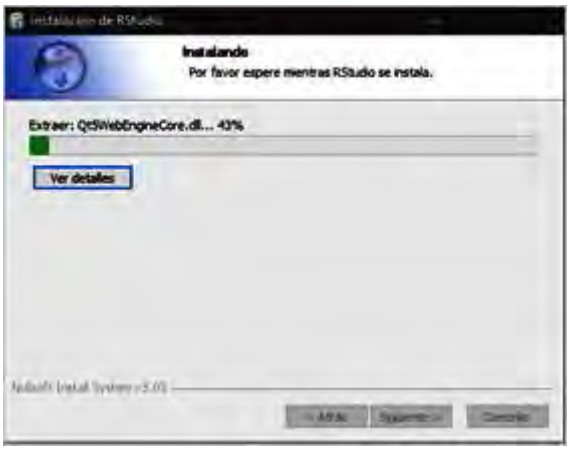


Figura 11.8.- Instalación RStudio

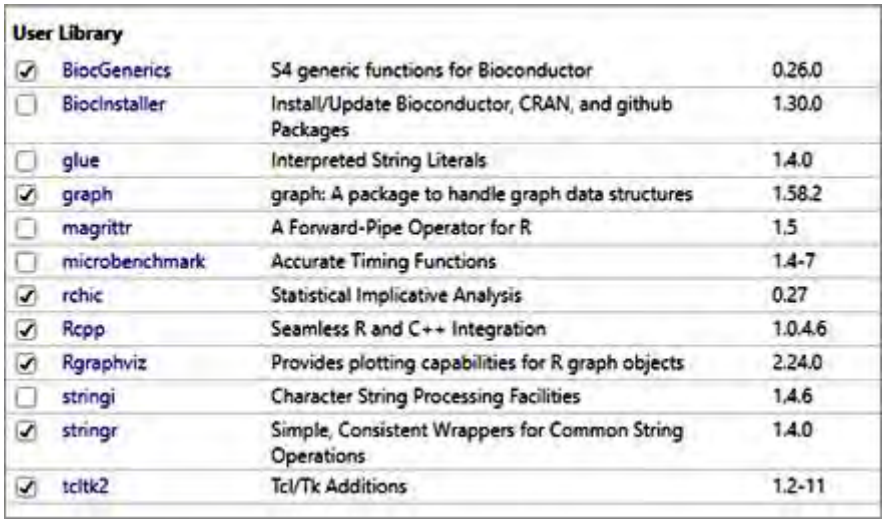
Una vez completada la instalación ya se tendrá disponible para su uso RStudio.

11.6 Apéndice F.- Principales paquetes de R utilizados

A continuación, se amplía la información sobre los paquetes de R utilizados en los capítulos en los cuales se realiza la comparación de los métodos clúster y reglas de asociación realizados.

11.6.1 Rchic

Rchic es un paquete de R que implementa la mayoría de las técnicas del Análisis Estadístico Implicativo (Raphael Couturier, 2016). Para instalar el paquete Rchic versión 0.27 encontrado en la página web de Raphael Couturier (*Rchic*, 2016), se requieren algunos paquetes los cuales deben ser previamente instalados y seguidamente activados (Figura 11.9):



User Library			
☑	BiocGenerics	S4 generic functions for Bioconductor	0.26.0
☐	BiocInstaller	Install/Update Bioconductor, CRAN, and github Packages	1.30.0
☐	glue	Interpreted String Literals	1.4.0
☑	graph	graph: A package to handle graph data structures	1.58.2
☐	magrittr	A Forward-Pipe Operator for R	1.5
☐	microbenchmark	Accurate Timing Functions	1.4-7
☑	rchic	Statistical Implicative Analysis	0.27
☑	Rcpp	Seamless R and C++ Integration	1.0.4.6
☑	Rgraphviz	Provides plotting capabilities for R graph objects	2.24.0
☐	stringi	Character String Processing Facilities	1.4.6
☑	stringr	Simple, Consistent Wrappers for Common String Operations	1.4.0
☑	tcltk2	Tcl/Tk Additions	1.2-11

Figura 11.9.- Paquetes necesarios antes de la instalación de Rchic

Para ingresar Rchic en la consola de RStudio, se usa el comando Rchic() el cual visualizará el área de trabajo de dicho paquete (Figura 11.10).

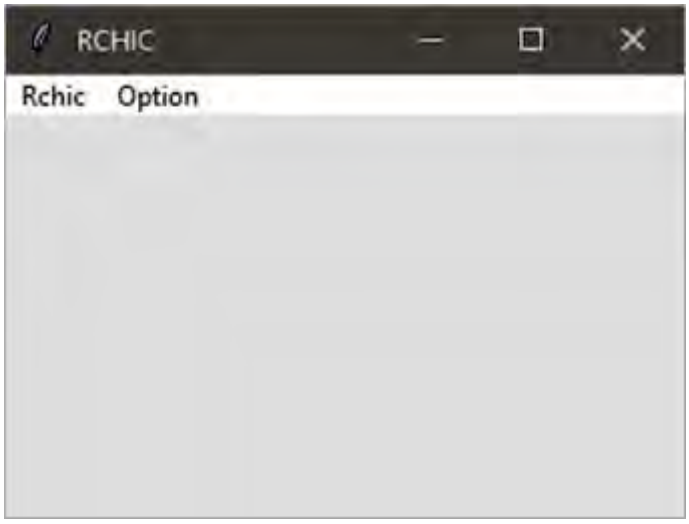


Figura 11.10.- Ventana de trabajo de Rchic

11.6.2 Microbenchmark

Proporciona infraestructura para medir y comparar con precisión el tiempo de ejecución las expresiones de R. Este paquete (microbenchmark) sirve como un reemplazo más preciso de la expresión system.time(replicate(1000,expr)). Se esfuerza por medir con precisión solo el tiempo que se tarda en evaluar expr. Para lograr esto, se utilizan las funciones de sincronización precisas de sub-milisegundos (supuestamente nanosegundos) que proporcionan la mayoría de los sistemas operativos modernos. Además, todas las evaluaciones de las expresiones se realizan en código C para minimizar cualquier sobrecarga (Mersmann et al., 2019).

11.6.3 Ggplot2

Permite la creación de gráficos sofisticados, el cual está basado en "La gramática de gráficos". Es difícil describir cómo funciona ggplot2 porque encarna una profunda filosofía de visualización. Sin embargo, en la mayoría de casos comienza con ggplot() un conjunto de datos y un mapeo estético (con aes()). Luego agrega capas (como geom_point() o geom_histogram()), escalas (como scale_colour_brewer()), especificaciones de facetas (como facet_wrap()) y sistemas de coordenadas (como coord_flip()) (Wickham et al., 2020).

11.6.4 Clúster

Técnicas de análisis clúster, ampliamente extendido es original de Peter Rousseeuw, Anja Struyf y Mia Hubert "Finding Groups in Data" (Maechler et al., 2019). Con respecto a los algoritmos de partición son enfoques de agrupamiento que dividen los conjuntos de datos, que contienen n observaciones, en un conjunto de k grupos (es decir, conglomerados). Entre las principales opciones se encuentran:

- K-means clúster, en el que, cada grupo está representado por el centro o los medios de los puntos de datos pertenecientes al grupo.
- K-medoids agrupación o pam en la que, cada grupo está representado por uno de los objetos en el grupo. Es una alternativa “no paramétrica” de la agrupación de k-means.

11.6.5 FactoExtra

Proporciona funciones para extraer y visualizar la salida de análisis de datos multivariados, incluidos 'PCA' (análisis de componentes principales), 'CA' (análisis de correspondencia), 'MCA' (análisis de correspondencia múltiple), 'FAMD' (Análisis factorial de datos mixtos), funciones 'MFA' (análisis factorial múltiple) y 'HMFA' (análisis factorial múltiple jerárquico) de diferentes paquetes R. También contiene funciones para simplificar algunos pasos de análisis de agrupamiento y proporciona una elegante visualización de datos basada en 'ggplot2' (Kassambara,Alboukadel, 2020).

11.6.6 Fastcluster

Este paquete proporciona interfaces tanto para R como para Python, implementa rutinas rápidas de agrupamiento en clúster jerárquicos y aglomerativos. Como parte de su funcionalidad está diseñado como reemplazo directo de las rutinas existentes: linkage () en el paquete 'SciPy' 'scipy.cluster.hierarchy', hclust () en el paquete stats de R y el paquete 'flashClust'. Proporciona la misma funcionalidad con el beneficio de una implementación mucho más rápida. Además, existen rutinas de ahorro de memoria para la agrupación de datos vectoriales, que van más allá de lo que proporcionan los paquetes existentes (Daniel Müllner, 2018).

11.6.7 CluMix

Proporciona utilidades para agrupar sujetos y variables de tipos de datos mixtos (*CRAN - Package CluMix*, 2016). Las similitudes entre sujetos se miden mediante el coeficiente de similaridad general de Gower con una extensión de Podani para las variables ordinales. Las similitudes entre las variables pueden evaluarse

- mediante la combinación de medidas de asociación apropiadas para diferentes pares de tipos de datos o
- basándose en la correlación de distancia. Alternativamente, las variables también se pueden agrupar mediante el enfoque 'ClustOfVar'.

La característica principal del paquete es la generación de un mapa de calor de datos mixtos. Para visualizar similitudes entre sujetos o variables, se dibuja un mapa de calor de la matriz de distancia correspondiente. Las asociaciones entre variables se exploran mediante un 'gráfico de confusión', que permite la detección visual de posibles factores de confusión, colineales o sustitutos para algunas variables de interés primario. Las matrices de distancia y dendrogramas para sujetos y variables se pueden derivar y utilizar para visualizaciones y aplicaciones adicionales (*CRAN - Package CluMix*, 2016)**.**

11.6.8 Dplyr

Es una herramienta rápida y coherente para trabajar con marcos de datos como objetos, tanto en la memoria como fuera de la memoria (Wickham,Hadley et al., 2020). Tiene tres objetivos principales:

- Identifica las herramientas de manipulación de datos más importantes necesarias para el análisis de datos y hacerlas fáciles de usar desde R.
- Proporciona un rendimiento ultrarrápido para los datos en memoria escribiendo piezas clave en C++.
- Utilizar la misma interfaz para trabajar con datos sin importar dónde estén almacenados, ya sea en un marco de datos, una tabla de datos o una base de datos.

11.6.9 Replyr

El paquete replyr, proporciona posibilidades prácticas de manipulación de datos para hacer que el código funcione de manera similar en datos locales o remotos (de gran tamaño). Replyr proporciona métodos para controlar el trabajo con tbl fuentes remotas (SQLbases de datos Spark) dplyr. La idea es agregar funciones de conveniencia para que tales tareas se parezcan más a trabajar con una memoria interna data.frame. Los resultados aún dependen del dplyr servicio que utilice, pero replyr tiene un acceso bastante uniforme a algunas funciones útiles. La regla general es: probar dplyr primero, y si eso no funciona, verificar si replyr ha investigado una solución alternativa. Replyr utiliza interfaces estándar a favor de la captura, para que pueda programar fácilmente sobre replyr (Mount, John, 2020).

11.6.10 Arules

Permite representar, manipular y analizar patrones y datos de transacciones, también proporciona implementaciones en C de los algoritmos de minería de asociaciones Apriori y Eclat (Hahsler,Michael et al., 2020). El paquete arules para R proporciona la infraestructura para representar, manipular y analizar patrones y datos de transacciones utilizando conjuntos de elementos frecuentes y reglas de asociación. También proporciona una amplia gama de medidas de interés y algoritmos de minería que incluyen interfaces y el código de las eficientes implementaciones en C de Borgelt de los algoritmos de reglas de asociación Apriori y Eclat.

11.6.11 ArulezViz

Extiende las 'arules' del paquete con varias técnicas de visualización para reglas de asociación y conjuntos de elementos, el paquete también incluye varias visualizaciones interactivas para la exploración de reglas (Hahsler, 2019). Tiene como principales características:

- Visualizaciones con gridy plotly.
- Visualizaciones interactivas con gridy plotly.
- Inspección de reglas interactivas usando datatable.
- Exploración de reglas interactivas integradas con ruleExplorer.

11.7 Apéndice G.- Funciones

A continuación, se identifican las principales funciones utilizadas en el análisis de datos, utilizados en los capítulos en los cuales se realiza la comparación de las técnicas clúster y reglas de asociación realizados.

11.7.1 Técnicas clúster

Las funciones clúster utilizadas en el capítulo 7 y que intervinieron en el análisis se muestran en la Tabla 11.2.

Tabla 11.2.- Funciones definidas por el usuario para las técnicas clúster

CLUSTER METHODS	Paquete	Función	Comando
hrarchy	Rchic 0,24 (2018)	callHierarchyTree	`callHierarchyTree (x, contribution.supp = FALSE, typicality.supp = FALSE, computing.mode = 3, verbose = FALSE )`
simlrty	Rchic 0,24	callSimilarityTree	`callSimilarityTree(x, contribution.supp=FALSE, typicality.supp=FALSE, verbose=FALSE)`
dendro_diana	cluster 2.0,7 (2018-04-05)	diana	`diana(df, metric = "euclidean", stand = TRUE)`
dendro_variables	CluMix 1.0,5 (2017-08)	dendro.variables	`dendro.variables(df, method="distcor")`
hclust_vector	Fastcluster 1.1.24 (2018-06-07)	hclust.vector	`hc<-hclust.vector(df, method="single", members=NULL, metric='euclidean', p=NULL)`

11.7.2 Técnicas de reglas de asociación

Las funciones clúster utilizadas en el capítulo 8 y que intervinieron en el análisis se muestran en la Tabla 11.3.

Tabla 11.3.- Funciones definidas por el usuario para las técnicas de asociación

ASSOCIATION METHODS	Paquete	Función	Comando
met_apriori	arules 1.6-2 (03/12/2018)	apriori	apriori(df, parameter = list(supp = 0,5, maxlen = 3, conf = 0,6, target = "rules"))
met_ASI	Rchic 0,24 (2018)	Implicative Graph	`callSimilarityTree(x, contribution.supp=FALSE, typicality.supp=FALSE, verbose=FALSE) sm<-similarity_matrix implicativeGraph(sm, list.variables = list.variables, computing.mode = 1, complete.graph = 0)`
met_eclat	arules 1.6-2 (03/12/2018)	eclat	`eclat(df, parameter = list(supp = 0,5, maxlen = 3))`
met_weclat	arules 1.6-2 (03/12/2018)	weclat	`weclat(df, parameter = list(support = 0,5, maxlen = 3), control = list(verbose = TRUE))`

11.8 Apéndice H.- Códigos de R

A continuación, se muestra el código completo de los principales programas utilizados en la generación de las bases de datos y en su análisis, aplicados en los capítulos en los cuales se realiza la comparación de las técnicas clúster y reglas de asociación realizadas.

11.8.1 Generador de bases de datos

```
#====================================================================#
# DATE: May 24, 2016
# AUTOR: Ruben Pazmino
# OBJETIVE: CATEGORIACAL File Generation (Var=100,Subjects=1000, DB = 10
a la 5)
# POST ANALYSIS: for Association Rule Minning and Cluster Time & Memory
Comparition
# OPERATING SYSTEMS:
#====================================================================#
rm(list = setdiff(ls(), lsf.str()))
closeAllConnections()
# memory.limit(100000) #is Windows-specific
# memory.limit()  #is Windows-specific
# memory.size()  #is Windows-specific

variablesNumber <- 100
subjetsNumber   <- 1000 #total 383 hacer 400
ownpath<-c("C:/Users/PC/Documents/R/ruben")#First move for ownpath
directory
beginFile<-1
endFile<-400
for (i in beginFile:endFile){
  nVar<- round(runif(1),2)*variablesNumber
  nFilas<- round(runif(1),3)*subjetsNumber

  if (nVar<3) {
    nVar<- round(runif(1),2)*100
    if (nVar<3) {
      nVar<- round(runif(1),2)*100
    }
  }

  if (nFilas<3) {
    nFilas<- round(runif(1),3)*1000
  }

  DataBase<-replicate(nVar, runif(nFilas) ) # Random matrix generate
  #DataBase<-round(DataBase,3) #Category Cuantity
  ncol(DataBase)

  # Data Base "name" creation
  if (nVar==ncol(DataBase)){
    rownames(DataBase)<-paste('C',1:nFilas,sep='')
    colnames(DataBase)<-c(';V1',paste('V',2:nVar,sep=''))
```

```
    f<-
paste('File_',toString(i),'_C',toString(nFilas),'xV',toString(nVar),'.csv
',sep="")  # SIA input file name
    write.table(DataBase, file=paste(ownpath,f,sep=""), sep=";" ,quote =
FALSE) # write SIA input file
    rm(DataBase)
  }
}
```

11.8.2 Para el análisis de los datos de las técnicas clúster

```
#================================================#
    # File   : Clustering.R
    # Created: july 2017
    # Updated: july 2019
    # Content: Clustering memory analysis
    # Author : Rubén Pazmiño-Maji
    #================================================#
    install.packages("Rcmdr")
    library(Rcmdr)
    #For Rcmdr clusterMemory1 <-
read.table("Clustering/MEMORY/clusterMemory1.csv",header=TRUE,
sep=",",dec=",")#For Rcmdr
#====================================== M E M O R Y
=======================================#
    # Read Clustering memory databases
    clusterMemory1 <-
read.table("Clustering/MEMORY/clusterMemory1.csv",header=TRUE,
sep=",",dec=".")
    head (clusterMemory1)
    #1) ==== Descriptive Statistics MEMORY ====
        #1.1) DataBase Characteristics MEMORY (ALL VARIABLES)
            summary(clusterMemory1)
        #1.2) NDATA & MEMORY
            numSummary(clusterMemory1[,"MEMORY", drop=FALSE],
           groups=clusterMemory1$NDATA,
           statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.3) OPERATING.SYSTEM & MEMORY
            numSummary(clusterMemory1[,"MEMORY", drop=FALSE],
            groups=clusterMemory1$OPERATING.SYSTEM,
            statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.4) CLUSTER.METHODS & MEMORY
            numSummary(clusterMemory1[,"MEMORY", drop=FALSE],
            groups=clusterMemory1$CLUSTER.METHODS,
            statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.5) (NDATA1=NDATAqualitative) & MEMORY
            numSummary(clusterMemory1[,"MEMORY", drop=FALSE],
           groups=clusterMemory1$NDATA1,
           statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.6) (OS_METHOD = OPERATING.SYSTEM , CLUSTER.METHODS) & MEMORY
            numSummary(clusterMemory1[,"MEMORY", drop=FALSE],
           groups=clusterMemory1$OS_METHOD,
           statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.7)(NDATA2 = NDATA1 , CLUSTER.METHODS)  & MEMORY
            numSummary(clusterMemory1[,"MEMORY", drop=FALSE],
           groups=clusterMemory1$NDATA2,
           statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.8 (COLUMNS1=COLUMNSqualitative) & MEMORY
```

```
            numSummary(clusterMemory1[,"MEMORY", drop=FALSE],
           groups=clusterMemory1$COLUMNS1,
           statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.9) (ROWS1=ROWSqualitative) & MEMORY
            numSummary(clusterMemory1[,"MEMORY", drop=FALSE],
           groups=clusterMemory1$ROWS1,
           statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.10) (COLUMNS2 = COLUMNSqualitative ,CLUSTER.METHODS) & MEMORY
            numSummary(clusterMemory1[,"MEMORY", drop=FALSE],
            groups=clusterMemory1$COLUMNS2,
            statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.11) (ROWS2 = ROWSqualitative ,CLUSTER.METHODS)& MEMORY
            numSummary(clusterMemory1[,"MEMORY", drop=FALSE],
            groups=clusterMemory1$ROWS2,
            statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))

    #2) ==== Descriptive Graphics MEMORY ====
        library(ggplot2)
        clusterMemory1 <-
read.table("Clustering/MEMORY/clusterMemory1.csv",header=TRUE,
sep=",",dec=".")
        head (clusterMemory1)
        #2.1) DataBase Characteristics MEMORY
        # No graphics are important
        #2.2) NDATA & MEMORY
            # No graphics, many data
        #2.3) OPERATING.SYSTEM & MEMORY
            attach(clusterMemory1)
            X <- OPERATING.SYSTEM
            Y <- MEMORY
            Xlabel  <- "OPERATING_SYSTEM"
            Ylabel  <- "MEMORY"
            df <- data.frame( Xlabel=factor(X), Ylabel=Y )
            G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
            G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
            T  <- theme(legend.position="top")
            G1 + geom_histogram( fill="white",position="dodge") + T
            G2 + geom_boxplot()+ T
            G2 + geom_violin() + T
        #2.4) CLUSTER.METHODS & MEMORY
            attach(clusterMemory1)
            X <- CLUSTER.METHODS
            Y <- MEMORY
            Xlabel  <- "CLUSTER_METHODS"
            Ylabel  <- "MEMORY"
            df <- data.frame( Xlabel=factor(X), Ylabel=Y )
            G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
            G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
            T  <- theme(legend.position="top")
            G1 + geom_histogram( fill="white",position="dodge") + T
```

```
    G2 + geom_boxplot()+ T
    G2 + geom_violin() + T
#2.5) (NDATA1=NDATAqualitative) & MEMORY
    attach(clusterMemory1)
    X <- NDATA1
    Y <- MEMORY
    Xlabel  <- "DATA_NUMBER"
    Ylabel  <- "MEMORY"
    df <- data.frame( Xlabel=factor(X), Ylabel=Y )
    G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
    G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
    T  <- theme(legend.position="top")
    G1 + geom_histogram( fill="white",position="dodge") + T
    G2 + geom_boxplot()+ T
    G2 + geom_violin() + T
#2.6) (OS_METHOD = OPERATING.SYSTEM , CLUSTER.METHODS) & MEMORY
    attach(clusterMemory1)
    X <- OS_METHOD
    Y <- MEMORY
    Xlabel  <- "OPERATING.SYSTEM&CLUSTER_METHODS"
    Ylabel  <- "MEMORY"
    df <- data.frame( Xlabel=factor(X), Ylabel=Y )
    G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
    G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
    T  <- theme(legend.position="top")
    G1 + geom_histogram( fill="white",position="dodge") + T
    G2 + geom_boxplot()+ T
    G2 + geom_violin() + T
#2.7) (NDATA2 = NDATAqualitative, CLUSTER_METHODS)  & MEMORY
    attach(clusterMemory1)
    X <- NDATA2
    Y <- MEMORY
    Xlabel  <- "NDATA&CLUSTER_METHODS"
    Ylabel  <- "MEMORY"
    df <- data.frame( Xlabel=factor(X), Ylabel=Y )
    G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
    G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
    T  <- theme(legend.position="top")
    G1 + geom_histogram( fill="white",position="dodge") + T
    G2 + geom_boxplot()+ T
    G2 + geom_violin() + T
#2.8 (COLUMNS1=COLUMNSqualitative) & MEMORY
    attach(clusterMemory1)
    X <- COLUMNS1
    Y <- MEMORY
    Xlabel  <- "COLUMNS_NUMBER"
    Ylabel  <- "MEMORY"
    df <- data.frame( Xlabel=factor(X), Ylabel=Y )
    G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
    G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
    T  <- theme(legend.position="top")
    G1 + geom_histogram( fill="white",position="dodge") + T
    G2 + geom_boxplot()+ T
    G2 + geom_violin() + T
```

```
        #2.9)(ROWS1 = ROWSqualitative) & MEMORY
            attach(clusterMemory1)
            X <- ROWS1
            Y <- MEMORY
            Xlabel  <- "ROWS_NUMBER"
            Ylabel  <- "MEMORY"
            df <- data.frame( Xlabel=factor(X), Ylabel=Y )
            G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
            G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
            T  <- theme(legend.position="top")
            G1 + geom_histogram( fill="white",position="dodge") + T
            G2 + geom_boxplot()+ T
            G2 + geom_violin() + T
        #2.10)(COLUMNS2 = COLUMNSqualitative ,CLUSTER.METHODS)& MEMORY
            attach(clusterMemory1)
            X <- COLUMNS2
            Y <- MEMORY
            Xlabel  <- "COLUMNS_NUMBER"
            Ylabel  <- "MEMORY"
            df <- data.frame( Xlabel=factor(X), Ylabel=Y )
            G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
            G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
            T  <- theme(legend.position="top")
            G1 + geom_histogram( fill="white",position="dodge") + T
            G2 + geom_boxplot()+ T
            G2 + geom_violin() + T
        #2.11)(ROWS2 = ROWSqualitative ,CLUSTER.METHODS)& MEMORY
            attach(clusterMemory1)
            X <- ROWS2
            Y <- MEMORY
            Xlabel  <- "ROWSNUMBER&CLUSTER_METHODS"
            Ylabel  <- "MEMORY"
            df <- data.frame( Xlabel=factor(X), Ylabel=Y )
            G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
            G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
            T  <- theme(legend.position="top")
            G1 + geom_histogram( fill="white",position="dodge") + T
            G2 + geom_boxplot()+ T
            G2 + geom_violin() + T
    #3) ==== Assumptions (NORMALITY, HOMOCEDASTICITY, INDEPENDENCE)
MEMORY ====
            library(nortest)  #Test de Anderson-Darling y Test de
Lilliefors (Kolmogorov-Smirnov)
            library(moments) #Test de Jarque Bera#
            library(Johnson) #JOHNSON TRANSFORM
            library(car)#Box-Cox TRANSFORM
        clusterMemory1 <-
read.table("Clustering/MEMORY/clusterMemory1.csv",header=TRUE,
sep=",",dec=".")
        #3.1) DataBase Characteristics MEMORY (ALL VARIABLES)
            #ASSUMPTION OF GLOBAL NORMALITY
                attach(clusterMemory1)
                Y <- MEMORY
                #Start Normality graphics
```

```
                hist(Y,probability=TRUE, main="Histogram of MEMORY in
Clustering data",xlab="Approximately normally distributed data")
                lines(density(Y),col=2)
                qqnorm(Y,main="QQ plot of MEMORY data",pch=19)
                qqline(Y)

                #Start Normality TESTS (FIVE NORMALITY TESTS)
                ad.test(Y) #Test de Anderson-Darling
                lillie.test(Y) #Test de Lilliefors (Kolmogorov-Smirnov)
                cvm.test(Y) #Cramer-von Mises normality test
                pearson.test(Y) #Pearson chi-square normality test
                jb.norm.test(Y)# #Test de Jarque Bera#

            #GLOBAL NORMALITY TRANSFORMS (FIVE NORMALITY TESTS, FOUR+TWO
TRANSFORMS)
                ad.test(log(Y)) #LOG TRANSFORM
                ad.test(1/Y) #EXP TRANSFORM
                ad.test(RE.Johnson(c(Y))$transformed) #JOHNSON TRANSFORM
                ad.test(boxCoxVariable(c(Y))) #Box-Cox TRANSFORM

                lillie.test(log(Y)) #LOG TRANSFORM
                lillie.test(1/Y) #EXP TRANSFORM
                lillie.test(RE.Johnson(c(Y))$transformed) #JOHNSON
TRANSFORM
                lillie.test(boxCoxVariable(c(Y))) #Box-Cox TRANSFORM

                cvm.test(log(Y)) #LOG TRANSFORM
                cvm.test(1/Y) #EXP TRANSFORM
                cvm.test(RE.Johnson(c(Y))$transformed) #JOHNSON TRANSFORM
                cvm.test(boxCoxVariable(c(Y))) #Box-Cox TRANSFORM

                pearson.test(log(Y)) #LOG TRANSFORM
                pearson.test(1/Y) #EXP TRANSFORM
                pearson.test(RE.Johnson(c(Y))$transformed) #JOHNSON
TRANSFORM
                pearson.test(boxCoxVariable(c(Y))) #Box-Cox TRANSFORM

                jb.norm.test(log(Y)) #LOG TRANSFORM
                jb.norm.test(1/Y) #EXP TRANSFORM
                jb.norm.test(RE.Johnson(c(Y))$transformed) #JOHNSON
TRANSFORM
                jb.norm.test(boxCoxVariable(c(Y))) #Box-Cox TRANSFORM

                summary(powerTransform(Y ~ 1, data=clusterMemory1,
family="bcPower")) #Box-Cox Power Transformation to Normality
                summary(powerTransform(Y ~ 1, data=clusterMemory1,
family="yjPower"))#Yeo-Johnson Power Transformation to Normality
        #3.2) NDATA & MEMORY
                # Many data to do normality test
        #3.3) OPERATING.SYSTEM & MEMORY
                library(Rcmdr)
                attach(clusterMemory1)
                X <- OPERATING.SYSTEM
                Y <- MEMORY
```

```
    # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
      normalityTest(Y~X, test="ad.test")
      normalityTest(Y~X, test="lillie.test")
      normalityTest(Y~X, test="cvm.test")
      normalityTest(Y~X, test="pearson.test")
      normalityTest(Y~X, test="shapiro.test")
      normalityTest(Y~X, test="sf.test")
    # NORMALITY TRANSFORMS BY GROUPS
      summary(powerTransform(Y ~ X, family="bcPower"))
      summary(powerTransform(Y ~ X, family="yjPower"))
    # HOMOCEDASTICITY ASSUMPTION
      Xfact<- factor(X)
      leveneTest(Y,Xfact, center = "median")
      bartlett.test(Y,Xfact, center = "median")
    # INDEPENDENCE ASSUMPTION
      chisq.test(X,MEMORY1)
      # install.packages("summarytools", dependencies = FALSE)
      # library("summarytools")
      # frequencyMEMORY<-freq(MEMORY)
      # frequencyMEMORY[,1]

#3.4) CLUSTER.METHODS & MEMORY
      library(Rcmdr)
      attach(clusterMemory1)
      X <- CLUSTER.METHODS
      Y <- MEMORY
    # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
      normalityTest(Y~X, test="ad.test")
      normalityTest(Y~X, test="lillie.test")
      normalityTest(Y~X, test="cvm.test")
      normalityTest(Y~X, test="pearson.test")
      normalityTest(Y~X, test="shapiro.test")
      normalityTest(Y~X, test="sf.test")
    # NORMALITY TRANSFORMS BY GROUPS
      summary(powerTransform(Y ~ X, family="bcPower"))
      summary(powerTransform(Y ~ X, family="yjPower"))
    # HOMOCEDASTICITY ASSUMPTION
      Xfact<- factor(X)
      leveneTest(Y,Xfact, center = "median")
      bartlett.test(Y,Xfact, center = "median")
    # INDEPENDENCE ASSUMPTION
      chisq.test(X,MEMORY1)

#3.5) (NDATA1 = NDATAqualitative) & MEMORY
      library(Rcmdr)
      attach(clusterMemory1)
      X <- NDATA1
      Y <- MEMORY
    # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
      normalityTest(Y~X, test="ad.test")
      normalityTest(Y~X, test="lillie.test")
      normalityTest(Y~X, test="cvm.test")
      normalityTest(Y~X, test="pearson.test")
      normalityTest(Y~X, test="shapiro.test")
```

```
            normalityTest(Y~X, test="sf.test")
        # NORMALITY TRANSFORMS BY GROUPS
            summary(powerTransform(Y ~ X, family="bcPower"))
            summary(powerTransform(Y ~ X, family="yjPower"))
        # HOMOCEDASTICITY ASSUMPTION
            Xfact<- factor(X)
            leveneTest(Y,Xfact, center = "median")
            bartlett.test(Y,Xfact, center = "median")
        # INDEPENDENCE ASSUMPTION
            chisq.test(X,MEMORY1)
    #3.6) (OS_METHOD = OPERATING.SYSTEM , CLUSTER.METHODS) & MEMORY
            library(Rcmdr)
            attach(clusterMemory1)
            X <- OS_METHOD
            Y <- MEMORY
        # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
            normalityTest(Y~X, test="ad.test")
            normalityTest(Y~X, test="lillie.test")
            normalityTest(Y~X, test="cvm.test")
            normalityTest(Y~X, test="pearson.test")
            normalityTest(Y~X, test="shapiro.test")
            normalityTest(Y~X, test="sf.test")
        # NORMALITY TRANSFORMS BY GROUPS
            summary(powerTransform(Y ~ X, family="bcPower"))
            summary(powerTransform(Y ~ X, family="yjPower"))
        # HOMOCEDASTICITY ASSUMPTION
            Xfact<- factor(X)
            leveneTest(Y,Xfact, center = "median")
            bartlett.test(Y,Xfact, center = "median")
        # INDEPENDENCE ASSUMPTION
            chisq.test(X,MEMORY1)
    #3.7) (NDATA2 = NDATAqualitative , ASSOCIATION.RULES.METHODS)  &
MEMORY
            library(Rcmdr)
            attach(clusterMemory1)
            X <- NDATA2
            Y <- MEMORY
        # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
            normalityTest(Y~X, test="ad.test")
            normalityTest(Y~X, test="lillie.test")
            normalityTest(Y~X, test="cvm.test")
            normalityTest(Y~X, test="pearson.test")
            normalityTest(Y~X, test="shapiro.test")
            normalityTest(Y~X, test="sf.test")
        # NORMALITY TRANSFORMS BY GROUPS
            summary(powerTransform(Y ~ X, family="bcPower"))
            summary(powerTransform(Y ~ X, family="yjPower"))
        # HOMOCEDASTICITY ASSUMPTION
            Xfact<- factor(X)
            leveneTest(Y,Xfact, center = "median")
            bartlett.test(Y,Xfact, center = "median")
        # INDEPENDENCE ASSUMPTION
            chisq.test(X,MEMORY1)
    #3.8 (COLUMNS1 = COLUMNSqualitative) & MEMORY
```

```
                library(Rcmdr)
                attach(clusterMemory1)
                X <- COLUMNS1
                Y <- MEMORY
            # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
                normalityTest(Y~X, test="ad.test")
                normalityTest(Y~X, test="lillie.test")
                normalityTest(Y~X, test="cvm.test")
                normalityTest(Y~X, test="pearson.test")
                normalityTest(Y~X, test="shapiro.test")
                normalityTest(Y~X, test="sf.test")
            # NORMALITY TRANSFORMS BY GROUPS
                summary(powerTransform(Y ~ X, family="bcPower"))
                summary(powerTransform(Y ~ X, family="yjPower"))
            # HOMOCEDASTICITY ASSUMPTION
                Xfact<- factor(X)
                leveneTest(Y,Xfact, center = "median")
                bartlett.test(Y,Xfact, center = "median")
            # INDEPENDENCE ASSUMPTION
                chisq.test(X,MEMORY1)
        #3.9) (ROWS1 = ROWSqualitative) & MEMORY
                library(Rcmdr)
                attach(clusterMemory1)
                X <- ROWS1
                Y <- MEMORY
            # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
                normalityTest(Y~X, test="ad.test")
                normalityTest(Y~X, test="lillie.test")
                normalityTest(Y~X, test="cvm.test")
                normalityTest(Y~X, test="pearson.test")
                normalityTest(Y~X, test="shapiro.test")
                normalityTest(Y~X, test="sf.test")
            # NORMALITY TRANSFORMS BY GROUPS
                summary(powerTransform(Y ~ X, family="bcPower"))
                summary(powerTransform(Y ~ X, family="yjPower"))
            # HOMOCEDASTICITY ASSUMPTION
                Xfact<- factor(X)
                leveneTest(Y,Xfact, center = "median")
                bartlett.test(Y,Xfact, center = "median")
            # INDEPENDENCE ASSUMPTION
                chisq.test(X,MEMORY1)
        #3.10) (COLUMNS2 = COLUMNSqualitative ,ASSOCIATION.RULES.METHODS)
& MEMORY
                library(Rcmdr)
                attach(clusterMemory1)
                X <- COLUMNS2
                Y <- MEMORY
            # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
                normalityTest(Y~X, test="ad.test")
                normalityTest(Y~X, test="lillie.test")
                normalityTest(Y~X, test="cvm.test")
                normalityTest(Y~X, test="pearson.test")
                normalityTest(Y~X, test="shapiro.test")
                normalityTest(Y~X, test="sf.test")
```

```
            # NORMALITY TRANSFORMS BY GROUPS
                summary(powerTransform(Y ~ X, family="bcPower"))
                summary(powerTransform(Y ~ X, family="yjPower"))
            # HOMOCEDASTICITY ASSUMPTION
                Xfact<- factor(X)
                leveneTest(Y,Xfact, center = "median")
                bartlett.test(Y,Xfact, center = "median")
            # INDEPENDENCE ASSUMPTION
                chisq.test(X,MEMORY1)
        #3.11) (ROWS2 = ROWSqualitative , ASSOCIATION.RULES.METHODS) &
MEMORY
                library(Rcmdr)
                attach(clusterMemory1)
                X <- ROWS2
                Y <- MEMORY
            # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
                normalityTest(Y~X, test="ad.test")
                normalityTest(Y~X, test="lillie.test")
                normalityTest(Y~X, test="cvm.test")
                normalityTest(Y~X, test="pearson.test")
                normalityTest(Y~X, test="shapiro.test")
                normalityTest(Y~X, test="sf.test")
            # NORMALITY TRANSFORMS BY GROUPS
                summary(powerTransform(Y ~ X, family="bcPower"))
                summary(powerTransform(Y ~ X, family="yjPower"))
            # HOMOCEDASTICITY ASSUMPTION
                Xfact<- factor(X)
                leveneTest(Y,Xfact, center = "median")
                bartlett.test(Y,Xfact, center = "median")
            # INDEPENDENCE ASSUMPTION
                chisq.test(X,MEMORY1)
    #4) ==== Hypothesis Test
        clusterMemory1 <-
read.table("Clustering/MEMORY/clusterMemory1.csv",header=TRUE,
sep=",",dec=".")
        library(Rfit) # non Parametric ANOVA
        names(clusterMemory1)# View variables names
        #4.1) DataBase Characteristics MEMORY
            # NO HYPOTESIS TEST
        #4.2) NDATA -> MEMORY
            # NO HYPOTESIS TEST BECAUSE NDATA IS NOT A FACTOR
        #4.3) OPERATING.SYSTEM -> MEMORY
            attach(clusterMemory1)
            X<-OPERATING.SYSTEM
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(clusterMemory1, oneway.rfit(Y,X)) # non
Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        #4.4) CLUSTER.METHODS -> MEMORY
```

```
            attach(clusterMemory1)
            X<-CLUSTER.METHODS
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(clusterMemory1, oneway.rfit(Y,X)) # non
Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        #4.5) NDATA1 -> MEMORY
            attach(clusterMemory1)
            X<-NDATA1
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(clusterMemory1, oneway.rfit(Y,X)) # non
Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        #4.6) (OS_METHOD = OPERATING.SYSTEM , CLUSTER.METHODS)  -> MEMORY
            attach(clusterMemory1)
            X<-OS_METHOD
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(clusterMemory1, oneway.rfit(Y,X)) # non
Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        #4.7) (NDATA2 = NDATAqualitative , CLUSTER.METHODS)  -> MEMORY
            attach(clusterMemory1)
            X<-NDATA2
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(clusterMemory1, oneway.rfit(Y,X)) # non
Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        #4.8 (COLUMNS1=COLUMNSqualitative) -> MEMORY
            attach(clusterMemory1)
            X<-COLUMNS1
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
```

```
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(clusterMemory1, oneway.rfit(Y,X)) # non
Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        #4.9) (ROWS1=ROWSqualitative) -> MEMORY
            attach(clusterMemory1)
            X<-ROWS1
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(clusterMemory1, oneway.rfit(Y,X)) # non
Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        #4.10) (COLUMNS2 = COLUMNSqualitative ,CLUSTER.METHODS) -> MEMORY
            attach(clusterMemory1)
            X<-COLUMNS2
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(clusterMemory1, oneway.rfit(Y,X)) # non
Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        #4.11) (ROWS2 = ROWSqualitative ,CLUSTER.METHODS)-> MEMORY
            attach(clusterMemory1)
            X<-ROWS2
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(clusterMemory1, oneway.rfit(Y,X)) # non
Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
```

11.8.3 Para el Análisis de los datos de las técnicas de reglas de asociación

```
#================================================#
    # File   : Clustering.R
    # Created: july 2017
    # Updated: july 2019
    # Content: Association Rules memory analysis
    # Author : Rubén Pazmiño-Maji
    #================================================#
    install.packages("Rcmdr")
    library(Rcmdr)
    #For Rcmdr  associationRulesMemory1 <-
read.table("AssociationRules/MEMORY/associationRulesMemory1.csv",header=T
RUE, sep=",",dec=",")
#======================================= M E M O R Y
=======================================#
# Read Clustering memory databases
    associationRulesMemory1 <-
read.table("AssociationRules/MEMORY/associationRulesMemory1.csv",header=T
RUE, sep=",",dec=".")
    head (associationRulesMemory1)
    #1) ==== Descriptive Statistics MEMORY ====
        #1.1) DataBase Characteristics MEMORY (ALL VARIABLES)
            summary(associationRulesMemory1)
        #1.2) NDATA & MEMORY
            numSummary(associationRulesMemory1[,"MEMORY", drop=FALSE],
           groups=associationRulesMemory1$NDATA,
           statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.3) OPERATING.SYSTEM & MEMORY
            numSummary(associationRulesMemory1[,"MEMORY", drop=FALSE],
            groups=associationRulesMemory1$OPERATING.SYSTEM,
            statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.4) ASSOCIATION.RULES.METHODS & MEMORY
            numSummary(associationRulesMemory1[,"MEMORY", drop=FALSE],
            groups=associationRulesMemory1$ASSOCIATION.RULES.METHODS,
            statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.5) (NDATA1 = NDATAqualitative) & MEMORY
            numSummary(associationRulesMemory1[,"MEMORY", drop=FALSE],
           groups=associationRulesMemory1$NDATA1,
           statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.6) (OS_METHOD = OPERATING.SYSTEM , ASSOCIATION.RULES.METHODS)
& MEMORY
            numSummary(associationRulesMemory1[,"MEMORY", drop=FALSE],
           groups=associationRulesMemory1$OS_METHOD,
           statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.7) (NDATA2 = NDATAqualitative , ASSOCIATION.RULES.METHODS)  &
MEMORY
            numSummary(associationRulesMemory1[,"MEMORY", drop=FALSE],
           groups=associationRulesMemory1$NDATA2,
```

```
            statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.8 (COLUMNS1 = COLUMNSqualitative) & MEMORY
             numSummary(associationRulesMemory1[,"MEMORY", drop=FALSE],
            groups=associationRulesMemory1$COLUMNS1,
            statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.9) (ROWS1 = ROWSqualitative) & MEMORY
             numSummary(associationRulesMemory1[,"MEMORY", drop=FALSE],
            groups=associationRulesMemory1$ROWS1,
            statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.10)(COLUMNS2 = COLUMNSqualitative ,ASSOCIATION.RULES.METHODS)
& MEMORY
             numSummary(associationRulesMemory1[,"MEMORY", drop=FALSE],
            groups=associationRulesMemory1$COLUMNS2,
            statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))
        #1.11)(ROWS2 = ROWSqualitative , ASSOCIATION.RULES.METHODS) &
MEMORY
             numSummary(associationRulesMemory1[,"MEMORY", drop=FALSE],
             groups=associationRulesMemory1$ROWS2,
            statistics=c("mean", "sd", "IQR", "quantiles","cv",
"skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1))

    #2) ==== Descriptive Graphics MEMORY ====
        library(ggplot2)
        associationRulesMemory1 <-
read.table("AssociationRules/MEMORY/associationRulesMemory1.csv",header=T
RUE, sep=",",dec=".")
        head (associationRulesMemory1)
        #2.1) DataBase Characteristics MEMORY
        # No graphics are important
        #2.2) NDATA & MEMORY
            # No graphics, many data
        #2.3) OPERATING.SYSTEM & MEMORY
            attach(associationRulesMemory1)
            X <- OPERATING.SYSTEM
            Y <- MEMORY
            Xlabel  <- "OPERATING_SYSTEM"
            Ylabel  <- "MEMORY"
            df <- data.frame( Xlabel=factor(X), Ylabel=Y )
            G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
            G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
            T  <- theme(legend.position="top")
            G1 + geom_histogram( fill="white",position="dodge") + T
            G2 + geom_boxplot()+ T
            G2 + geom_violin() + T
        #2.4) ASSOCIATION.RULES.METHODS & MEMORY
            attach(associationRulesMemory1)
            X <- ASSOCIATION.RULES.METHODS
            Y <- MEMORY
            Xlabel  <- "ASSOCIATION.RULES.METHODS"
            Ylabel  <- "MEMORY"
```

```
            df <- data.frame( Xlabel=factor(X), Ylabel=Y )
            G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
            G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
            T  <- theme(legend.position="top")
            G1 + geom_histogram( fill="white",position="dodge") + T
            G2 + geom_boxplot()+ T
            G2 + geom_violin() + T
        #2.5) (NDATA1=NDATAqualitative) & MEMORY
            attach(associationRulesMemory1)
            X <- NDATA1
            Y <- MEMORY
            Xlabel  <- "DATA_NUMBER"
            Ylabel  <- "MEMORY"
            df <- data.frame( Xlabel=factor(X), Ylabel=Y )
            G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
            G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
            T  <- theme(legend.position="top")
            G1 + geom_histogram( fill="white",position="dodge") + T
            G2 + geom_boxplot()+ T
            G2 + geom_violin() + T
        #2.6) (OS_METHOD = OPERATING.SYSTEM , ASSOCIATION.RULES.METHODS)
& MEMORY
            attach(associationRulesMemory1)
            X <- OS_METHOD
            Y <- MEMORY
            Xlabel  <- "OPERATING.SYSTEM&ASSOCIATION.RULES.METHODS"
            Ylabel  <- "MEMORY"
            df <- data.frame( Xlabel=factor(X), Ylabel=Y )
            G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
            G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
            T  <- theme(legend.position="top")
            G1 + geom_histogram( fill="white",position="dodge") + T
            G2 + geom_boxplot()+ T
            G2 + geom_violin() + T
        #2.7) (NDATA2 = NDATAqualitative, CLUSTER_METHODS)  & MEMORY
            attach(associationRulesMemory1)
            X <- NDATA2
            Y <- MEMORY
            Xlabel  <- "NDATA&ASSOCIATION.RULES.METHODS"
            Ylabel  <- "MEMORY"
            df <- data.frame( Xlabel=factor(X), Ylabel=Y )
            G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
            G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
            T  <- theme(legend.position="top")
            G1 + geom_histogram( fill="white",position="dodge") + T
            G2 + geom_boxplot()+ T
            G2 + geom_violin() + T
        #2.8 (COLUMNS1=COLUMNSqualitative) & MEMORY
            attach(associationRulesMemory1)
            X <- COLUMNS1
            Y <- MEMORY
            Xlabel  <- "COLUMNS_NUMBER"
            Ylabel  <- "MEMORY"
            df <- data.frame( Xlabel=factor(X), Ylabel=Y )
```

```
            G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
            G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
            T  <- theme(legend.position="top")
            G1 + geom_histogram( fill="white",position="dodge") + T
            G2 + geom_boxplot()+ T
            G2 + geom_violin() + T
        #2.9)(ROWS1 = ROWSqualitative) & MEMORY
            attach(associationRulesMemory1)
            X <- ROWS1
            Y <- MEMORY
            Xlabel  <- "ROWS_NUMBER"
            Ylabel  <- "MEMORY"
            df <- data.frame( Xlabel=factor(X), Ylabel=Y )
            G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
            G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
            T  <- theme(legend.position="top")
            G1 + geom_histogram( fill="white",position="dodge") + T
            G2 + geom_boxplot()+ T
            G2 + geom_violin() + T
        #2.10)(COLUMNS2 = COLUMNSqualitative ,ASSOCIATION.RULES.METHODS)&
MEMORY
            attach(associationRulesMemory1)
            X <- COLUMNS2
            Y <- MEMORY
            Xlabel  <- "COLUMNS_NUMBER"
            Ylabel  <- "MEMORY"
            df <- data.frame( Xlabel=factor(X), Ylabel=Y )
            G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
            G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
            T  <- theme(legend.position="top")
            G1 + geom_histogram( fill="white",position="dodge") + T
            G2 + geom_boxplot()+ T
            G2 + geom_violin() + T
        #2.11)(ROWS2 = ROWSqualitative ,ASSOCIATION.RULES.METHODS)&
MEMORY
            attach(associationRulesMemory1)
            X <- ROWS2
            Y <- MEMORY
            Xlabel  <- "ROWSNUMBER&CLUSTER_METHODS"
            Ylabel  <- "MEMORY"
            df <- data.frame( Xlabel=factor(X), Ylabel=Y )
            G1 <- ggplot(df, aes(Ylabel, color=Xlabel))
            G2 <- ggplot(df, aes(Xlabel,Ylabel, color=Xlabel))
            T  <- theme(legend.position="top")
            G1 + geom_histogram( fill="white",position="dodge") + T
            G2 + geom_boxplot()+ T
            G2 + geom_violin() + T
    #3) ==== Assumptions (NORMALITY, HOMOCEDASTICITY, INDEPENDENCE)
MEMORY ====
            library(nortest)  #Test de Anderson-Darling y Test de
Lilliefors (Kolmogorov-Smirnov)
            library(moments) #Test de Jarque Bera#
            library(Johnson) #JOHNSON TRANSFORM
            library(car)#Box-Cox TRANSFORM
```

```
        associationRulesMemory1 <-
read.table("AssociationRules/MEMORY/associationRulesMemory1.csv",header=T
RUE, sep=",",dec=".")
        #3.1) DataBase Characteristics MEMORY (ALL VARIABLES)
            #ASSUMPTION OF GLOBAL NORMALITY
                attach(associationRulesMemory1)
                Y <- MEMORY
                #Start Normality graphics
                hist(Y,probability=TRUE, main="Histogram of MEMORY in
Association Rule Minnig data",xlab="Approximately normally distributed
data")
                lines(density(Y),col=2)
                qqnorm(Y,main="QQ plot of MEMORY data",pch=19)
                qqline(Y)

                #Start Normality TESTS (FIVE NORMALITY TESTS)
                ad.test(Y) #Test de Anderson-Darling
                lillie.test(Y) #Test de Lilliefors (Kolmogorov-Smirnov)
                cvm.test(Y) #Cramer-von Mises normality test
                pearson.test(Y) #Pearson chi-square normality test
                jb.norm.test(Y)# #Test de Jarque Bera#

            #GLOBAL NORMALITY TRANSFORMS (FIVE NORMALITY TESTS, FOUR+TWO
TRANSFORMS)
                ad.test(log(Y)) #LOG TRANSFORM
                ad.test(1/Y) #EXP TRANSFORM
                ad.test(RE.Johnson(c(Y))$transformed) #JOHNSON TRANSFORM
                ad.test(boxCoxVariable(c(Y))) #Box-Cox TRANSFORM

                lillie.test(log(Y)) #LOG TRANSFORM
                lillie.test(1/Y) #EXP TRANSFORM
                lillie.test(RE.Johnson(c(Y))$transformed) #JOHNSON
TRANSFORM
                lillie.test(boxCoxVariable(c(Y))) #Box-Cox TRANSFORM

                cvm.test(log(Y)) #LOG TRANSFORM
                cvm.test(1/Y) #EXP TRANSFORM
                cvm.test(RE.Johnson(c(Y))$transformed) #JOHNSON TRANSFORM
                cvm.test(boxCoxVariable(c(Y))) #Box-Cox TRANSFORM

                pearson.test(log(Y)) #LOG TRANSFORM
                pearson.test(1/Y) #EXP TRANSFORM
                pearson.test(RE.Johnson(c(Y))$transformed) #JOHNSON
TRANSFORM
                pearson.test(boxCoxVariable(c(Y))) #Box-Cox TRANSFORM

                jb.norm.test(log(Y)) #LOG TRANSFORM
                jb.norm.test(1/Y) #EXP TRANSFORM
                jb.norm.test(RE.Johnson(c(Y))$transformed) #JOHNSON
TRANSFORM
                jb.norm.test(boxCoxVariable(c(Y))) #Box-Cox TRANSFORM
```

```
                summary(powerTransform(Y ~ 1,
data=associationRulesMemory1, family="bcPower")) #Box-Cox Power
Transformation to Normality
                summary(powerTransform(Y ~ 1,
data=associationRulesMemory1, family="yjPower"))#Yeo-Johnson Power
Transformation to Normality
        #3.2) NDATA & MEMORY
              # Many data to do normality test
        #3.3) OPERATING.SYSTEM & MEMORY
                library(Rcmdr)
                attach(associationRulesMemory1)
                X <- OPERATING.SYSTEM
                Y <- MEMORY
            # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
                normalityTest(Y~X, test="ad.test")
                normalityTest(Y~X, test="lillie.test")
                normalityTest(Y~X, test="cvm.test")
                normalityTest(Y~X, test="pearson.test")
                normalityTest(Y~X, test="shapiro.test")
                normalityTest(Y~X, test="sf.test")
            # NORMALITY TRANSFORMS BY GROUPS
                summary(powerTransform(Y ~ X, family="bcPower"))
                summary(powerTransform(Y ~ X, family="yjPower"))
            # HOMOCEDASTICITY ASSUMPTION
                Xfact<- factor(X)
                leveneTest(Y,Xfact, center = "median")
                bartlett.test(Y,Xfact, center = "median")
            # INDEPENDENCE ASSUMPTION
                chisq.test(X,MEMORY1)
                # install.packages("summarytools", dependencies = FALSE)
                # library("summarytools")
                # frequencyMEMORY<-freq(MEMORY)
                # frequencyMEMORY[,1]

        #3.4) ASSOCIATION.RULES.METHODS & MEMORY
                library(Rcmdr)
                attach(associationRulesMemory1)
                X <- ASSOCIATION.RULES.METHODS
                Y <- MEMORY
            # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
                normalityTest(Y~X, test="ad.test")
                normalityTest(Y~X, test="lillie.test")
                normalityTest(Y~X, test="cvm.test")
                normalityTest(Y~X, test="pearson.test")
                normalityTest(Y~X, test="shapiro.test")
                normalityTest(Y~X, test="sf.test")
            # NORMALITY TRANSFORMS BY GROUPS
                summary(powerTransform(Y ~ X, family="bcPower"))
                summary(powerTransform(Y ~ X, family="yjPower"))
            # HOMOCEDASTICITY ASSUMPTION
                Xfact<- factor(X)
                leveneTest(Y,Xfact, center = "median")
                bartlett.test(Y,Xfact, center = "median")
            # INDEPENDENCE ASSUMPTION
```

```
                chisq.test(X,MEMORY1)

        #3.5) (NDATA1 = NDATAqualitative) & MEMORY
                library(Rcmdr)
                attach(associationRulesMemory1)
                X <- NDATA1
                Y <- MEMORY
            # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
                normalityTest(Y~X, test="ad.test")
                normalityTest(Y~X, test="lillie.test")
                normalityTest(Y~X, test="cvm.test")
                normalityTest(Y~X, test="pearson.test")
                normalityTest(Y~X, test="shapiro.test")
                normalityTest(Y~X, test="sf.test")
            # NORMALITY TRANSFORMS BY GROUPS
                summary(powerTransform(Y ~ X, family="bcPower"))
                summary(powerTransform(Y ~ X, family="yjPower"))
            # HOMOCEDASTICITY ASSUMPTION
                Xfact<- factor(X)
                leveneTest(Y,Xfact, center = "median")
                bartlett.test(Y,Xfact, center = "median")
            # INDEPENDENCE ASSUMPTION
                chisq.test(X,MEMORY1)
        #3.6) (OS_METHOD = OPERATING.SYSTEM , ASSOCIATION.RULES.METHODS)
& MEMORY
                library(Rcmdr)
                attach(associationRulesMemory1)
                X <- OS_METHOD
                Y <- MEMORY
            # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
                normalityTest(Y~X, test="ad.test")
                normalityTest(Y~X, test="lillie.test")
                normalityTest(Y~X, test="cvm.test")
                normalityTest(Y~X, test="pearson.test")
                normalityTest(Y~X, test="shapiro.test")
                normalityTest(Y~X, test="sf.test")
            # NORMALITY TRANSFORMS BY GROUPS
                summary(powerTransform(Y ~ X, family="bcPower"))
                summary(powerTransform(Y ~ X, family="yjPower"))
            # HOMOCEDASTICITY ASSUMPTION
                Xfact<- factor(X)
                leveneTest(Y,Xfact, center = "median")
                bartlett.test(Y,Xfact, center = "median")
            # INDEPENDENCE ASSUMPTION
                chisq.test(X,MEMORY1)
        #3.7) (NDATA2 = NDATAqualitative , ASSOCIATION.RULES.METHODS)  &
MEMORY
                library(Rcmdr)
                attach(associationRulesMemory1)
                X <- NDATA2
                Y <- MEMORY
            # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
                normalityTest(Y~X, test="ad.test")
                normalityTest(Y~X, test="lillie.test")
```

```
        normalityTest(Y~X, test="cvm.test")
        normalityTest(Y~X, test="pearson.test")
        normalityTest(Y~X, test="shapiro.test")
        normalityTest(Y~X, test="sf.test")
    # NORMALITY TRANSFORMS BY GROUPS
        summary(powerTransform(Y ~ X, family="bcPower"))
        summary(powerTransform(Y ~ X, family="yjPower"))
    # HOMOCEDASTICITY ASSUMPTION
        Xfact<- factor(X)
        leveneTest(Y,Xfact, center = "median")
        bartlett.test(Y,Xfact, center = "median")
    # INDEPENDENCE ASSUMPTION
        chisq.test(X,MEMORY1)
#3.8 (COLUMNS1 = COLUMNSqualitative) & MEMORY
        library(Rcmdr)
        attach(associationRulesMemory1)
        X <- COLUMNS1
        Y <- MEMORY
    # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
        normalityTest(Y~X, test="ad.test")
        normalityTest(Y~X, test="lillie.test")
        normalityTest(Y~X, test="cvm.test")
        normalityTest(Y~X, test="pearson.test")
        normalityTest(Y~X, test="shapiro.test")
        normalityTest(Y~X, test="sf.test")
    # NORMALITY TRANSFORMS BY GROUPS
        summary(powerTransform(Y ~ X, family="bcPower"))
        summary(powerTransform(Y ~ X, family="yjPower"))
    # HOMOCEDASTICITY ASSUMPTION
        Xfact<- factor(X)
        leveneTest(Y,Xfact, center = "median")
        bartlett.test(Y,Xfact, center = "median")
    # INDEPENDENCE ASSUMPTION
        chisq.test(X,MEMORY1)
#3.9) (ROWS1 = ROWSqualitative) & MEMORY
        library(Rcmdr)
        attach(associationRulesMemory1)
        X <- ROWS1
        Y <- MEMORY
    # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
        normalityTest(Y~X, test="ad.test")
        normalityTest(Y~X, test="lillie.test")
        normalityTest(Y~X, test="cvm.test")
        normalityTest(Y~X, test="pearson.test")
        normalityTest(Y~X, test="shapiro.test")
        normalityTest(Y~X, test="sf.test")
    # NORMALITY TRANSFORMS BY GROUPS
        summary(powerTransform(Y ~ X, family="bcPower"))
        summary(powerTransform(Y ~ X, family="yjPower"))
    # HOMOCEDASTICITY ASSUMPTION
        Xfact<- factor(X)
        leveneTest(Y,Xfact, center = "median")
        bartlett.test(Y,Xfact, center = "median")
    # INDEPENDENCE ASSUMPTION
```

```
                chisq.test(X,MEMORY1)
        #3.10) (COLUMNS2 = COLUMNSqualitative ,ASSOCIATION.RULES.METHODS)
& MEMORY
                library(Rcmdr)
                attach(associationRulesMemory1)
                X <- COLUMNS2
                Y <- MEMORY
            # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
                normalityTest(Y~X, test="ad.test")
                normalityTest(Y~X, test="lillie.test")
                normalityTest(Y~X, test="cvm.test")
                normalityTest(Y~X, test="pearson.test")
                normalityTest(Y~X, test="shapiro.test")
                normalityTest(Y~X, test="sf.test")
            # NORMALITY TRANSFORMS BY GROUPS
                summary(powerTransform(Y ~ X, family="bcPower"))
                summary(powerTransform(Y ~ X, family="yjPower"))
            # HOMOCEDASTICITY ASSUMPTION
                Xfact<- factor(X)
                leveneTest(Y,Xfact, center = "median")
                bartlett.test(Y,Xfact, center = "median")
            # INDEPENDENCE ASSUMPTION
                chisq.test(X,MEMORY1)
        #3.11) (ROWS2 = ROWSqualitative , ASSOCIATION.RULES.METHODS) &
MEMORY
                library(Rcmdr)
                attach(associationRulesMemory1)
                X <- ROWS2
                Y <- MEMORY
            # NORMALITY ASSUMPTION BY GROUPS (SIX NORMALITY TESTS)
                normalityTest(Y~X, test="ad.test")
                normalityTest(Y~X, test="lillie.test")
                normalityTest(Y~X, test="cvm.test")
                normalityTest(Y~X, test="pearson.test")
                normalityTest(Y~X, test="shapiro.test")
                normalityTest(Y~X, test="sf.test")
            # NORMALITY TRANSFORMS BY GROUPS
                summary(powerTransform(Y ~ X, family="bcPower"))
                summary(powerTransform(Y ~ X, family="yjPower"))
            # HOMOCEDASTICITY ASSUMPTION
                Xfact<- factor(X)
                leveneTest(Y,Xfact, center = "median")
                bartlett.test(Y,Xfact, center = "median")
            # INDEPENDENCE ASSUMPTION
                chisq.test(X,MEMORY1)

    #4) ==== Hypothesis Test
        library(Rfit) # non Parametric ANOVA
        associationRulesMemory1 <-
read.table("AssociationRules/MEMORY/associationRulesMemory1.csv",header=T
RUE, sep=",",dec=".")
        head (associationRulesMemory1)
        names(associationRulesMemory1)# View variables names
        # 4.1   DataBase Characteristics MEMORY (ALL VARIABLES)
```

```
            # No hipotesys test
        # 4.2   NDATA -> MEMORY
            # NO HYPOTESIS TEST BECAUSE NDATA IS NOT A FACTOR
        # 4.3   OPERATING.SYSTEM -> MEMORY
            attach(associationRulesMemory1)
            X<-OPERATING.SYSTEM
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(associationRulesMemory1, oneway.rfit(Y,X)) #
non Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        # 4.4   ASSOCIATION.RULES.METHODS -> MEMORY
            attach(associationRulesMemory1)
            X<-ASSOCIATION.RULES.METHODS
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(associationRulesMemory1, oneway.rfit(Y,X)) #
non Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        # 4.5   (NDATA1 = NDATAqualitative) -> MEMORY
            attach(associationRulesMemory1)
            X<-NDATA1
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(associationRulesMemory1, oneway.rfit(Y,X)) #
non Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        # 4.6(OS_METHOD = OPERATING.SYSTEM , ASSOCIATION.RULES.METHODS) -
> MEMORY
            attach(associationRulesMemory1)
            X<-OS_METHOD
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(associationRulesMemory1, oneway.rfit(Y,X)) #
non Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        # 4.7   (NDATA2 = NDATAqualitative , ASSOCIATION.RULES.METHODS)
-> MEMORY
```

```
            attach(associationRulesMemory1)
            X<-NDATA2
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(associationRulesMemory1, oneway.rfit(Y,X)) #
non Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        # 4.8   (COLUMNS1 = COLUMNSqualitative) -> MEMORY
            attach(associationRulesMemory1)
            X<-COLUMNS1
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(associationRulesMemory1, oneway.rfit(Y,X)) #
non Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        # 4.9(ROWS1 = ROWSqualitative) -> MEMORY
            attach(associationRulesMemory1)
            X<-ROWS1
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(associationRulesMemory1, oneway.rfit(Y,X)) #
non Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        # 4.10(COLUMNS2 = COLUMNSqualitative ,ASSOCIATION.RULES.METHODS)
-> MEMORY
            attach(associationRulesMemory1)
            X<-COLUMNS2
            Y<-MEMORY
            tapply(Y,X,summary)
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(associationRulesMemory1, oneway.rfit(Y,X)) #
non Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
        # 4.11(ROWS2 = ROWSqualitative ,ASSOCIATION.RULES.METHODS)->
MEMORY
            attach(associationRulesMemory1)
            X<-ROWS2
            Y<-MEMORY
            tapply(Y,X,summary)
```

```
            kruskal.test(Y~X)
            pairwise.wilcox.test(Y,X,p.adj='bonferroni',exact=F)
            pairwise.wilcox.test(Y,X,p.adj='holm',exact=F)
            nPanova<- with(associationRulesMemory1, oneway.rfit(Y,X)) #
non Parametric ANOVA
            nPanova
            summary(nPanova, method = "tukey")
```

11.9 Apéndice I.- Bases de Datos

A continuación, se presentan las bases de datos reducidas utilizadas en el análisis de datos y consideradas en los capítulos en los cuales se realiza la comparación de las técnicas clúster y reglas de asociación realizados.

11.9.1 Para el análisis de los datos de las técnicas clúster

A continuación, se muestra una parte de las bases de datos utilizadas para la comparación de la complejidad temporal y espacial de las técnicas clúster de LA y ASI (Tabla 11.4).

Tabla 11.4.- Ejemplo de bases de datos parciales de las técnicas clúster de LA y ASI.

COLUMNS1	ROWS1	NDATA1	CLUSTER METHODS	COLUMNS2	MEMORY1
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)

20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)

20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)

20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_diana,[1	20000)_dendro_diana,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_dendro_variables,[1	20000)_dendro_variables,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hclust_vector,[1	20000)_hclust_vector,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_hrarchy,[1	20000)_hrarchy,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)
20),[1	200),[1	20000),[1	200)_simlrty,[1	20000)_simlrty,[103,0	145,0)

11.9.2 Para el análisis de los datos de las técnicas de reglas de asociación

A continuación, se muestra una parte de la base de datos utilizada para la comparación de la complejidad temporal y espacial de las técnicas de reglas de asociación de LA y ASI (Tabla 11.5).

Tabla 11.5.- Ejemplo de bases de datos parciales de las técnicas de reglas de asociación de LA y ASI.

COLUMNS1	ROWS1	NDATA1	ASSOCIATI RULES	COLUMNS2	MEMORY1
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)

100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)

100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)

100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)

100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)

100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[148,2	156,8)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[136	148,2)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]

100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]

100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)

100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]

100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_apriori,[1	20000)_met_apriori,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_ASI,[1	20000)_met_ASI,[163,7	241]
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_eclat,[1	20000)_met_eclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)
100],[1	200),[1	20000),[80	200)_met_weclat,[1	20000)_met_weclat,[156,8	163,7)

11.10 Apéndice J.- Indicadores educativos

A continuación, se muestran los indicadores educativos universitarios utilizados en el caso de estudio.

Tabla 11.6.- Indicadores utilizados en el caso de estudio

ORDEN	INDICADORES	CÓDIGO
1	II1_ESTADO ACTUAL PROSPECTIVO	II1ESACP
2	II2_PROYECTOS PROGRAMAS VINCULACIÓN	II2PRPRV
3	II3_PERFIL PROFESIONAL	II3PERPR
4	II4_PERFIL DE EGRESO	II4PEREG
5	II5_ESTRUCTURA CURRICULAR	II5ESTCU
6	II6_PLAN ESTUDIOS	II6PLAES
7	II7_PROGRAMAS-ASIGNATURAS	II7PROAS
8	II8_PRÁCTICAS RELACIÓN ASIGNATURAS	II8PRARA
9	II9_AFINIDAD FORMACIÓN POSTGRADO	II9AFIFP
10	II10_ACTUALIZACIÓN CIENTÍFICA PEDAGÓGICA	II10ACCP
11	II11_TITULARIDAD	II11TTLR
12	II12_PROFESOR TC-MT-TP	II12PCMP
13	II13_ESTUDIANTE PROFESOR	II13ESPR
14	II14_DISTRIBUCIÓN HORARIA	II14DIHO
15	II15_PRODUCCIÓN ACADÉMICA CIENTÍFICA	II15PRAC
16	II16_PRODUCCIÓN GENERAL	II16PRGE
17	II17_LIBROS O CAPÍTULOS DE LIBROS	II17LOCL
18	II18_PONENCIAS	II18PNNC
19	II19_DIRECCIÓN COORDINACIÓN ACADÉMICA	II19DICA
20	II20_EVALUACIÓN DESEMPEÑO DOCENTE	II20EVDD
21	II21_SEGUIMIENTO SÍLABO	II21SESI
22	II22_SEGUIMIENTO PROCESO DE TITULACIÓN	II22SPDT
23	II23_SEGUIMIENTO GRADUADOS	II23SEGR
24	II24_SEGUIMIENTO PRÁCTICAS PRE-PROFESIONALES	II24SPPR
25	II25_CALIDAD DE INFORMACIÓN	II25CAIN
26	II26_BIBLIOGRAFÍA BÁSICA	II26BIBA
27	II27_CALIDAD BIBLIOGRÁFICA	II27CABI
28	II28_FUNCIONALIDAD	II28FNCN
29	II29_EQUIPAMIENTO	II29EQPM
30	II30_DISPONIBILIDAD (EQUIPOS, LABORATORIOS)	II30DSPN
31	II31_TUTORÍAS	II31TTRS
32	II32_ACTIVIDADES COMPLEMENTARIAS	II32ACCO
33	II33_ACTIVIDADES VINCULACIÓN CON LA COLECTIVIDAD	II33ACVC

34	II34_BIENESTAR ESTUDIANTIL	II34BIES
35	II35_PARTICIPACIÓN ACREDITACIÓN	II35PAAC
36	II36_TASA RETENCIÓN	II36TARE
37	II37_TASA TITULACIÓN	II37TATI

REFERENCIAS

Se adjuntan las referencias bibliográficas utilizadas en las revisiones sistemáticas en particular y en esta tesis en general

REFERENCIAS

Aaten, A. B., van den Heuvel-Panhuizen, M., y Elia, I. (2016). Kindergartners' perspective taking abilities. *in Proceedings of the Seventh Congress of the European Society for Research in Mathematics Education, University of Rzeszów, Poland*, *Vol. 1822*.

Acevedo, J. C. B. (2006). *Matemáticas avanzadas y estadística para ciencias e ingenierías*. Universidad de Sevilla.

Acioly-Regnier, N., y Regnier, J.-C. (2007). *Analyse cohésitive et interprétations des données dans le champ de l'éducation*.

Agrawal, R., Imieliński, T., y Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 207-216. https://doi.org/10.1145/170035.170072

Aguilar, J., Cordero, J., y Buendía, O. (2017). Specification of the Autonomic Cycles of Learning Analytic Tasks for a Smart Classroom. *Journal of Educational Computing Research*, *56*(6), 866-891.

Aguilar, J., Sánchez, M., Cordero, J., Valdiviezo-Díaz, P., Barba-Guamán, L., y Chamba-Eras, L. (2018). Learning analytics tasks as services in smart classrooms. *Universal Access in the Information Society*, *17*(4), 693-709.

Aguilar, J., Valdiviezo, P., Cordero, J., Riofrio, G., Encalada, E., y Lagos-Ortiz K., del C. J. V.-L. N. A.-M. G. V.-G. R. (2016). A general framework for learning analytic in a smart classroom. *Communications in Computer and Information Science*, *658*, 214-225. https://doi.org/10.1007/978-3-319-48024-4_17

Aguilar, J., Valdiviezo-Díaz, P., y Teran L., T. L. M. A. (2017). Learning analytic in a smart classroom to improve the eEducation. *2017 4th International Conference on*

EDemocracy and EGovernment, ICEDEG 2017. https://doi.org/10.1109/ICEDEG.2017.7962510

Allaire, J. (2012). RStudio: Integrated development environment for R. *Boston, MA*, *770*(394), 165-171.

Almenara Barrios, J., Silva Ayçaguer, L. C., Benavides Rodríguez, A., García Ortega, C., y González Caballero, J. L. (2004). Historia de la bioestadística, la génesis, la normalidad y la crisis. *Revista Española de Salud Pública*, *78*(1), 115-116.

Analítica de aprendizaje. (2021). En *Wikipedia, la enciclopedia libre*. https://es.wikipedia.org/w/index.php?title=Anal%C3%ADtica_de_aprendizaje&oldid=137026001

Anastasiadou, S. D. (2019). Comparison of multivariate pattering methods in group/cluster identification regarding the science of educational research: Implicative Statistical Analysis vs. Lâ€™ Analysee Factorielle des Correspondances. *New Trends and Issues Proceedings on Humanities and Social Sciences*, *6*(1), 238-245.

Anastasiadou, S. D., Anastasiadis, L., Vandikas, I., y Angeletos, T. (2011). Implicative Statistical Analysis and Principal Components Analysis in Recording Students' Attitudes to Electronics and Electrical Construction Subjects. *International Journal of Technology, Knowledge & Society*, *7*(1).

Anastasiadou, S. D., y Karakos, A. S. (2011). The beliefs of electrical and computer engineering students' regarding computer programming. *The International Journal of Technology, Knowledge and Society*, *7*(1), 37-51.

Aporte—Sinónimos y antónimos—WordReference.com. (2021). https://www.wordreference.com/sinonimos/aporte

Arabie, P., Baier, N. D., Critchley, C. F., y Keynes, M. (2006). *Studies in Classification, Data Analysis, and Knowledge Organization*.

Arifin, W. N. (2019). *Introduction to R and RStudio IDE*.

ASALE, R.-, y RAE. (2021). *Aporte | Diccionario de la lengua española*. «Diccionario de la lengua española» - Edición del Tricentenario. https://dle.rae.es/aporte

Axelsen, M., Redmond, P., Heinrich, E., y Henderson, M. (2020). The evolving field of learning analytics research in higher education. *Australasian Journal of Educational Technology*, *36*(2), 1-7.

Bachelard, G. (1993). La formation de l'esprit scientifique (1938). *Paris, Vrin*, *8*, 123.

Bailleul, M. (2000). *Le rôle des variables supplémentaires dans l'analyse statistique implicative. Une recherche sur la professionnalisation des enseignants.*

Bailleul, M., y Gras, R. (1994). L'implication statistique entre variables modales. *Mathématiques et sciences humaines*, *128*, 41-57.

Baker, R., y Inventado, P. (2014). Educational data mining and learning analytics. En *Learning analytics* (pp. 61-75). Springer.

Barragán-Pazmiño, B. M., y Pazmiño-Maji, R. (2018). Literatura Científica sobre Análisis Estadístico Implicativo: Un mapeo sistemático de la década que transcurre. *CIENCIA DIGITAL*, *2*, 16.

Berenson, M. L., Levine, D. M., y Krehbiel, T. C. (2006). *Statistics for administration*. Pearson Educación.

Bernard, J.-M., y Charron, C. (1996). L'analyse implicative bayésienne, une méthode pour l'étude des dépendances orientées. I: données binaires. *Mathématiques et Sciences humaines*, *134*, 5-38.

Blanchard, J., Guillet, F., Gras, R., y Briand, H. (2005). Using information-theoretic measures to assess association rule interestingness. *Fifth IEEE International Conference on Data Mining (ICDM'05)*, 8 pp.

Blanchard, J., Kuntz, P., Guillet, F., y Gras, R. (2003). Implication intensity: From the basic statistical definition to the entropic version. *Statistical data mining and knowledge discovery*, 473-485.

Bodin, A. (1997). Analyse implicative: Modèles sous-jacents à l'analyse implicative et outils complémentaires. *Publications mathématiques et informatique de Rennes*, *3*, 1-23.

Bonneton-Botte, N., Hili, H., De La Haye, F., y Noel, Y. (2015). Drawings of the hand and numerical skills in children of preschool age. *Canadian Journal of Behavioural Science-Revue Canadienne Des Sciences Du Comportement*, *47*(3), 207-215.

Brousseau, G. (2013). *The importance of supplementary variables in a case of an educational research*.

Campbell, D. T., y Stanley, J. C. (2015). *Experimental and quasi-experimental designs for research*. Ravenio Books.

Canto de Gante, Á. G., Sosa González, W. E., Bautista Ortega, J., Escobar Castillo, J., y Santillán Fernández, A. (2020). Escala de Likert: Una alternativa para elaborar e interpretar un instrumento de percepción social. *Revista de La Alta Tecnología y Sociedad*, *12*(1).

Casanovas Muñoz, J. (2016). *Ontology for modelling and understanding educational data and concepts: An application to Learning Analytics for Secondary project*. Universitat Politècnica de Catalunya.

Castor, G., Antonio, V., y Aldo, B. (2013). *Tratamientos de datos con r,statistica y spss* (Diaz de Santos). file:///C:/Users/Personal/Downloads/tratamiento-de-datos-con-r-estatistica-y-spss.pdf

Censos, I. N. de E. y. (2020). *Población y Demografía*. Instituto Nacional de Estadística y Censos. https://www.ecuadorencifras.gob.ec/censo-de-poblacion-y-vivienda/

CEUR-WS.org—CEUR Workshop Proceedings. (2020). http://ceur-ws.org/

Chamba-Eras, L., Labanda-Jaramillo, M., Coronel-Romero, E., Roman-Sanchez, M., y del Mar Perez Sanagustin M., O. X. (2018). Learning analytics in continuing training in higher education. Case study. *CEUR Workshop Proceedings*, *2231*.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., y Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, *4*(5-6), 318-331.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., y Thüs, H. (2013). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, *4*(5-6), 318-331.

Cided. (2019). http://cided.espoch.edu.ec/

Citavi—Reference Management and Knowledge Organization. (2016). Citavi.Com. https://www.citavi.com/en

Conde, M. Á., García-Peñalvo, F. J., Rodríguez-Conde, M. J., Alier, M., Casany, M. J., y Piguillem, J. (2014). An evolving Learning Management System for new educational environments using 2.0 tools. *Interactive learning environments*, *22*(2), 188-204.

Conde, M. Á., y Hernández-García, Á. (2017). Learning analytics: Expanding the frontier. *Proceedings of the 5th International Conference on Technological Ecosystems for Enhancing Multiculturality*, 1-5.

Connaway, L. S. (2015). Retos de la investigación: El camino hacia el compromiso y el progreso. *BiD: textos universitaris de biblioteconomia i documentació*, *35*.

Connaway, L. S., y Radford, M. L. (2016). *Research methods in library and information science*. ABC-CLIO.

Cook, T. D., Campbell, D. T., y Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA.

Correa, J. C., Iral, R., y Rojas, L. (2006). Estudio de potencia de pruebas de homogeneidad de varianza. *Revista colombiana de estadística*, *29*(1), 57-76.

Coutrier, R., Pazmiño Maji, R. A., Conde González, M. Á., y García-Peñalvo, F. J. (2015). *Statistical implicative analysis for educational data sets: 2 analysis with RCHIC*.

Couturier, R., y Almouloud, S. A. (2009). *Historique et fonctionnalités de CHIC*.

Couturier, R., y Gras, R. (2005a). CHIC: traitement de données avec l'analyse implicative. *EGC*, 679-684.

Couturier, R., y Gras, R. (2005b). *CHIC: traitement de données avec l'analyse implicative*. 679-684.

Couturier, R., y Pazmiño-Maji, R. (2016). Use of Statistical Implicative Analysis in Complement of Item Analysis. *International Journal of Information and Education Technology*, *6*(1), 39.

Couturier, R., Pazmiño-Maji, R., García-Peñalvo, F., y Conde-González, M. (2015). *Statistical implicative analysis for educational data sets: 2 analysis with RCHIC*.

CRAN - Package CluMix. (2016). https://cran.r-project.org/web/packages/CluMix/index.html

Crisol-Moya, E., Herrera-Nieves, L., y Montes-Soldado, R. (2020). Educación virtual para todos: Una revisión sistemática. *Education in the knowledge society (EKS)*, *21*, 13.

Daniel Müllner. (2018, junio 7). *CRAN - Paquete fastcluster*. https://cran.r-project.org/web/packages/fastcluster/index.html

David, J., Guillet, F., Briand, H., y Gras, R. (2008). On the use of Implication Intensity for matching ontologies and textual taxonomies. En *Statistical Implicative Analysis* (pp. 227-245). Springer.

de León, C. G. D., y de la Garza, E. L. (2014). Método comparativo. *Métodos y técnicas cualitativas y cuantitativas aplicables a investigación en Ciencias Sociales*, 223-251.

Delacroix, T., y Boubekki, A. (2014). An application of multiple behavior SIA for analyzing data from student exams Applications multiples de l´ ASI pour l´ analyse des données des examens d´ étudiants. *Educação Matemática Pesquisa: Revista do Programa de Estudos Pós-Graduados em Educação Matemática*, *16*(3), 795-812.

Díaz, G. V. (2006). Situación de la educación en el Ecuador. *Observatorio de la Economía Latinoamericana*, *70*.

Díaz Nafría, J. M., Alfonso Cendón, J., y Panizo Alonso, L. (2015). Building up eParticipatory decision-making from the local to the global scale. Study case at the European Higher Education Area. *Computers in Human Behavior*, *47*, 26-41.

Dorta, I., León, C., Rodríguez, C., Rodríguez, G., y Rojas, A. (2003). Complejidad Algorítmica: De la Teoría a la Práctica. *III Jornadas de Enseñanza Universitaria de Informática*.

Dos Santos, H. L., Cechinel, C., Nunes, J. B. C., y Ochoa, X. (2017). An initial review of learning analytics in Latin America. *2017 Twelfth Latin American Conference on Learning Technologies (LACLO)*, 1-9.

Douglas, L., y Marchal, W. (2018). *Estadistica aplicada a los Negocios y la Economia* (S. Wathen, Trad.; Maria Teresa Zapata Terranzas). Mc Graw Hill.

Download R-3.5.2 for Windows. The R-project for statistical computing. (2021). https://cran.r-project.org/bin/windows/base/old/3.5.2/

Draper, N. R., y Cox, D. R. (1969). On Distributions and Their Transformation to Normality. *Journal of the Royal Statistical Society: Series B (Methodological)*, *31*(3), 472-476. https://doi.org/10.1111/j.2517-6161.1969.tb00806.x

DSpace: An Open Source Dynamic Digital Repository. (2018). http://www.dlib.org/dlib/january03/smith/01smith.html

Elia, I., Özel, S., Gagatsis, A., Panaoura, A., y Özel, Z. E. Y. (2016). Students' mathematical work on absolute value: Focusing on conceptions, errors and obstacles. *ZDM*, *48*(6), 895-907.

Espinoza Freire, E. E. (2018). El problema de investigación. *Conrado*, *14*(64), 22-32.

ESPOL. (2019). http://www.espol.edu.ec/

Fallas, J. (2012). Prueba de Hipótesis. *Recuperado de: http://www. ucipfg. com/Repositorio/MGAP/MGAP*, *5*.

Fayyad, U., Piatetsky-Shapiro, G., y Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27-34.

Fazio, C., Battaglia, O. R., y Di Paola, B. (2013). Investigating the quality of mental models deployed by undergraduate engineering students in creating explanations: The case of thermally activated phenomena. *Physical Review Special Topics-Physics Education Research*, *9*(2), 020101.

Fazio, C., Battaglia, O. R., y Sperandeo-Mineo, R. M. (2017). Quantitative and qualitative analysis of the mental models deployed by undergraduate students in explaining thermally activated phenomena. *Scientia in educatione*, *8*.

Ferguson, R. (2014). Learning Analytics: Drivers, developments and challenges. *Italian Journal of Educational Technology*, *22*(3), 138-147.

Fernandez, D. B., y Lujan-Mora, S. (2016). Exploring approaches to educational data mining and learning analytics, to measure the level of acquisition of student's learning outcome. En L. G. Chova, A. L. Martinez, y I. C. Torres (Eds.), *Edulearn16: 8th International Conference on Education and New Learning*

Technologies (pp. 1845-1850). Iated-Int Assoc Technology Education a& Development. http://gateway.isiknowledge.com/gateway/Gateway.cgi?GWVersion=2&SrcAuth=ResearchSoft&SrcApp=EndNote&DestLinkType=FullRecord&DestApp=WOS&KeyUT=WOS:000402955901129

Fernández, S. F., Sánchez, J. M. C., Córdoba, A., y Largo, A. C. (2002). *Estadística Descriptiva*. ESIC Editorial.

Fiallos, A., y Ochoa, X. (2019). Semi-automatic generation of intelligent curricula to facilitate learning analytics. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3303772.3303834

Fillottrani, P. R. (2009). Algoritmos y Complejidad. *Algoritmos sobre grafos, http://www. cs. uns. edu. ar/prf/teaching/AyC09/clase15. pdf*.

Fonseca, D., Garcia-Penalvo, F. J., y Camba, J. D. (2021). *New methods and technologies for enhancing usability and accessibility of educational data*. Springer.

Fotiadis, T. A., y Anastasiadou, S. (2019). *Contemporary advanced statistical methods for the science of marketing: Implicative Statistical Analysis vs Principal Components Analysis*.

Fox, J., y Weisberg, S. (2011). *An R Companion to Applied Regression*. SAGE Publications.

Gaboardi, M., y Rogers, R. (2018). Local private hypothesis testing: Chi-square tests. *International Conference on Machine Learning*, 1626-1635.

Gallerati, P. (2008). Health, Safety & Environmental Performance Data Mining Using Statistical Implicative Analysis. *SPE International Conference on Health, Safety, and Environment in Oil and Gas Exploration and Production*.

García Holgado, A., Marcos Pablos, S., y García Peñalvo, F. J. (2020). Guidelines for performing systematic research projects reviews. *International Journal of Interactive Multimedia and Artificial Intelligence*, *6*(2), 9.

García Peñalvo, F. J., Rodríguez Conde, M. J., Therón Sánchez, R., García Holgado, A., Benito Santos, A., y Martínez Abad, F. (2019). Grupo GRIAL. *IE Comunicaciones. Revista Iberoamericana de Informática Educativa*, *30*, 33-48.

García Peñalvo, F. J., y Seoane Pardo, A. M. (2015). Una revisión actualizada del concepto de eLearning: Décimo Aniversario= An updated review of the concept of eLearning: Tenth anniversary. *Una revisión actualizada del concepto de eLearning: décimo Aniversario= An updated review of the concept of eLearning: tenth anniversary*, 119-144.

García-Peñalvo, F. J. (2019). *Conociendo a GRIAL*.

García-Peñalvo, F. J., y Alier Forment, M. (2014). *Learning management system: Evolving from silos to structures*. Taylor & Francis.

García-Tinisaray, D., Mejias, J. L. P., Pichardo, J. M. M., y del Mar Perez Sanagustin M., O. X. (2018). Learning Analytics as an analysis factor of university academic performance. *CEUR Workshop Proceedings*, *2231*.

Gomes, da S. J. C., y Régnier, J.-C. (2005). *Nouveaux Apports Théoriques à l'Analyse Statistique Implicative et Applications. Critérios de adoção e utilização do livro didático de matemática no ensino fundamental do nordeste brasileiro. P. 145-161.(Critères d'adoption et utilisation du livre didactique de mathématiques dans l'enseignement fondamental nord-est Brésilien.)*.

Gonzalez, A.-B., Rodriguez, M.-J., Olmos, S., Borham, M., y García, F. (2013). Experimental evaluation of the impact of b-learning methodologies on engineering students in Spain. *Computers in Human Behavior*, *29*(2), 370-377.

González, E. R. (1968). *Estadística general* (Número 16). Universidad Central de Venezuela.

González, I. F., Urrútia, G., y Alonso-Coello, P. (2011). Revisiones sistemáticas y metaanálisis: Bases conceptuales e interpretación. *Revista española de cardiología*, *64*(8), 688-696.

Graham, C. R. (2006). Blended learning systems: Definition, current trends, and future directions, Handbook of blended learning: Global perspectives, local designs. *Local Designs*, *2*, 3-18.

Gras, R. (1991). L'analyse de données: Une méthodologie de traitement de questions de didactique. *Publications mathématiques et informatique de Rennes*, *S6*, 115-118.

Gras, R. (2005). Panorama du développement de l'ASI à partir de situations fondatrices. *Actes des Troisièmes Rencontres Internationale ASI Analyse Statistique Implicative, Volume Secondo supplemento al*, *15*, 9-33.

Gras, R. (2014). Genese et developpement de l'analyse statistique implicative: Retrospective Historique. *Educ Matem Pesq São Paulo*, *16*(3), 645-661.

Gras, R., Ag Almouloud, S., Bailleul, M., Larher, A., Polo, M., Ratsimba-Rajohn, H., y Totohasina, A. (1996). *L'implication statistique, nouvelle méthode exploratoire de données*.

Gras, R., Couturier, R., y Gregori, P. (2015). Un mariage arrange entre l'implication et la confiance. *8th International Meeting Statistical Implicative Analysis. Tunisia: Institut Supérieur des Études Technologiques de Radès*.

Gras, R., David, J., Régnier, J.-C., y Guillet, F. (2006). *Typicalité et contribution des sujets et des variables supplémentaires en Analyse Statistique Implicative*. 359-370.

Gras, R., y Kuntz, P. (2009). El Análisis Estadístico Implicativo (ASI) en respuesta a problemas que le dieron origen. *Teoría y aplicaciones del Análisis Estadístico*

Implicativo: primera aproximación en lengua hispana. Castellón: Departamento de Matemática de la Universitat Jaume I, 3-51.

Gras, R., Kuntz, P., y Briand, H. (2001). Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données. *Mathématiques et sciences humaines. Mathematics and social sciences*, *154*.

Gras, R., Kuntz, P., Briand, H., y Couturier, R. (2002). Hiérarchie de règles généralisées et notion de variable supplémentaire en analyse statistique implicative. *Actes des IXèmes Rencontres de la Société Francophone de Classification, Université de Toulouse*, 211-214.

Gras, R., y Régnier, J.-C. (2017). *Dualité entre variables actives et variables supplémentaires: Typicalité et contribution*. Cépaduès Editions.

Gras, R., Regnier, J.-C., y Guillet, F. (2009). *Analyse statistique implicative*.

Gros, B., y García-Peñalvo, F. J. (2016). *Future trends in the design strategies and technological affordances of e-learning*. Springer.

Gruzd, A., y Conroy, N. (2020). Learning Analytics Dashboard for Teaching with Twitter. *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

Gutiérrez, F., Seipp, K., Ochoa, X., Chiluiza, K., De Laet, T., y Verbert, K. (2018). LADA: A learning analytics dashboard for academic advising. *Computers in Human Behavior*.

Gutiérrez, R. B., y Pere, G. C. (2010). *55 respuestas a dudas típicas de estadística*. Ediciones Díaz de Santos.

Hahsler, M. (2019, mayo 20). *Visualizing Association Rules and Frequent Itemsets [R package arulesViz version 1.3-3]*. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=arulesViz

Hahsler,Michael, Buchta,Christian, Gruen,Bettina, Hornik,Kurt, Johnson,Ian, y Borgelt,Christian. (2020, mayo 15). *CRAN - Arules del paquete*. https://cran.r-project.org/web/packages/arules/index.html

Harris, A. D., McGregor, J. C., Perencevich, E. N., Furuno, J. P., Zhu, J., Peterson, D. E., y Finkelstein, J. (2006). The use and interpretation of quasi-experimental studies in medical informatics. *Journal of the American Medical Informatics Association*, *13*(1), 16-23.

Hernández, R., Fernández, C., y Baptista, P. (2010). Metodología de la. *Ciudad de México: Mc Graw Hill*, *12*, 20.

Hernández-García, Á., y Conde, M. Á. (2016). Learning analytics: Needs and opportunities. *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 309-312.

Hider, P., y Pymm, B. (2008). Empirical research methods reported in high-profile LIS journal literature. *Library & information science research*, *30*(2), 108-114.

IEEE - The world's largest technical professional organization dedicated to advancing technology for the benefit of humanity. (2020). https://www.ieee.org/

IEEE Xplore Digital Library. (2019). https://ieeexplore.ieee.org/Xplore/home.jsp

IES, listado provisional | CES - Consejo de Educación Superior | Ecuador. (2020). http://www.ces.gob.ec/index.php?option=com_sobipro&sid=159&Itemid=335

Iglesias Pedrejón, A. (2018). *Réplica y agregación de resultados de un experimento verdadero sobre el impacto de los mecanismos de usabilidad de preferencias, retroalimentación de progreso y abortar operación en un entorno web*.

Ihaka, R., y Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, *5*(3), 299-314.

Inicio | Universidad de Cuenca. (2019). https://www.ucuenca.edu.ec/

Inicio—Escuela Superior Politécnica de Chimborazo. (2016). https://www.espoch.edu.ec/

Inzunsa Cazares, S., y Jiménez Ramírez, J. V. (2013). Caracterización del razonamiento estadístico de estudiantes universitarios acerca de las pruebas de hipótesis. *Revista latinoamericana de investigación en matemática educativa*, *16*(2), 179-211.

Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., y Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods*, *6*(1), 9.

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning* (Vol. 1). Sthda.

Kassambara,Alboukadel. (2020). *factoextra: Extraiga y visualice los resultados de análisis de datos multivariados*. https://CRAN.R-project.org/package=factoextra

Kaufman, L., y Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, *344*, 68-125.

Khaled, H., y Couturier, R. (2015). Apport de la Combinaison de la Méthode d'Analyse Statistique Implicative (ASI) avec la Théorie de Réponse aux Items (IRT). *Actes du 8ème Colloque International sur Analyse Statistique Implicative*, 243-262.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., y Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, *51*(1), 7-15.

Kitchenham, B., y Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering*.

Kitchenham, B., Pretorius, R., Budgen, D., Brereton, O. P., Turner, M., Niazi, M., y Linkman, S. (2010). Systematic literature reviews in software engineering–a tertiary study. *Information and Software Technology*, *52*(8), 792-805.

Kizilcec, R. F., Pérez-Sanagustín, M., y Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & education*, *104*, 18-33.

Kloke, J., y McKean, J. (2020). *Rfit: Rank-Based Estimation for Linear Models* (0.24.2) [Computer software]. https://CRAN.R-project.org/package=Rfit

Kloos, C. D., Alario-Hoyos, C., Fernández-Panadero, C., Estévez-Ayres, I., Muñoz-Merino, P. J., Cobos, R., Moreno, J., Tovar, E., Cabedo, R., Piedra, N., Chicaiza, J., López, J., y Mendes A.J., G.-P. F. J. (2016). EMadrid project. *2016 International Symposium on Computers in Education, SIIE 2016: Learning Analytics Technologies*. https://doi.org/10.1109/SIIE.2016.7751870

Kohanova, I. (2012). Analysis of university entrance test from mathematics. *Acta Didactica Universitatis Comenianae Mathematics*, *12*, 31-46.

Kortenkamp, U., y Ladel, S. (2014). Flexible Use and Understanding of Place Value via Traditional and Digital Tools. *North American Chapter of the International Group for the Psychology of Mathematics Education*.

Koufogiannakis, D., Slater, L., y Crumley, E. (2004). A content analysis of librarianship research. *Journal of information science*, *30*(3), 227-239.

Kwecko, V., de Tôledo, F. P., Devincenzi, S., Ortiz, J. O. de S., y Botelho, S. S. da C. (2020). Analysis of the feelings of the population's opinion in social media: A look at education. *2020 IEEE Frontiers in Education Conference (FIE)*, 1-9.

Lahanier-Reuter, D. (2008). Didactics of mathematics and implicative statistical analysis. En *Statistical Implicative Analysis* (pp. 277-298). Springer.

Lahanier-Reuter, D., Gras, R., y Bailleul, M. (2017). Variable nodale et cône implicatif. *Gras, R., Régnier, J.-C., Lahanier-Reuter, D., Marinica, C., Guillet, F. Analyse*

Statistique Implicative. Des Sciences dures aux sciences humaines et sociales, Toulouse: Éditions Cépaduès.

LAK 2011: 1st International Conference Learning Analytics and Knowledge. (2011). http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=11606

Lang, C., Siemens, G., Wise, A., y Gasevic, D. (2017). *Handbook of learning analytics*. SOLAR, Society for Learning Analytics and Research.

Larher, A. (1991). *Implication statistique et applications a l'analyse des demarches de preuve mathematique*. Rennes 1.

Lasa, A., Abaurrea, J., y Iribas, H. (2020). Mathematical Content on STEM Activities. *Journal on Mathematics Education*, *11*(3), 333-346.

Laxmi, K. R., Ramya, N., Pallavi, S., y Madhuravani, K. (2020). Study and Analysis of Apriori and K-Means Algorithms for Web Mining. En *Innovations in Electronics and Communication Engineering* (pp. 693-701). Springer.

Lerman, I., Chantrel, T., y Cohen, I. (1981). *Classification et analyse ordinale des données* (Vol. 15). Dunod Paris.

Lerman, I., Gras, R., y Rostam, H. (1981). Élaboration et évaluation d'un indice d'implication pour des données binaires. 2. *Mathématiques et sciences humaines*, *75*, 5-47.

Lind, D. A., Marchal, W. G., Wathen, S. A., Obón León, M. del P., y León Cárdenas, J. (2012). *Estadística aplicada a los negocios y la economía*. México: McGraw-Hill/Interamericana Editores.

Lozano, X. B., y Fuentes, M. M. (2012). *Análisis y selección de inversiones en mercados financieros*. Profit Editorial.

Luzardo, G., Guamán, B., Chiluiza, K., Castells, J., y Ochoa, X. (2014). Estimation of presentations skills based on slides and audio features. *MLA 2014 - Proceedings of*

the 2014 ACM Multimodal Learning Analytics Workshop and Grand Challenge, Co-Located with ICMI 2014. https://doi.org/10.1145/2666633.2666639

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, *1*(14), 281-297.

Madrigal, L. M. de. (1996). *Introducción a la estadística de la salud*. Editorial Universidad de Costa Rica.

Maechler, M., original), P. R. (Fortran, original), A. S. (S, original), M. H. (S, Hornik [trl, K., maintenance(1999-2000)), ctb] (port to R., Studer, M., Roudier, P., Gonzalez, J., Kozlowski, K., pam()), E. S. (fastpam options for, y Murphy (volume.ellipsoid({d >= 3})), K. (2019). *cluster: «Finding Groups in Data»: Cluster Analysis Extended Rousseeuw et al.* (2.1.0) [Computer software]. https://CRAN.R-project.org/package=cluster

Marangunić, N., y Granić, A. (2015). Technology acceptance model. *Universal Access in the Information Society*, *14*(1), 81-95. https://doi.org/10.1007/s10209-014-0348-1

Margolinas, C. (2014). Teaching Fractions through Situations: A Fundamental Experiment. *REDIMAT*, *3*(2), 186-188.

Marín Martínez, P. (2017). *Probability distribution of the classical implication intensity seen as a random variable in Statistical Implicative Analysis*.

Melusova, J., y Vidermanova, K. (2015). Upper-secondary students' strategies for solving combinatorial problems. *Procedia-Social and Behavioral Sciences*, *197*, 1703-1709.

Mereología | Qué es, Definición y Concepto. (2021). *Enciclopedia Online*. https://enciclopediaonline.com/es/mereologia/

Mersmann, O., Beleites, C., Hurling, R., Friedman, A., y Ulrich, J. M. (2019). *microbenchmark: Accurate Timing Functions* (1.4-7) [Computer software]. https://CRAN.R-project.org/package=microbenchmark

Michael, P., Elia, I., Gagatsis, A., y Kalogirou, P. (2010). *Examining primary school students' operative apprehension of geometrical figures through a comparison between the hierarchical clustering of variables, implicative statistical analysis and confirmatory factor analysis*. Citeseer.

Mode, E. B. (1990). *Elementos de probabilidad y estática*. Reverte.

Montgomery, D. C., Runger, G. C., y Medal, E. G. U. (1996). *Probabilidad y estadística aplicadas a la ingeniería* (Números 968-18-5914-6. 01-A1 LU. AL-PyE. 1.). McGraw-Hill México DF.

Montilla, J.-M., y Kromrey, J. (2010). Robustez de las pruebas T en comparación de medias, ante violación de supuestos de normalidad y homocedasticidad. *Ciencia e Ingeniería*, *31*(2), 101-107.

Moore, D. S. (2005). *Estadística aplicada básica, 2a ed.* Antoni Bosch editor.

Mount, John. (2020). *GitHub—WinVector / replyr: Parches para usar dplyr con bases de datos y Big Data*. https://github.com/WinVector/replyr

Muhammad, R. N., Tasmin, R., y Aziati, A. N. (2020). Sustainable Competitive Advantage of Big Data Analytics in Higher Education Sector: An Overview. *Journal of Physics: Conference Series*, *1529*(4), 042100.

Naranjo, M., Pazmiño-Maji, R., Conde, M., y Peñalvo, F. (2018). *LA&SIA cluster methods: Computational comparison*.

Naranjo Serrano, M. M., y Pazmiño Maji, R. A. (2018a). *Estudio comparativo del anàlisis estadìstico implicativoy el Learning Analytics en relaciòn al uso de las tècnicas de*

exploracoòn de datos educativos. http://repositorio.pucesa.edu.ec/handle/123456789/2387

Naranjo Serrano, M. M., y Pazmiño Maji, R. A. (2018b). *Estudio comparativo del anàlisis estadìstico implicativoy el Learning Analytics en relaciòn al uso de las tècnicas de exploracoòn de datos educativos*. http://repositorio.pucesa.edu.ec/handle/123456789/2387

Navarrete, D. (2019). *Learning Analytics Perú: Plataforma de desarrollo para la Analítica del Aprendizaje en el Perú*. https://repositorioacademico.upc.edu.pe/handle/10757/624844

Neiva, F. W., David, J. M. N., Braga, R., y Campos, F. (2016). Towards pragmatic interoperability to support collaboration: A systematic review and mapping of the literature. *Information and Software Technology*, *72*, 137-150.

Nikolantonakis, K., y Vivier, L. (2013). Positions numeration in any base for future elementary school teachers in France and Greece: One discussion via registers and praxis. *Menon, Florina*, *2*, 99-114.

Nordstokke, D. W., y Zumbo, B. D. (2010). A new nonparametric Levene test for equal variances. *Psicológica*, *31*(2), 401-430.

NVivo qualitative data analysis software | QSR International. (2016). https://www.qsrinternational.com/nvivo/home

NYU Steinhardt. (2020). NYU Steinhardt. https://steinhardt.nyu.edu/

Ochoa, X. (2019). Learning analytics in Latin America present an opportunity not to be missed. *Nature human behaviour*, *3*(1), 6.

Ochoa, X., Chiluiza, K., Méndez, G., Luzardo, G., Guamán, B., y Castells, J. (2013). Expertise estimation based on simple multimodal features. *ICMI 2013 -*

Proceedings of the 2013 ACM International Conference on Multimodal Interaction, 583-590. https://doi.org/10.1145/2522848.2533789

Ochoa, X., McKay, T., Molinaro, y Greer. (2016). Simple metrics for curricular analytics. *CEUR Workshop Proceedings*, *1590*.

Okada, A., Whitelock, D., Holmes, W., y Edwards, C. (2017). Student acceptance of online assessment with e-authentication in the UK. *International Conference on Technology Enhanced Assessment*, 109-122.

Okoli, C., y Schabram, K. (2010). *A guide to conducting a systematic literature review of information systems research*.

Orús, P., Gregori, P., y Castellón, C. R. S.-E. (2005). Des variables supplémentaires et des élèves «fictifs», dans la fouille didactique de données avec CHIC. *Troisième rencontre internationale de l'Analyse Statistique Implicative (ASI3)*, 279-291.

Panitsides, E. A., y Anastasiadou, S. (2015). Lifelong learning policy agenda in the European union: A bi-level analysis. *Open Review of Educational Research*, *2*(1), 128-142.

Papamitsiou, Z., y Economides, A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, *17*(4), 49-64.

Patten, M. L., y Newhart, M. (2017). *Understanding research methods: An overview of the essentials*. Taylor & Francis.

Pavlovičová, G., y Vargová, L. (2020). Investigation of Selected Aspects of Fraction Understanding. *TEM Journal*, *9*(2), 702.

Pavlovicova, G., y Zahorska, J. (2015). The attitudes of students to the geometry and their concepts about square. *Procedia-social and behavioral sciences*, *197*, 1907-1912.

Pazmiño Maji, R. A., Ortega, J. R. L., González, M. Á. C., y Peñalvo, F. J. G. (2019). Las analíticas de aprendizaje en el Ecuador: Un análisis inicial basado en el mapeo sistemático de los trabajos de graduación. *Explorador Digital*, *3*(3.1), 224-245.

Pazmiño Maji, R., García Peñalvo, F. J., y Conde González, M. Á. (2017). *Is it possible to apply Statistical Implicative Analysis in hierarchical cluster Analysis? Firsts issues and answers*.

Pazmiño, R., García-Peñalvo, F. J., Coutrier, R., y Conde-González, M. (2015). *Statistical implicative analysis for educational data sets: 2 analysis with RCHIC*.

Pazmiño-Maji, García-Peñalvo, F. J., y Conde-González, M. A. (2016). Approximation of statistical implicative analysis to learning analytics: A systematic review. *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*, 355-376.

Pazmiño-Maji, R. (2014a). *Aporte del Análisis Estadístico Implicativo a las Analíticas de Aprendizaje*. http://hdl.handle.net/10366/124058

Pazmiño-Maji, R. (2014b). *Aproximación al Análisis Estadístico Implicativo desde sus Aplicaciones Educativas*.

Pazmiño-Maji, R. A., Conde González, M. Á., y García Peñalvo, F. J. (2019). *Learning Analytics in Ecuador: Analysis based in a Mapping Review*. Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality. D:\8 Citavi5\Projects\LAecuador_JCR\Citavi Attachments\Learning Analytics in Ecuador 2019.pdf

Pazmiño-Maji, R. A., García-Peñalvo, F. J., y Conde-González, M. A. (2017). Statistical Implicative Analysis Approximation to KDD and Data Mining: A Systematic and Mapping Review in Knowledge Discovery Database Framework. *DBKDA 2017*, 79.

Pazmiño-Maji, R., Bonilla, M., Baquero, J., y Miguez, R. (2018). Software estadístico chic: Descubriendo sus potencialidades mediante el análisis de percepción sexual universitaria. *Ciencia digital*, *2*, 17.

Pazmiño-Maji, R., Conde González, M. Á., y García Peñalvo, F. J. (2019). Las analíticas de aprendizaje en el Ecuador: Un análisis inicial basado en el mapeo sistemático de los trabajos de graduación. *Explorador Digital*, *3*(3.1), 224-245. https://doi.org/10.33262/exploradordigital.v3i3.1.885

Pazmiño-Maji, R., Conde, M. Á., y García-Peñalvo, F. (2021). Learning analytics in Ecuador: A systematic review supported by statistical implicative analysis. *Universal Access in the Information Society*, 1-18.

Pazmiño-Maji, R., García-Peñalvo, F., Conde-González, M., y Solis Benavides, C. (2019). *La investigación de pregrado en la Escuela Superior Politécnica de Chimborazo: Mapeo sistemático y analíticas*.

Pazmiño-Maji, R., García-Peñalvo, F. J., y Conde-González, M. A. (2017a). *Association rules with SIA in B-Learning Courses: A mapping review*.

Pazmiño-Maji, R., García-Peñalvo, F. J., y Conde-González, M. A. (2017b). *Comparing Hierarchical Trees in Statistical Implicative Analysis & Hierarchical Cluster in Learning Analytics*. 1-7.

Pazmiño-Maji, R., García-Peñalvo, F. J., y Conde-González, M. A. (2017c). *Statistical Implicative Analysis aproximation to KDD and Data Mining: A systematic and mapping review in Knowledge Discovery Database framework.*

Pazmiño-Maji, R., López-Ortega, J., González, M. Á. C., y Peñalvo, F. J. G. (2019). Las analíticas de aprendizaje en el Ecuador: Un análisis inicial basado en el mapeo sistemático de los trabajos de graduación. *Explorador Digital*, *3*(3.1), 224-245.

Pazmiño-Maji, R., Naranjo-Ordoñez, L., Conde-González, M., y García-Peñalvo, F. (2019). Learning analytics in Ecuador: An initial analysis based in a mapping review. En Conde-Gonzalez M.A., Rodriguez-Sedano F.J., Fernandez-Llamas C., y Garcia-Penalvo F.J. (Eds.), *ACM Int. Conf. Proc. Ser.* (pp. 304-311). Association for Computing Machinery; Scopus. https://doi.org/10.1145/3362789.3362913

Pazmiño-Maji, R., Pérez, M. G., y Andaluz, V. (2014). Cuasi-implicación estadística y determinación automática de clases de equivalencia en imágenes de resonancia magnética de cerebro. *Revista Politécnica*, *34*(1).

Pazmiño-Maji Rubén, Conde-Gonzales M.A., y Garcia-Penalvo F.J. (2021). *What are Learning Analytics ?: Analysis from its definition*. *1*, 10.

Pérez-Álvarez, R., Maldonado-Mahauad, J., Pérez-Sanagustín, M., y Elferink R., D. H. P.-S. V. P.-S. M. S. M. (2018). Tools to Support Self-Regulated Learning in Online Environments. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11082 LNCS*, 16-30. https://doi.org/10.1007/978-3-319-98572-5_2

Pérez-Caraballo, G., Acioly-Régnier, N. M., y Régnier, J.-C. (2014). Competences professionnelles et linguistiques de professionnels de santé dans l'espace frontalier Uruguayen-Bresilien Professional and linguistic competence of health care providers in the Uruguay-Brazil border region: SIA'S contribution. *Educação Matemática Pesquisa: Revista do Programa de Estudos Pós-Graduados em Educação Matemática*, *16*(3), 813-853.

Phan, L. P., Huynh, H. X., Huynh, H. H., y Nguyen, K. M. (2016). Association-based recommender system using statistical implicative cohesion measure. *2016 Eighth International Conference on Knowledge and Systems Engineering (KSE)*, 144-149.

Piedra, N., Chicaiza, J., López, J., y Tovar Caro, E. (2015). Towards a learning analytics approach for supporting discovery and reuse of OER an approach based on Social Networks Analysis and Linked Open Data. *IEEE Global Engineering Education Conference, EDUCON*, *2015-April*. https://doi.org/10.1109/EDUCON.2015.7096092

Pizzolato, N., Fazio, C., Mineo, R. M. S., y Adorno, D. P. (2014). Open-inquiry driven overcoming of epistemological difficulties in engineering undergraduates: A case study in the context of thermal science. *Physical Review Special Topics-Physics Education Research*, *10*(1), 010107.

Quine, W. V., y Carnap, R. (2020). Homage to Rudolf Carnap. En *Dear Carnap, Dear Van* (pp. 463-466). University of California Press.

R. (2021). https://www.r-project.org/about.html

R: Pairwise Wilcoxon Rank Sum Tests. (2021). https://stat.ethz.ch/R-manual/R-devel/library/stats/html/pairwise.wilcox.test.html

Rakotomalala, H. F., Ralahady, B. B., y Totohasina, A. (2019). A Novel Cohesitive Implicative Classification Based on Application on Diagnostic on Informatics Literacy of Students of Higher Education in Madagascar. *Third International Congress on Information and Communication Technology*, 161-174.

Rakotomalala, H. F., y Totohasina, A. (2020). On Hierarchical Classification Implicative and Cohesive Application on Based: Application on Analysis of the Computing Curricula and Students Abilities According the Anglo-Saxon Model. *Fourth International Congress on Information and Communication Technology*, 83-90.

Raphael Couturier. (2016, agosto 26). *Rchic*. Couturier,Raphael. https://members.femto-st.fr/raphael-couturier/en/rchic

Razali, N. M., y Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, *2*(1), 21-33.

Rchic. (2016, agosto 26). Raphael Couturier. https://members.femto-st.fr/raphael-couturier/en/rchic

Real Academia Española. (2021). http://www.asale.org/academias/real-academia-espanola

Régnier, J.-C., Gras, R., Couturier, R., y Bodin, A. (2019). *Analyse Statistique Implicative.*

Régnier, J.-C., Gras, R., Henry, M., Couturier, R., y Brousseau, G. (2020). *Analyse Statistique Implicative.*

Repositorio Digital Senescyt: Página de inicio. (2019). http://repositorio.educacionsuperior.gob.ec/

REUTERS, T. (2017, abril 29). *Web of Science [v.5.24]—Colección principal de Web of Science*. apps.webofknowledge.com

RRAAE Home. (2019). http://www.rraae.org.ec/

RStudio. (2020). https://rstudio.com/products/rstudio/

RStudio | Open source & professional software for data science teams. (2020). https://rstudio.com/

Ruipérez-Valiente, J. A., Muñoz-Merino, P. J., Leony, D., y Kloos, C. D. (2015). ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform. *Computers in Human Behavior*, *47*, 139-148.

Sagaró del Campo, N. M., y Zamora Matamoros, L. (2019). Análisis estadístico implicativo versus Regresión logística binaria para el estudio de la causalidad en salud. *Multimed*, *23*(6), 1416-1440.

Salazar Pinto, C., Castillo Galarza, S. del, y Del Castillo Galarza, S. (2017). *Fundamentos básicos de estadística*.

Santos, M. C. M. E., Santos, P. C. M. D. A., Acioly-Régnier, N. M., y Régnier, J.-C. (2014). Motivações e competências interculturais para a mobilidade acadêmica França-Brasil: O caso de estudantes da Universidade Lumière Lyon 2 Motivations and intercultural skills for France-Brazil academic mobility: The case of students from the University. *Educação Matemática Pesquisa: Revista do Programa de Estudos Pós-Graduados em Educação Matemática*, *16*(3), 723-744.

Sarguera, C. R. B., y Rebustillo, C. M. R. (2017). Estructura del problema de investigación, contradicciones inherentes y exigencias metodológicas para su formulación/structure of the investigation problem, inherent contradictions and methodological requirements for their formulation. *Pedagogía Universitaria*, *22*(2), 1-19.

Scopus—Document search. (2020). https://www.scopus.com/search/form.uri?display=basic

Scopus—Welcome to Scopus. (2017, abril 29). https://www.scopus.com/home.uri

Senescyt – Secretaría de Educación Superior, Ciencia, Tecnología e Innovación – Ser Bachiller, Becas, Investigación, Innovación Ecuador. (2020). https://www.educacionsuperior.gob.ec/

Silva, K. A. de G., y de Almeida, M. E. B. (2017). Combined use of software that supports research and qualitative data analysis: Potential applications for researches in education. En *Computer Supported Qualitative Research* (pp. 25-37). Springer.

Sinónimos de aporte. (2020). https://trovami.altervista.org/es/sinonimi/aporte

Soto, P. J. L. (2013). Contraste de hipótesis. Comparación de más de dos medias independientes mediante pruebas no paramétricas: Prueba de Kruskal-Wallis. *Revista Enfermería del Trabajo*, *3*(4), 166-171.

Spagnolo, F. (2005). *L'analyse statistique implicative: Une des méthodes d'analyse des données en didactique*.

Sprock, A. S., Vicari, R. M., Rincón, M. R., Silveira, I. F., Gallegos, J. P., Maldonado, J., Toscano, A., y Casali A., R. M. C. D. A. S. A. S. (2017). Latin-American Network of Learning Analytics—LALA. *12th Latin American Conference on Learning Objects and Technologies, LACLO 2017*, *2017-January*. https://doi.org/10.1109/LACLO.2017.8120916

Team, R. C. (2013). *R: A language and environment for statistical computing*.

The Comprehensive R Archive Network. (2015). https://cran.r-project.org/

Tinisaray, G., y Karina, D. (2016). *Construcción de un modelo para determinar el rendimiento académico de los estudiantes basado en learning analytics (análisis del aprendizaje), mediante el uso de técnicas multivariantes*.

Torre, R. D. de la. (2004). *Iniciación a la probabilidad y la estadística*. Univ. Autònoma de Barcelona.

Torres, J. C. (2003). Diagnóstico de la Educación Superior Virtual en Ecuador. *La educación superior virtual en américa latina y el caribe*, 269.

Triola, M. F. (2004). *Estadística*. Pearson Educación.

Turcios, R. S. (2015). Prueba de Wilcoxon-Mann-Whitney: Mitos y realidades. *Rev Mex Endocrinol Metab Nutr*, *2*, 18-21.

Turgut, M. (2018). Synergies among students' thinking modes and representation types in linear algebra: Employing statistical implicative analysis. *International journal of mathematical education in science and technology*, *49*(8), 1181-1202.

Using-data-improve-student-learning. (2016). http://www.eqao.com/en/Our_Data_in_Action/articles/Pages/%20using-data-improve-student-learning.aspx

UTPL | Decide ser más. (2012). https://www.utpl.edu.ec/

Valls, X. (2014). *Diseño de un paquete R para el Análisis Estadístico Implicativo*.

Van den Heuvel-Panhuizen, M., y Elia, I. (2020). Mapping kindergartners' quantitative competence. *ZDM*, *52*(4), 805-819.

van den Heuvel-Panhuizen, M., Elia, I., y Robitzsch, A. (2015). Kindergartners' performance in two types of imaginary perspective-taking. *ZDM*, *47*(3), 345-362.

Vásquez, A. C. (2004). Teoría de la complejidad computacional y teoría de la computabilidad. *Revista de investigación de Sistemas e Informática*, *1*(1), 102-105.

Vidal Ledo, M., Oramas Díaz, J., y Borroto Cruz, R. (2015). Revisiones sistemáticas. *Educación Médica Superior*, *29*(1), 198-207.

Web of Science—Web of Science Group. (2010). https://clarivate.com/webofsciencegroup/solutions/web-of-science/

Wei, T., y Simko, V. (2017). *R package «corrplot»: Visualization of a Correlation Matrix*. https://github.com/taiyun/corrplot

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., y RStudio. (2020). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* (3.3.2) [Computer software]. https://CRAN.R-project.org/package=ggplot2

Wickham,Hadley, François,Romain, Henry,Lionel, Müller,Kirill, y RStudio. (2020, agosto 18). *CRAN - Package dplyr*. https://cran.r-project.org/web/packages/dplyr/index.html

Wilches, O. E. C., y Grisales-Palacio, V. H. (2017). *Learning Analytics en Colombia*. D:\8 Citavi5\Projects\LAecuador_JCR\Citavi Attachments\Wilches, Grisales-Palacio - Learning Analytics en Colombia.pdf

Wiley, K. J., Dimitriadis, Y., Bradford, A., y Linn, M. C. (2020). From theory to action: Developing and evaluating learning analytics for learning design. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 569-578.

Wolfram Mathematica: Computación técnica moderna. (2015). https://www.wolfram.com/mathematica/

Yang, X.-S., Sherratt, S., Dey, N., y Joshi, A. (2015). *Fourth International Congress on Information and Communication Technology*.

Zamora, L., Gregori, P., y Orús, P. (2009). Conceptos fundamentales del Análisis Estadístico Implicativo (ASI) y su soporte computacional CHIC. *Contribuciones al ASI*, *4*, 65-101.

Zamora-Matamoros, L., Díaz-Silvera, J. R., y Portuondo-Mallet, L. (2015). Fundamental Concepts on Classification and Statistical Implicative Analysis for Modal Variables. *Revista Colombiana de Estadística*, *38*(2), 335-351.

Zhang, H., y Ali Babar, M. (2010). *On searching relevant studies in software engineering*.

Žilková, K. (2015). Misconceptions in pre-service primary education teachers about quadrilaterals. *Journal of Education, Psychology and Social Sciences*, *1*, 2015.