VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

FACULTAD DE CIENCIAS

GRADO EN ESTADÍSTICA

TRABAJO DE FIN DE GRADO

APROXIMACIÓN MULTIVARIANTE A LA CONSECUCIÓN DE LOS OBJETIVOS DE DESARROLLO SOSTENIBLE

Autor: Carlos Nieto Gutiérrez

Tutoras: Nerea González García y Ana Belén Nieto Librero

FACULTAD DE CIENCIAS

GRADO EN ESTADÍSTICA

TRABAJO DE FIN DE GRADO

APROXIMACIÓN MULTIVARIANTE A LA CONSECUCIÓN DE LOS OBJETIVOS DE DESARROLLO SOSTENIBLE

Autor: Carlos Nieto Gutiérrez

Tutoras: Nerea González García y Ana Belén Nieto Librero

Índice:

Índice de ilustraciones:

Índice de tablas:

1. Introducción:

Los *Objetivos de Desarrollo Sostenible (ODS)* son un bien necesario para poder conseguir a largo plazo una sostenibilidad global que nos permita avanzar como sociedad de una manera justa y en beneficio de toda la humanidad. Los *ODS* son herederos directos de *los Objetivos de Desarrollo del Milenio (ODM)* y existen para favorecer la consecución de estos últimos mediante una serie de propuestas a corto plazo (Barrero-Barrero & Baquero-Valdés, 2020).

Para la evaluación de la mayoría de los ámbitos que incumben a estos dos planes, nos servimos del *Índice de Desarrollo Humano (IDH),* que sirve para cuantificar este aspecto en concreto, dentro del desarrollo global del planeta (Rosenberg, 1994).

1.1. Antecedentes de los ODS:

En el año 2015 tuvo lugar un importante acontecimiento para las políticas, los actores y la gobernanza de la cooperación internacional. La Asamblea General de las Naciones Unidas ratificó una resolución que definía una serie de objetivos que se denominarán "*Objetivos de Desarrollo sostenible"*, cuya aplicación se extiende hasta el año 2030 (Perales, 2014).

En primer lugar, hay que hablar de los antecesores de estos; los *ODM.* Estos objetivos plasman y resumen las metas de desarrollo que fueron planteadas en las conferencias internacionales y cumbres mundiales que tuvieron lugar durante la década de los noventa. Fue el 8 de septiembre del año 2000 cuando la Asamblea General de las Naciones Unidas compendió los objetivos y las metas principales en lo que se nombró cómo la Declaración del Milenio (Robles-Llamazares, 2006).

Los *ODM* han sido un hito histórico en el que todo el globo suscribió una promesa concreta: un acuerdo entre todos los países para conseguir reducir la pobreza y las privatizaciones humanas a una velocidad nunca vista a través de una acción colaborativa. Los *ODM* se distinguen de otro tipo de promesas globales en su naturaleza integral y los esfuerzos sistémicos realizados para la financiación, el control y el desarrollo de estos (C. Torres & Mújica, 2004).

La implementación de estos objetivos no estuvo exenta de debate entre los círculos académicos y profesionales. Aparecen diferentes posturas: en primer lugar, contamos con una visión optimista del proyecto que lo considera una acción necesaria para conseguir un cambio en la transformación humana, por otra parte, existían círculos donde no se comprendían los *ODM* como un plan de acción, pero sí creían que eran necesarios para agrandar las ambiciones y conseguir por ello un mayor compromiso político. En último lugar mencionamos un sector crítico más duro que entendía los objetivos como una conspiración; servía únicamente para eclipsar los verdaderos problemas que arrasaban el mundo. No obstante, a pesar de lo antagónico de los diferentes puntos de vista, había algo compartido; tomar a los *ODM* como un hecho latente y real (Hulme, 2009).

La Declaración del Milenio se dividía en 8 objetivos, estos objetivos comprendían a su vez una serie de 28 metas, cuantificables mediante 48 indicadores. Los objetivos eran los siguientes:

- Erradicar la pobreza extrema y el hambre
- Lograr la enseñanza primaria universal

- Promover la igualdad entre los géneros y la autonomía de la mujer
- Reducir la mortalidad infantil
- Combatir el VIH, el paludismo y otras enfermedades
- Garantizar la sostenibilidad del medio ambiente
- Fomentar la asociación mundial para el desarrollo

Desde que se establecieron estas metas era recurrente la numerosa aparición de informes para ir evaluando el desarrollo de estos objetivos; para saber cómo se estaban realizando y saber mediante una serie de indicadores si esta era la manera correcta para conseguirlos. Una vez llegados al año 2015, año que estaba marcado para consecución de todos los objetivos propuestos, era necesario hacer un balance general acerca del desarrollo de estos. Así, tanto numerosas revistas científicas como los departamentos oficiales de los diferentes países empezaron a realizar informes para evaluar como había sido el desarrollo de los *ODM* (Hulme, 2009).

La mayoría de ellos coincidían en lo mismo: los *ODM* han sido un recorrido de luces y sombras. Esto se debe a que se han conseguido importantes avances, pero a su vez ha habido otra serie de puntos en los que claramente no se han cumplido las expectativas, ya que los indicadores actuales se encuentran muy lejos de los esperados en el año 2000. La prestigiosa revista *"British Medical Journal" (BMJ)* publicó un análisis muy interesante respecto a esta cuestión.

En este artículo reconoce grandes avances en algunos de los campos de los *ODM* como por ejemplo en el que se refiere a la mortalidad infantil (cuarto objetivo), ésta se ha conseguido reducir a la mitad en todo el mundo, y aunque en este objetivo se recogía reducir esta lacra un 75% y no ha sido así, no se puede negar que se haya producido un importante avance (Beattie et al., 2015).

También se ha conseguido alcanzar el sexto objetivo, que hacía referencia a la lucha contra la epidemia del virus de la inmunodeficiencia humana, esto se ve con claridad cuando actualmente la prevención y el tratamiento frente esta enfermedad es una realidad. "La infección neonatal por VIH es ahora prevenible, como resultado del gran éxito de los programas de transmisión entre madres e hijos. Esto era algo inconcebible cuando de establecieron los *ODM"* (Prendergast et al., 2015).

Además de esto también ha habido grandes avances en el segundo objetivo, que trata de la necesidad de mejorar el acceso a la educación. La consecución de este objetivo se ve reflejada actualmente en cómo los menores hijos de madres con acceso a educación primaria cuentan con un porcentaje más elevado de supervivencia que aquellos con madres analfabetas (V. E. R. Torres et al., 2018).

Existe también un reverso negativo analizado dentro del artículo: los puntos en los que se ha avanzado muy insuficientemente o no se ha avanzado siquiera. Estos son los que se refieren a: la tasa de mortalidad materna y neonatal, la malnutrición infantil y la igualdad de género. Estos fracasos muchos los achacan a que, en realidad, los *ODM* eran demasiado ambiciosos y que, de todas formas, estos cambios se habían producido de forma natural. Sin embargo, un estudio económico reciente revela que aproximadamente 13,6 millones de vidas adicionales de niños han sido salvadas desde 2001 (Beattie et al., 2015).

Blanca Carazo, responsable de programas de cooperación de UNICEF Comité Español, afirma que para ella el balance de los *ODM* ha sido positivo: "han sido un instrumento muy válido

para aunar voluntades, y el hecho de tener objetivos medibles posibilita saber si se estaba avanzando o no, y también te permitían hacer incidencia política con los países implicados" (Marín, 2015).

A pesar de los avances conseguidos gracias a los *ODM*, para la comunidad internacional estos no eran suficientes y querían buscar otro proyecto a largo plazo que garantizara el desarrollo sostenible del planeta, por eso en el año 2015 fueron aprobados los *"Objetivos de Desarrollo Sostenible"* (Gil, 2017).

1.2. Historia de los ODS:

El 1 de octubre de 2013, cuando faltaban 850 días para que se diera por finalizado el plazo del año 2015, tuvo lugar una reunión entre todos los países para poder lograr el cumplimiento de los *ODM*.

Debido a ello se estableció la necesidad de reforzar aquellos objetivos que distaban más de lograr cumplirse principalmente aquellos relacionados con el hambre, la pobreza, el acceso a una educación primaria, etc. Se confeccionó una agenda sólida para el desarrollo una vez se llegará al 2015, la cual se apoyaba en los *ODM*, estableciendo además que esta agenda debía reforzar el compromiso internacional para la erradicación de la pobreza y el desarrollo sostenible. Incluyendo también la pretensión de que este proceso debería de ser inclusivo y estar centrado para todas las personas, contando con un desarrollo gubernamental completamente transparente y que recibiera aportes de los distintos agentes sociales (Griggs, 2013).

Fue en base a esto que se decidió dar inicio al sexagésimo noveno período de sesiones de la Asamblea General. Periodo encargado de que se iniciase un proceso de negociaciones que consiguiera alcanzar la aprobación de la agenda posterior a 2015 para el desarrollo, estableciendo como fecha para la aprobación de esta agenda septiembre de 2015.

En el año 2014, el grupo de trabajo abierto redactó una carta que fue enviada al presidente de la Asamblea General, adjuntando un informe sobre los *ODS*, con el propósito de que fuera revisada y que se tomaran las medidas propias de la Asamblea General.

El 25 de septiembre del 2015 los 193 países miembros de las Naciones Unidas se reunieron en Nueva York, siendo en esta Asamblea General donde se votaría la aprobación del documento llamado "Transformar nuestro mundo: la Agenda 2030 para el Desarrollo Sostenible", más conocido actualmente como "La Agenda 2030". Esta fue aprobada por unanimidad de los países. Esta agenda fue fruto del trabajo realizado por el grupo de trabajo abierto que se inició en el año 2013 y contó con la participación de 70 países; también incluyó varias consultas a académicos, además que las Naciones Unidas colaboró ayudando a conducir una serie de conversaciones y consultas técnicas. Esta agenda se considera la continuación de los *ODM*, para así abordar los problemas que no se pudieron resolver e incorporando 17 objetivos y 169 metas (Cosme-Casulo, 2018).

En el contenido del preámbulo de la Agenda 2030 podemos identificar en beneficio de quienes se ha llevado a cabo, en él podemos leer lo siguiente: "La presente Agenda es un plan de acción en favor de las personas, el planeta y la prosperidad. También tiene por objeto fortalecer la paz universal dentro de un concepto más amplio de la libertad. Reconocemos que la erradicación de la pobreza en todas sus formas y dimensiones, incluida la pobreza

extrema, es el mayor desafío a que se enfrenta el mundo y constituye un requisito indispensable para el desarrollo sostenible (Aravena-Jara, 2021).

Esta agenda nos presenta que su enfoque son las personas, el planeta, además de luchar por un fortalecimiento de la paz, dando principal importancia a la erradicación de la pobreza.

La agenda tendrá un alcance mundial y una importancia sin precedentes, todos los países la aceptan y la aplican, aunque se tiene en cuenta las diferencias entre ellos, sus capacidades y los distintos niveles de desarrollo, respetando también cada política territorial y las prioridades nacionales. Estas metas afectan a todos, tanto a los países desarrollados como a los que están en vías de desarrollo, son de carácter único e indivisible (Robaina-Romero, 2019).

A pesar de que en la declaración se expresa que tiene un alcance global e involucra a todos los países, hay que tener en cuenta que la agenda es una mera declaración, debido a que queda a cargo de ellos su cumplimiento e implementación ya que no es un acuerdo vinculante. Los 17 objetivos de la Agenda 2030 hacen referencia a diferentes esferas que se intentan estimular. Entre ellas encontramos la que se refiere a las personas, y que pretende erradicar la pobreza y el hambre. Otra esfera referente a la prosperidad en la que se lucha por que todos los seres humanos podamos vivir de manera próspera y plena, que el progreso venga dado de la mano con la naturaleza. En la esfera de la paz encontramos la misión de que se consigan sociedades pacíficas, justas y libres de todo temor y violencia. Y por último en la de las alianzas se establece el compromiso de movilizar los medios que fueran necesarios para poder implementar la agenda (Aravena-Jara, 2021).

Los 17 objetivos que fueron planteados finalmente fueron:

- *Objetivo 1:* Fin de la pobreza
- *Objetivo 2***:** Hambre cero
- *Objetivo 3:* Salud y bienestar
- *Objetivo 4:* Educación de calidad
- *Objetivo 5:* Igualdad de género
- *Objetivo 6:* Agua limpia y saneamiento
- *Objetivo 7:* Energía asequible y no contaminante
- *Objetivo 8:* Trabajo decente y crecimiento económico
- *Objetivo 9:* Agua, industria e innovación e infraestructuras
- *Objetivo 10:* Reducción de las desigualdades
- *Objetivo 12:* Ciudades y comunidades sostenibles
- *Objetivo 13:* Producción y consumos responsables
- *Objetivo 14:* Acción por el clima
- *Objetivo 15:* Vida submarina
- *Objetivo 16:* Vida de ecosistemas terrestres
- *Objetivo 17:* Paz, justicia e instituciones sólidas

1.3. Desarrollo humano:

Además de los ODS, otro factor importante del que habría que hablar y es de gran peso en el trabajo es el Desarrollo Humano. Este se denomina como el proceso por el cual una sociedad, a partir de un buen desarrollo económico, mejora de manera sustancial las condiciones de vida de cada uno de sus miembros (*Significado de Desarrollo Humano*, 2019). Por lo tanto, el desarrollo humano no solo significa que las personas opten a los recursos mínimos para cubrir sus necesidades básicas, si no a su vez, que tengan acceso a los sistemas de salud y educación, pertinentes niveles de seguridad personal, plenas libertades políticas y culturales. Precisamente por esto, uno de los principales objetivos del desarrollo humano es crear las condiciones necesarias para que todas las personas gocen de una numerosa gama de oportunidades como, por ejemplo: empleo, educación, desarrollo productivo etc. Además de que lleven una vida que sea valorada y que esté acorde con sus metas y capacidades personales (Delval, 1994).

De esta forma, el desarrollo humano también significa calidad de vida, una participación activa de las decisiones que son relevantes en nuestro mundo, vías para desarrollar al máximo nuestras capacidades y un total respeto a los derechos humanos y a la dignidad de la vida (Papalia, 2009). Este va más allá del nivel de renta o la riqueza de un país, al contrario, este se centra en el recurso con mayor valor y riqueza con el que cuenta una nación: el capital humano.

Conseguir este desarrollo es una de las metas de la Organización de las Naciones Unidas, más en concreto, el organismo que tiene como misión coordinar todas las políticas al acerca del desarrollo humano es el Programa de Naciones Unidas para el Desarrollo (PNUD). Este presenta periódicamente el Programa Anual Mundial sobre el Desarrollo Humano. Es este informe se exponen los datos estadísticos que calculan, según una serie de indicadores, el nivel de desarrollo humano. El más conocido y en el que nos vamos a centrar es el *IDH*. (Rosenberg, 1994).

El IDH se propuso en 1990 debido a la preocupación que se tenía por medir el desarrollo humano. El Programa de las Naciones Unidas para el Desarrollo, en su informe expone que lo perfecto sería poder incluir muchas variables pero que un exceso de indicadores podría llegar a dar una imagen distorsionada y confusa y que esta fuera el motivo de desvío para las personas encargadas de diseñar las políticas públicas. Por ello se propuso la medición del desarrollo humano centrándose en tres elementos fundamentales de la vida humana: la salud, los conocimientos y los niveles de vida dignos (Conceição, 2019).

Para el primer elemento referido a la salud, se tomó como medida la esperanza de vida al nacer. Se tuvo en cuenta dado que la esperanza de vida es crucial y va directamente relacionada con la salud y la nutrición.

Para el segundo, se tuvo en cuenta los años esperados de escolarización y promedio de años de escolarización, ya que esto es algo básico para iniciar la vida educativa de las personas y así llevar una vida productiva.

Por último, el tercer elemento hace referencia al manejo de los recursos para conseguir un nivel de vida digno, según el PNUD es probablemente el indicador más complicado de medir debido a que son necesarios datos sobre el acceso a la tierra, el crédito, el ingreso y otros recursos. El ingreso se crea a partir de las cifras reales del PIB ajustadas al poder monetario,

ya que son las cifras que brindan unas mejores aproximaciones de los datos (Molina Salazar & Pascual-García, 2015).

El propósito de este trabajo será entonces, sirviéndose de distintos métodos estadísticos, realizar una clasificación en dos grupos de los países en base a un nuevo *"Índice de Desarrollo Humano"*, contando con los indicadores provistos por los *ODS* para que posteriormente se pueda comparar con los datos del *IDH*; así como crear también una clasificación propia más concreta en varios grupos.

2. Objetivos:

Objetivo principal:

El propósito general de este trabajo es proponer un método multivariante alternativo al *IDH* para medir el Desarrollo humano y que divida a los países en dos grupos: los países que cuentan con más desarrollo y los que cuentan con menos. Para finalmente comparar estos resultados con el *ranking* oficial del *IDH*.

Objetivo secundario:

Una vez conseguida la agrupación en dos bloques buscada en el objetivo principal, se procederá a realizar una clasificación en un mayor número de grupos para así conseguir una agrupación más específica.

3. Material:

Para la realización de este estudio se compuso una base de datos, a través de los datos obtenidos de un repositorio público de la página oficial de la ONU, esta está compuesta por 19 variables correspondientes a 19 indicadores de los *ODS* repartidos entre los objetivos número 3 referente a la salud, el 4 referente a la educación y por último el 8 referente a la economía. Se eligieron estos indicadores debido a que forman parte de los elementos necesarios para calcular el *IDH* se vieron anteriormente.

En cuanto a los sujetos utilizados se escogieron 43 países del continente europeo: Albania, Armenia, Austria, Azerbaiyán, Bielorrusia, Bélgica, Bosnia y Herzegovina, Bulgaria, Croacia, Chipre, República Checa, Dinamarca, Estonia, Finlandia, Francia, Georgia, Alemania, Grecia, Hungría, Islandia, Irlanda, Italia, Letonia, Lituania, Luxemburgo, Malta, Montenegro, Países Bajos, Macedonia del Norte, Noruega, Polonia Portugal, Moldavia, Rusia, Serbia, Eslovaquia, Eslovenia, España, Suecia, Suiza, Ucrania y Reino Unido.

Las variables utilizadas son las siguientes:

- Objetivo número 3:
    - *X.3.1.1:* Tasa de mortalidad materna
    - *X.3.2.1:* Tasa de mortalidad de niños menores de 5 años
    - *X.3.2.2:* Tasa de mortalidad neonatal
    - *X.3.3.2:* Incidencia de la tuberculosis por cada 100.000 habitantes
    - *X.3.4.1:* Tasa de mortalidad atribuida a las enfermedades cardiovasculares,cáncer etc.
    - *X.3.4.2:* Tasa de mortalidad por suicidio
    - *X.3.5.2:* Consumo nocivo de alcohol per cápita
    - *X.3.7.2:* Tasa de fecundidad de las adolescentes
    - *X.3.8.1:* Cobertura de los servicios de salud esenciales
    - *X.3.9.3:* Tasa de mortalidad atribuida a intoxicaciones involuntarias

- Objetivo número 4:
    - *X.4.1.2:* Volumen de producción por unidad de trabajo
    - *X.4.2.2:* Tasa de participación en el aprendizaje organizado
    - *X.4.5.1***:** Índices de paridad

- Objetivo número 8:
    - *X.8.1.1:* Tasa de crecimiento anual del PIB real per cápita
    - *X8.2.1:* Tasa de crecimiento anual del PIB real por persona empleada
    - *X8.6.1:* Proporción de jóvenes desempleados
    - *X.8.5.2:* Tasa de desempleo
    - *X.8.10.1:* Número de sucursales de bancos comerciales por cada 100.000 adultos
    - *X.8.10.2:* Proporción de adultos con cuenta bancaria

Para el procesamiento de la base de datos y la organización de la misma, se utilizó la herramienta Excel. Por otro lado, para el análisis de los datos se utilizaron los *softwares* SPSS y R.

4. *Métodos:*

Tanto como para encontrar esa nueva medida para la clasificación según el Desarrollo Humano, como para la nueva clasificación de los países se utilizarán una serie de métodos estadísticos como son el Análisis de Componentes Principales para la reducción de la dimensión de las variables y diferentes métodos de *clustering* para la agrupación de los países.

A continuación, se procederá a una explicación teórica de los métodos que se han utilizado para el desarrollo del estudio.

4.1. *Álgebra:*

En primer lugar, se van a definir algunos términos algebraicos necesarios para entender las demostraciones que se realizarán posteriormente (Clapham, 2004)

Sea $B \in M_k(\mathbb{R})$ se dirá que:

I. La matriz B es invertible si existe otra matriz que se denominará como B^{-1} de tal forma que $BB^{-1} = B^{-1}B = I_k$, donde I_k denomina a la matriz identidad de tamaño k

II. B es una matriz simétrica si $B = B^t \; donde \; B^t$ es la matriz traspuesta de B

III. B es ortogonal si $B^t = B^{-1}$

IV. B es semidefinida positiva si para un vector $x \in \mathbb{R}^k$ se comprueba que $x^t Bx \geq 0$

V. Se dirá que $\lambda \in \mathbb{R}$, es un valor propio de B si existe un vector $v \in \mathbb{R}^k$ de tal forma que $Bv = \lambda v$. A v se denominará vector propio de B asociado a λ

VI. B se podrá diagonalizar si existen $P, D \in M_k(\mathbb{R})$ de tal forma que $B = PDP^{-1}$, donde las columnas pertenecientes a P están formadas por lo vectores propios $B \; y \; D$ es una matriz diagonal formada por los valores propios de B. Si P es ortogonal, se dirá que B se diagonaliza ortogonalmente.

4.2. *Varianza y covarianza:*

A continuación, se recordará el concepto de esperanza matemática y algunas de las propiedades que esta posee, las cuales se usarán cuando se introduzca la definición de varianza.

Dada una variable X discreta dentro de un espacio de probabilidad $(\Omega; S; Pr)$ con soporte $S = (x_1, x_2, \dots, x_k)$ y con una función de probabilidad $p(x)$, se definirá su esperanza matemática o media, denominada $E(x)$ o μ_x como:

$$E(X) = \sum_{i=1}^{+\infty} x_i p(x_i), \tag{4}$$

siempre que se cumpla lo determinado en *4*.

$$\sum_{i=1}^{+\infty} |x_i| p(x_i) < +\infty \quad (5)$$

En el caso de que la variable X fuera continua con una función de densidad de probabilidad $f(x)$, se denominará su media como:

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx \quad (6)$$

Siempre que

$$\int_{-\infty}^{+\infty} |x| f(x) dx < +\infty \quad (7)$$

La media cumple las siguientes propiedades:

Sean dos variables aleatorias $X\ e\ Y$, estando ambas definidas en el mismo espacio de probabilidad y $a, b \in \mathbb{R}$ se tendrá:

i. $E(a) = a$

ii. En el caso de que exista $E(x)\ entonces\ E(aX + b) = aE(x) + b$

iii. $E(X + Y) = E(x) + E(Y)$

Ahora se definirá el concepto de varianza y se enunciarán algunas las propiedades que esta cumple (Fisher, 1919).

Sea X una variable aleatoria con media μ_x se definirá como su varianza la fórmula *8*:

$$\sigma^2 \underset{x}{=} Var(X) = E((X - \mu x)^2) \quad (8)$$

Teniendo esta variable X con esperanza $\mu_x = E(x)$, una varianza $\sigma^2 = Var(x)$ y sean a, $b \in \mathbb{R}$. Entonces cumplirá las siguientes propiedades:

i. $Var(a) = 0$

ii. $Var(X) = E(X^2) - E^2(X)$

iii. $Var(aX + b) = a^2 Var(X)$

iv. $Si\ Z = (X - \mu x)/\sigma x, entonces\ E(Z) = 0\ y\ Var(Z) = 1$

A continuación, se definirá el concepto de covarianza, y se verán algunas de sus propiedades (Spiegel, 1992).

Tenemos dos variables X e Y con medias $\mu x\ y\ \mu y$. Se definirá entonces la covarianza entre $X\ e\ Y$ como sigue:

$$Cov(X,Y) = \sigma x,y = E((X - \mu x)(Y - \mu y)) \quad (9)$$

Cumpliendo esta las siguientes propiedades:

Siendo variables aleatorias $X, Y\ y\ Z$ con $a \in \mathbb{R}$

i. $Cov(X,Y) = E(XY) = E(X)E(Y)$
ii. $Var(X) = Cov(X,X)$
iii. $Cov(X,Y) = Cov(Y,X)$
iv. $Cov(X + Y,Z) = Cov(X,Z) + Cov(Y,Z)$
v. $Cov(aX,Y) = aCov(X,Y)$

A partir de la definición de covarianza se puede definir un concepto que aparecerá en demostraciones posteriores y este es el de las variables sin correlación. Dos variables estarán *incorrelacionadas* si $Cov(X,Y) = 0$.

4.3. Matriz de covarianzas:

Para expresar la variabilidad de los datos y la información referente a las relaciones lineales entre las variables se utiliza la *matriz de covarianzas* (Snedecor & Cochran, 1980).

Sea un vector aleatorio $X = (X_1, X_2, \dots, X_k)^t$ cuyo vector de medias es $\mu x = (\mu_1, \mu_2, \dots \mu_k)^t$ tal que $E(X^2) < \infty$ para todo $i = 1,2, \dots, k$ su *matriz de covarianzas* será:

$$S = Cov(X) = E\big((X - \mu_x)(X - \mu_x)^t\big) = \begin{pmatrix} \sigma_1^2 & \sigma_1^2 & \dots & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \cdots & \sigma_1^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1^2 & \sigma_1^2 & \cdots & \sigma_1^2 \end{pmatrix} \quad (10)$$

Donde $\sigma_i^2 = Var(X_i)\ y\ \sigma_{ij} = Cov\big(X_i, X_j\big)\ para\ i,j = 1,2, \dots, k$

La matriz de covarianza cumple las siguientes propiedades:

Se tiene un vector aleatorio $X = (X_1, X_2, \dots, X_k)^t$ y su matriz de covarianza S entonces se cumple que:

i. Es *semidefinida* positiva y simétrica
ii. Si A es una matriz de dimensión $n\ x\ k\ y\ B$ es una matriz de dimensión $n\ x\ 1$, ambas son reales y, por lo tanto, se cumple la igualdad *11*.

$$Y = AX + B\ , \quad (11)$$

Entonces:

$$Cov(Y) = ASA^t\ , \quad \boldsymbol{q.e.d} \quad (12)$$

En primer lugar, se demostrará (ii). Debido a que $E(Y = E(AX + B) = AE(X) + B = A\mu_x + B$ usando la definición de *matriz de covarianzas* se tendrá lo siguiente:

$$
\begin{aligned}
S &= Cov(Y) \\
&= Cov(AX + B) \\
&= E\big((AX + B - A\mu_x - B)(AX + B - A\mu_x - B)^t\big) \\
&= E(A(X - \mu_x)(X - \mu_x)^t A^t) \\
&= AE((X - \mu_x)(X - \mu_x)^t A^t \\
&= ASA^t
\end{aligned}
$$

$$\boldsymbol{q.e.d} \quad (13)$$

Ahora se demostrará (i). Tenemos que S es simétrica por su definición. Si se tiene la variable $Y = a^t X \; donde \; a = (a_1, a_2, \dots a_k)^t \in \mathbb{R}^k$ entonces:

$$Var(Y) = a^t S a \geq 0\,, \quad (14)$$

Por lo tanto, se tiene que S es semidefinida positiva.

4.4. Matriz de correlación:

Sean dos variables aleatoria *X* e *Y* con una varianza estrictamente positiva, se denominará *coeficiente de correlación* entre las dos variables a la fórmula definida en *15* (Kenney & Keeping, 1951).

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}\,, \quad (15)$$

Se definirá como *matriz de correlación* a la matriz de forma cuadrada y simétrica cuyos valores de la diagonal son unos y en el resto de los valores, los coeficientes de correlación entre las variables. Se denota como sigue en *15*:

$$R = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1k} \\ \rho_{21} & 1 & \dots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \dots & 1 \end{pmatrix}, \quad (16)$$

Si se estandarizan las variables:

$$X_i^* = \frac{X_i - \mu_i}{\sigma_i}\,, \quad (17)$$

para los valores de $i = 1,2, \dots, k$, se comprueba que la *matriz de covarianzas* S concuerda con la *matriz de correlaciones* de las variables sin estandarizar, R. Como $Var(X_i^*) = 1$ y se cumple lo definido en *18*,

$$
\begin{aligned}
Cov\left(X_i^*, X_j^*\right) &= E\left(X_i^* X_j^*\right) - E\left(X_i^*\right) \cdot E\left(X_j^*\right) \\
&= E\left(\left(\frac{X_i - \mu_i}{\sigma_i}\right)\left(\frac{X_j - \mu_j}{\sigma_j}\right)\right) \\
&= \frac{1}{\sigma_i \sigma_j} E\left(X_i X_j - X_i \mu_j - X_j \mu_i + \mu_i \mu_j\right) \\
&= \frac{1}{\sigma_i \sigma_j}\left(E\left(X_i X_j\right) - E\left(X_i \mu_j\right) - E\left(X_j \mu_i\right) + E\left(\mu_i \mu_j\right)\right) \\
&= \frac{1}{\sigma_i \sigma_j}\left(E\left(X_i X_j\right) - \mu_i \mu_j - \mu_j \mu_i + \mu_i \mu_j\right) \\
&= \frac{1}{\sigma_i \sigma_j}\left(E\left(X_i X_j\right) - \mu_i \mu_j\right) \\
&= \frac{\sigma_{ij}}{\sigma_i \sigma_j} \\
&= \rho_{ij}
\end{aligned}
$$

$\boldsymbol{q.e.d}$ (18)

se tiene por tanto que:

$$
S^* = \begin{pmatrix} 1 & \sigma_{12}^* & \dots & \sigma_{1k}^* \\ \sigma_{21}^* & 1 & \dots & \sigma_{2k}^* \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1}^* & \sigma_{k2}^* & \dots & 1 \end{pmatrix} = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1k} \\ \rho_{21} & 1 & \dots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \dots & 1 \end{pmatrix} = R \qquad (19)
$$

Una consecuencia directa de esto es lo siguiente:

La *matriz de correlación* es *semidefinida* positiva. Como se comprobó anteriormente en la *matriz de covarianzas* y en este caso al R coincidir con S^* esta será *semidefinida* positiva.

4.5. Análisis de Componentes Principales:

La técnica del *Análisis de Componentes Principales (ACP)* fue desarrollada inicialmente por *Pearson* en el siglo XIX y a posteriori estudiadas por *Hotelling* a principios del siglo XX. A pesar de esto, no fue hasta la aparición de los computadores no se empezó a popularizar esta técnica (Shlens, 2005)

Para poder estudiar las relaciones que existen entre n variables correlacionadas se debe transformar el conjunto inicial en otro de variables nuevas sin correlación, llamado conjuntos de componentes principales.

Estas variables nuevas son combinaciones lineales de las anteriores y se construyen siguiendo el orden de relevancia referido a su variabilidad total. De este modo, se quieren buscar $m < p$ variables que sean combinaciones lineales de las variables iniciales y que estén sin correlacionar, intentando recoger la mayor parte posible de la información (Navarro-Céspedes, Casas-Cardoso, & González-Rodríguez, 2010).

Cabe destacar que, si las variables originales no tienen correlación, entonces la realización del *ACP* carecerá de sentido (Marín, J.M. 2015).

4.5.1. Cálculo general de las componentes principales:

Para el cálculo de las componentes se tiene un vector aleatorio $X = (X_1, X_2, \dots, X_k)$ con dimensión k con *matriz de covarianzas* S definida positiva. Entonces las *componentes principales (CPs)* estarán definidas como se muestra en *20*:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{pmatrix} = P^t X = \begin{pmatrix} p_{11} & p_{21} & \cdots & p_{k1} \\ p_{12} & p_{21} & \cdots & p_{k2} \\ \vdots & \ddots & \ddots & \vdots \\ p_{1k} & p_{2k} & \cdots & p_{kk} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix} \tag{20}$$

dónde se tiene una matriz ortogonal P que diagonaliza S, siendo $D = (\lambda_1, \lambda_2, \dots, \lambda_k)$ esta matriz diagonal con $\lambda_1 > \lambda_2 > \cdots > \lambda_k > 0$ (Shlens, 2005).

En primer lugar, se verá que al ser S simétrica se diagonaliza ortogonalmente y también, como es definida positiva sus valores propios verificarán $\lambda_1 > \lambda_2 > \cdots > \lambda_k > 0$ si ahora se define que $Y = P^t X$ se tiene:

$$Cov(Y) = P^t S P = D \tag{21}$$

Por lo tanto:

- $Cov(Y_i, Y_j) = 0 \; si \; i \neq j$
- $Var(Y_i) = \lambda_i > 0 \; para \; todo \; i = 1,2, \dots, k$
- Las filas P^t forman una base ortonormal de $\mathbb{R}^k$ debido a que son las mismas que las columnas de P, las cuales están formadas por los vectores propios $p_i = (p_1, p_2, \dots, p_{ki})^t$ de S

Ahora se observa que:

$$Y_1 = p_{11} X_1 + p_{21} X_2 + \cdots + p_{k1} X_k \tag{22}$$

Esta es la primera componente principal. Sea $a_1 = (a_{11}, a_{12}, \dots a_{1k})^t \in \mathbb{R}$ con $\|a_1\| = 1$ de tal forma que:

$$a_1^t X = a_{11} X_1 + a_{12} X_2 + \cdots + a_{1k} X_k \tag{23}$$

Debido a que los vectores propios forman una base de $\mathbb{R}^k$ existirán escalares, no todos nulos, $c_1, c_2, \dots, c_k$ tal que:

$$a_j = c_1 p_1 + c_2 p_2 + \cdots + c_k p_k \tag{24}$$

Por lo tanto, su varianza tendrá el valor de:

$$Var(a_1^t X) = a_1^t S a_1$$

$$= \left(\sum_{i=1}^{k} c_i p_i^t \right) S \left(\sum_{i=1}^{k} c_i p_i \right)$$

$$= \left(\sum_{i=1}^{k} c_i p_i^t\right)\left(\sum_{i=1}^{k} c_i S p_i\right) \tag{25}$$

Ya que cada p_i es un vector propio de S se puede escribir la expresión anterior como sigue:

$$\left(\sum_{i=1}^{k} c_i p_i^t\right)\left(\sum_{i=1}^{k} c_i V p_i\right) = \left(\sum_{i=1}^{k} c_i p_i^t\right)\left(\sum_{i=1}^{k} c_i \lambda_i p_i\right) = \sum_{i,j=1}^{k} c_i c_j \lambda_j p_i^t p_j \tag{26}$$

Y al tener que $p_i^t p_j = 0 \; si \; i \neq j$ se tiene que:

$$Var(a_1^t X) = \sum_{i,j=1}^{k} c_i c_j \lambda_j p_i^t p_j = \sum_{i=1}^{k} c_i^2 \lambda_i \tag{27}$$

Como también se quiere que se cumpla que $\|a_1\| = 1$ se tiene que verificar lo siguiente:

$$a_1^t a_1 = \left(\sum_{i=1}^{k} c_i p_i^t\right)\left(\sum_{i=1}^{k} c_i p_i\right) = \sum_{i,j=1}^{k} c_i c_j \lambda_j p_i^t p_j = \sum_{i=1}^{k} c_i^2 = 1 \tag{28}$$

$$\boldsymbol{q.e.d}$$

Por lo que si $c_1^2 = 1, c_2 = c_3 = \cdots = c_k = 0$ la varianza de Y_1, que si se recuerda coincide con λ_1, será máxima como consecuencia de *29*:

$$Var(Y_1) = \sum_{i=1}^{k} c_i^2 \lambda_1 \geq \sum_{i=1}^{k} c_i^2 \lambda_i = Var(a_1^t X) \tag{29}$$

Debido a esto, $Y_1 = \pm p_1^t X$ es la componente principal número 1, que puede ser no única si $\lambda_1 = \lambda_2$.

Ahora se supondrá que por inducción que $Y_1, Y_2, \ldots, Y_{j-1} \; son \; las \; j-1$ primeras *CPs.*

En primer lugar, como tiene que verificarse que $Cov\left(a_j^t X, Y_i\right) = 0$ para $i = 1,2, \ldots, j-1$ debe verificar lo siguiente:

$$Cov\left(a_j^t X, Y_i\right) = Cov\left(a_j^t X, p_i^t X_i\right)$$

$$= E\left(\left(a_j^t X - E\left(a_j^t X\right)\right)\left(p_i^t X - E(p_i^t X)\right)^t\right)$$

$$= E\left(\left(a_j^t X - a_j^t E(X)\right)\left(p_i^t X - p_i^t E(X)\right)^t\right)$$

$$= E(a_j^t\left(X - E(X)\right)\left(X - E\left((X)\right)^t p_i\right)$$

$$= a_j^t E\left(\left(X - E(X)\right)\left(X - E(X)\right)^t\right) p_i$$

$$= a_j^t S p_i$$

$$= 0$$

$$\boldsymbol{q.e.d} \tag{30}$$

Ahora como los *vectores propios* forman una base, existirán entonces escalares no nulos $c_1, c_2, \dots, c_k$ de tal forma que:

$$a_j = c_1 p_1 + c_2 p_2 + \cdots + c_k p_k \tag{31}$$

Por lo que:

$$Cov\left(a_j^t X, Y_i\right) = a_j^t S p_i = a_j^t \lambda_i p_i = \lambda_i a_j^t p_i = \lambda_i \left(\sum_{v=1}^{k} c_v p_v^t \right) p_i = \lambda_i \sum_{v=1}^{k} c_v p_v^t \, p_i = 0 \tag{32}$$

Y debido a que $p_i^t p_l = 0 \; si \; i \neq j$, ya que son ortonormales, se tiene lo que sigue:

$$Cov\left(a_j^t X, Y_i\right) = \lambda_i \sum_{v=1}^{k} c_v p_v^t \, p_i = \lambda_i c_i p_i^t = \lambda_i c_i = 0 \tag{33}$$

Lo que con lleva que $c_1 = c_2 = \cdots = c_{j-1} = 0$ ya que $\lambda_1, \lambda_2, \dots, \lambda_{j-1} > 0$. Además,

$$Var\left(a_j^t X\right) = \sum_{i,v=1}^{k} c_i c_v \lambda_v p_v^t \, p_i \sum_{i,v=1}^{k} c_i^2 \lambda_i \tag{34}$$

Por esto la varianza será máxima si $c_j = 1 \; y \; c_i = 0$ para todo $i > j$ debido a que:

$$Var\left(Y_j\right) = \lambda_j = \sum_{i,v=1}^{k} c_i^2 \lambda_j \geq \sum_{i,v=1}^{k} c_i^2 \lambda_i = Var\left(a_j^t X\right) \tag{35}$$

$$\boldsymbol{q.e.d}$$

Por ello $Y_j = \pm p_j^t X$ es una componente principal *j-ésima*.

Como consecuencia directa de este resultado se tiene lo siguiente:

Siendo $X = (X_1, X_2, \dots X_k)^t$ un vector aleatorio de dimensión k, cuya *matriz de covarianzas* S está definida positiva. Si S tiene unos *valores propios* que verifican que $\lambda_1 > \lambda_2 > \cdots > \lambda_k > 0$ quiere decir que las *CPs* serán únicas, salvo el signo.

Esto se demuestra sabiendo que, si todos los *valores propios* son distintos, también los respectivos *vectores propios* lo serán, lo que conlleva que las *CPs* son únicas salvo signo.

En los resultados anteriores se supuso que los valores propios de S son todos positivos, pero no quiere decir que esto siempre sea así. Si S es definida *semipositiva*, algunos de sus *valores propios* serán igual a 0, esto sería debido a que alguna de las variables iniciales sería combinación lineal de $p < k$ variables. Lo que llevaría a ignorar aquellas componentes que

están asociadas a los *valores propios* nulos debido a que estas no aportarían ninguna información ya que su varianza sería igual a 0.

4.5.2. Propiedades de las componentes:

Una vez que se ha explicado el método del cálculo de las *CPs* del vector aleatorio $X = (X_1, X_2, \dots X_k)^t$ con la *matriz de covarianzas* S, se procederá a estudiar sus propiedades (Lozares-Colina & López-Roldán, 1991).

Las *CPs* mantienen su varianza inicial, lo que quiere decir que la suma de las varianzas de las variables originales es la misma que la suma de las varianzas de las componentes

Esto se demosrará de la siguiente forma:

Como $Var(Y_i) = \lambda_i$ y la suma de las varianzas originales coincide con $traza(S)$ ya que si $A \in M_k(\mathbb{R})$ y $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ son sus *valores propios*, entonces $\sum_{i=1}^{p} \lambda_i = traza(A)$. Por lo tanto, tras esta demostración se cumple que:

$$\sum_{i=1}^{k} Var(X_i) = traza(S) = \sum_{i=1}^{k} \lambda_i = \sum_{i=1}^{k} Var(Y_i) \tag{36}$$

Como consecuencia de esto se tiene que la proporción de varianza de la componente *j-ésima* vale lo determinado en *37*:

$$\frac{Var(Y_j)}{\sum_{i=1}^{k} Var(Y_i)} = \frac{\lambda_j}{\sum_{i=1}^{k} \lambda_i} \tag{37}$$

Esto se demuestra aplicando la proposición anterior, y se tiene que:

$$\frac{Var(Y_j)}{\sum_{i=1}^{k} Var(Y_i)} = \frac{\lambda_j}{traza(S)} = \frac{\lambda_j}{\sum_{i=1}^{k} \lambda_i} \tag{38}$$

De esta forma se puede definir la cantidad de información que se conserva en cada componente de la siguiente manera:

$$I_j = 100\% * \frac{\lambda_j}{\sum_{i=1}^{k} \lambda_i} \tag{39}$$

Una vez que se ha determinado la varianza de cada componente se dispondrá a estudiar la relación que existen entre las componentes y las variables iniciales

Las componentes principales verifican que:

$$Cov(X, Y) = PD \tag{40}$$

$$Corr(X, Y) = diag(S)^{-1/2} PD^{1/2} \tag{41}$$

Donde $diag(S) = diag(\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2)$.

Se puede demostrar si aplicando la definición de correlación entre $X_i e\, Y_j$, obteniendo lo siguiente:

$$Corr(X_i, Y_j) = \frac{Cov(X_i, Y_j)}{\sqrt{Var(X_i)Var(Y_j)}} = \frac{Cov(X_i, Y_j)}{\sqrt{\sigma_i^2 \lambda_j}} = \frac{Cov(X_i, Y_j)}{\sigma_i \lambda^{1/2}} = \sigma_i^{-1} Cov(X_i, Y_j) \lambda_j^{-1/2} \quad (42)$$

Por lo tanto $Corr(X, Y) = diag(S)^{-1/2} Cov(X, Y) D^{1/2}$. Aplicando la propiedad (v) de la covarianza se tiene que:

$$Cov(X, Y) = Cov(X, P^t X) = Cov(X, X)P = SP \quad (43)$$

Y debido a que se cumple que $S = PDP^t$ queda la expresión anterior como:

$$Corr(X, Y) = PD \quad (44)$$

Y entonces:

$$Corr(X, Y) = diag(S)^{-1/2} PDD^{-1/2} = diag(S)^{-1/2} PD^{1/2} \qquad \boldsymbol{q.e.d} \quad (45)$$

Ahora si se quiere establecer la covarianza o la correlación entre una variable X_i y una componente Y_j se tiene como resultado lo siguiente, lo cual es consecuencia directa de lo anteriormente demostrado.

Por lo tanto, se cumple lo que sigue:

$$Cov(X_i, Y_j) = p_{ij} \lambda_j \quad (46)$$

$$Corr(X_i, Y_j) = \frac{p_{ij}}{\sigma_i} \lambda_j \quad (47)$$

Estos resultados permiten definir la *matriz de saturaciones*

Sean Y las *CPs* pertenecientes a X, se llamará *matriz de saturaciones* a $A = Corr(X, Y)$

Esta matriz verifica que:

$$AA^t = R \quad (48)$$

Esto se demostrará de la siguiente manera

$$\begin{aligned} AA^t &= diag(S)^{-1/2} PD^{1/2} \left(diag(S)^{1/2} PD^{1/2} \right)^t \\ &= diag(S)^{1/2} PD^{1/2} D^{1/2} P^t diag(S)^{-1/2} \\ &= diag(S)^{-1/2} PDP^t diag(S)^{-1/2} \\ &= diag(S)^{-1/2} S diag(S)^{-1/2} \\ &= R \end{aligned}$$

$\boldsymbol{q.e.d}$ (49)

4.5.3. Cálculo de las componentes principales a través de la matriz de correlaciones:

Si se opta por calcular las *CPs* a partir de la *matriz de covarianzas* S las variables que tengan mayor varianza serán aquellas que tendrán más influencia (Gurrea, 2000). Esto puede ser un problema cuando las variables tienes diferentes escalas. Para solventar ese problema se trabajará con las variables escaladas definidas en *50*:

$$X_i^* = \frac{X_i - \mu_{xi}}{\sigma_i} \tag{50}$$

Estas variables escaladas tendrán media nula y varianza 1, como se demostró anteriormente la *matriz de correlaciones* R, de las variables originales coincidirás con la m*atriz de covarianzas* S^* de las variables escaladas, por lo tanto, se tiene que:

Las *CPs* basadas en $X^* = (X_1^*, X_2^*, \dots, X_k^*)^t$ están determinadas por lo siguiente:

$$\tilde{Y} = \begin{pmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \\ \vdots \\ \tilde{Y}_k \end{pmatrix} = \tilde{P}^t X = \begin{pmatrix} \tilde{p}_{11} & \tilde{p}_{21} & \dots & \tilde{p}_{k1} \\ \tilde{p}_{12} & \tilde{p}_{22} & \dots & \tilde{p}_{k2} \\ \vdots & \ddots & \ddots & \vdots \\ \tilde{p}_{1k} & \tilde{p}_{2k} & \dots & \tilde{p}_{kk} \end{pmatrix} = \begin{pmatrix} X_1^* \\ X_2^* \\ \vdots \\ X_k^* \end{pmatrix} \tag{51}$$

Donde la matriz $\tilde{P}$ es ortogonal diagonaliza a R sabiendo que $R = S^*$ y $siendo\, \tilde{D} = diag(\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_k)$ esta matriz diagonal tendrá $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \dots \tilde{\lambda}_k > 0$.

De igual forma que cuando se vio el cálculo con la *matriz de covarianzas* se sabe que siendo $X^* = (X_1^*, X_2^*, \dots, X_k^*)$ una variable aleatoria de dimensión k cuya *matriz de correlación* R que está definida positiva como se demostró anteriormente. Si R tiene unos *valores propios* que verifican que $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \tilde{\lambda}_k > 0$ quiere decir que las *CPs* serán únicas, salvo el signo.

Ahora centrándose en el estudio de las propiedades de las componentes que se han obtenido a partir de C, las cuales son consecuencia directa de los resultados de la sección *4.4.2*

Las *CPs* que han sido obtenidas a través de la matriz de correlaciones mantienen la varianza inicial de las variables escaladas, es decir:

$$\sum_{i=1}^{k} Var(\tilde{Y}_i) = k \tag{52}$$

Como consecuencia de esto se tiene que:

$$\frac{Var(\tilde{Y}_j)}{\sum_{i=1}^{k} Var(\tilde{Y}_j)} = \frac{\tilde{\lambda}_j}{k} \tag{53}$$

Una vez que se determina la varianza de cada componente, es preciso analizar la relación existente entre las componentes y las variables originales. Las *CPs* quese han obtenido a partir de la *matriz de correlaciones* verifican:

$$Corr(X, \tilde{Y}) = \tilde{P}\tilde{D}^{1/2} \tag{54}$$

Demostración: en primer lugar, se puede ver que es posible escribir:

$$\tilde{Y} = \tilde{P}^t X^* = \tilde{P} diag(S)^{-1/2}(X - \mu) \tag{55}$$

Se cumple que:

$$Cov(X^*, \tilde{Y}) = Cov(X^*, X^*)\tilde{P} = Var(X^*)\tilde{P} = C\tilde{P} = \tilde{P}\tilde{D} \tag{56}$$

Y por tanto se verifica también que:

$$Corr(X, \tilde{Y}) = Corr(X^*, \tilde{Y}) = Cov(X^*, \tilde{Y})\tilde{D}^{-1/2} = \tilde{P}\tilde{D}\tilde{D}^{-1/2} = \tilde{P}\tilde{D}^{-1/2} \quad \boldsymbol{q.e.d} \tag{57}$$

En las condiciones anteriores se cumple lo siguiente:

$$Corr(X_i \tilde{Y}_j) = \tilde{p}_{ij}\tilde{\lambda}_j^{1/2} \tag{58}$$

Tras esto se puede entonces definir la *matriz de saturaciones* como ya se hizo en la sección *4.4.2*.

Sean $\tilde{Y}$ las *CPs* obtenidas a través de la matriz de correlaciones de X, se define como *matriz de saturaciones* a $\tilde{A} = Corr(X, \tilde{Y})$.

Esta matriz verifica que:

$$\tilde{A}\tilde{A}^t = R \tag{59}$$

$$\tilde{A}^t\tilde{A} = \tilde{D} \tag{60}$$

Demostración:

$$\tilde{A}\tilde{A}^t = \tilde{P}\tilde{D}^{\frac{1}{2}}\left(\tilde{P}\tilde{D}^{\frac{1}{2}}\right)^t$$

$$= \tilde{P}\tilde{D}^{\frac{1}{2}}\tilde{D}^{\frac{1}{2}}\tilde{P}^t$$

$$= \tilde{P}\tilde{D}\tilde{P}^t$$

$$= S^*$$

$$= R$$

$$\tilde{A}^t\tilde{A} = \left(\tilde{P}\tilde{D}^{\frac{1}{2}}\right)^t \tilde{P}\tilde{D}^{\frac{1}{2}}$$

$$= \widetilde{D}^{1/2}\tilde{P}^{t}\tilde{P}\widetilde{D}^{1/2}$$

$$= \widetilde{D}^{1/2}\widetilde{D}^{1/2}$$

$$= \widetilde{D} \tag{61}$$

$$\boldsymbol{q.e.d}$$

4.5.4. Interpretación de las componentes principales:

A continuación, se definirán las cargas y las puntuaciones (Carmona, 2014).

Sean $Y = P^{t}X$ las componentes principales pertenecientes a un vector, entonces se definirá como:

i. Las cargas de la componente *j-ésima*, como aquellos valores pertenecientes a la fila *j-ésima* de una matriz P^{t}.

ii. Las puntuaciones en la componente *j-ésima* de un determinado individuo, al resultado de la sustitución de los valores que posee éste en cada variable aleatoria en las componentes.

Cuando se encuentra una correlación positiva entre todas las variables, es posible escoger los valores propios de tal manera que las cargas de la primera componente sean positivas, por lo tanto, se puede interpretar como una media ponderada de las variables y servirá para descubrir a aquellos individuos que poseen valores grandes en todas las variables.

Cuando todas las variables que se someten a estudio no tengan correlación o la tengan, pero esta sea muy baja en valor absoluto entre sí, no tendrá mucha utilidad realizar el *ACP* ya que éstas serían escogidas al azar.

4.5.5. Selección del número de componentes:

Una vez calculadas las componentes principales, el siguiente paso al que enfrentarse es el ver cuántas de esas componentes se seleccionarán para el estudio (Peres-Neto, Jackson, & Somers, 2005).

Aunque para esto hay numerosos métodos, en este trabajo se han utilizado los dos siguientes:

- Regla del codo: Este método consiste en realizar un gráfico de los *valores propios,* λ_i frente a i, para la selección hay que fijarse en cuando la gráfica forme aproximadamente una línea recta. La idea por la cual se llama método del codo es porque en esta forma de "codo" donde una i a partir de la cual, los λ_i empiezan a formar una línea recta. Este método también es conocido como gráfico de sedimentación. (Syakur, Khotimah, Rochman, & Satoto, 2018).

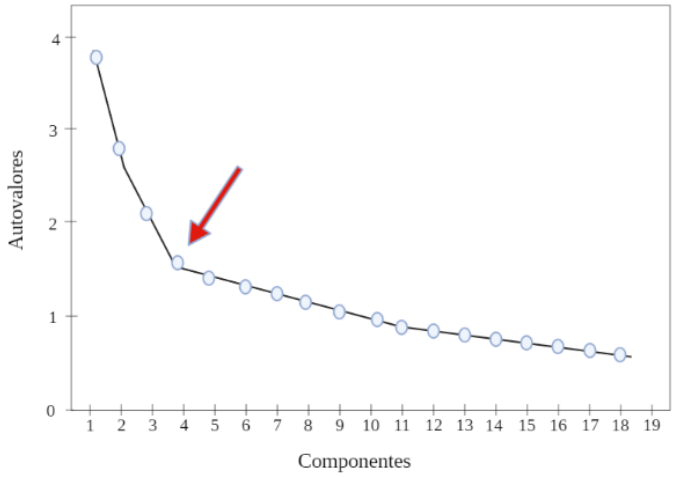


Ilustración 1. Ejemplo gráfico de sedimentación / Fuente: Elaboración propia con Biorender

- Criterio de *Kaiser*: Al obtener las componentes principales a través de la *matriz de correlaciones* equivale a que las varianzas de las variables observadas sea 1. Lo que quiere decir que, si una componente tiene una varianza menos que 1, esta explica menos variabilidad que una variable observable (Kaiser, 1960).

  Por lo tanto, con el criterio de *Kaiser* se elegirán las m componentes que cumplan que $\lambda_m \geq 1$ donde $\lambda_1 \geq \ldots \lambda_p$ son los valores propios de R que también son las varianzas de las componentes (Méndez-Peña & Sánchez-Gutiérrez, 2014).

4.6. Análisis de Cluster:

Los métodos de *Clustering* (o análisis de conglomerados) son técnicas de Análisis Exploratorio de datos utilizadas para resolver problemas de clasificación. Su principal objetivo consiste en ordenar objetos (cosas, animales, personas, variables, etc.) en diferentes grupos (*clusters* o conglomerados) de tal forma que el grado de homogeneidad entre los miembros de los mismos *cluster* sea mayor que la que existe entre los miembros de *clusters* diferentes. Cada *cluster* viene definido como la clase a la que pertenecen sus miembros (Blashfield & Albenderfer, 1978).

Estos métodos de asociación permiten descubrir relaciones y estructuras en los datos que a simple vista no son evidentes pero que una vez encontradas, estas pueden resultar muy útiles. Los resultados que se obtienen tras realizar un *Análisis de Cluster* pueden contribuir a establecer una definición formal de un esquema clasificatorio, de igual forma que una taxonomía para un grupo de objetos, también para poder describir poblaciones, a asignar nuevos individuos etc (Xu & Wunsch, 2008).

Los métodos de agrupación que se verán entran dentro del aprendizaje no supervisado, ya que consisten en la búsqueda de la división óptima del conjunto de entrada en función de las similitudes y diferencias entre sus ejemplos, es decir, busca la estructura principal de los datos (Celebi & Aydin, 2016). Normalmente el conjunto de datos empleado en este tipo de modelos no cuenta con una variable respuesta Y a diferencia de los algoritmos predictivos. Por tanto, se busca la extracción de patrones de la base de datos sin tener una variable que pueda contrastar el ajuste (por eso se denomina no supervisado) (García & Gómez, 2006). Esta división puede ser simplificada mediante técnicas de reducción de la dimensionalidad (por ejemplo, el *ACP*). Algunos de los métodos no supervisados y en los que se centrará el trabajo son: algoritmo *k-means* y *agrupamiento jerárquico* (Kettenring, 2006).

4.6.1. Algoritmo k-means:

El algoritmo *k-means*, que fue creado por *MacQueen* en 1967, es el algoritmo de *clustering* más popular y utilizado, debido a que su aplicación es muy simple y eficiente. Es un proceso de clasificación simple de un conjunto de objetos en un k número determinado de *clusteres*, que son determinados a priori (MacQueen, 1967).

El nombre de este algoritmo se debe a que cada uno de los *clusters* es representado por la media o media ponderada de sus puntos, es decir, por su centroide. Esta representación por centroides tiene como ventaja que se tiene un significado gráfico y estadístico de forma inmediata. Por lo tanto, cada *cluster* se caracteriza por su centro (García & Gómez, 2006).

En la *ilustración 2* se puede observar cómo están distribuidos los datos antes y después de aplicar el algoritmo

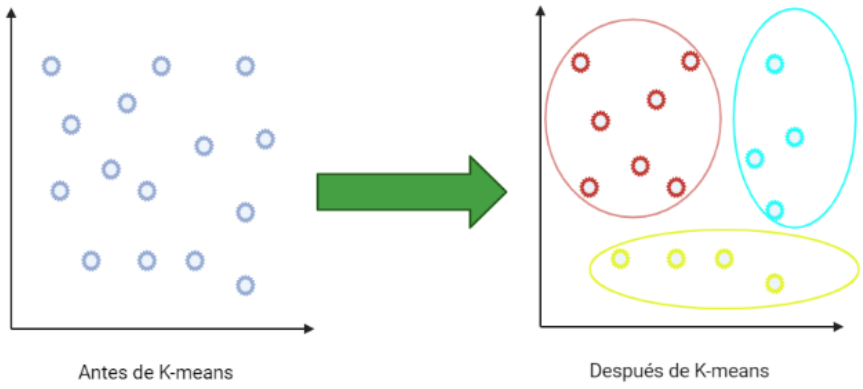


Ilustración 2. Ejemplo aplicación K-means / Fuente: Elaboración propia con Biorender

Este algoritmo consta de tres pasos (Wu et al., 2008):

I. *Iniciación:* una vez que se escoge el número de grupos k, se procede a establecer los k centroides en el espacio de los datos. Esto se puede hacer por ejemplo de forma

aleatoria.

II. *Asignación de los objetos a los centroides:* cada objeto de los datos se asigna a su centroide más cercano.

III. *Actualización de los centroides:* la posición del centroide de cada grupo se va actualizando y se establece como nuevo centroide la posición media de los objetos pertenecientes a dicho grupo.

Los pasos 2 y 3 se repetirán hasta que los centroides no se muevan, o se muevan por debajo de una distancia establecida.

El algoritmo *k-means* sirve para resolver un problema de optimización, siendo esta función de optimizar la suma de las distancias cuadráticas de cada objeto al centroide de su *cluster* (Ochoa-Reyes, Orellana-García, Sánchez-Corales, & Davila-Hernández, 2014).

Estos objetos son representados mediante vectores reales de dimensiones $(x_1, x_2, \dots, x_n)$ y el algoritmo *k-means* se encarga de construir k grupos en los cuales se minimiza la suma de distancias de los objetos, dentro de cada grupo $S = \{S_1, S_2, \dots, S_k\}$ a su centroide. Este problema se puede representar de la siguiente manera

$$\min_S E\,(\mu_i) = \min_S \sum_{i=1}^{k} \sum_{x_j \,\epsilon\, S_i} \left\| x_j - \mu_i \right\|^2 \tag{62}$$

Donde S es el conjunto de datos cuyos elementos son los objetos x_j que están representados por vectores, donde cada uno de esos vectores representan una característica. Por lo tanto, se tendrán k grupos con su respectivo centroide μ_i.

En cada una de las actualizaciones de los centroides, desde un punto de vista matemático, se impondrá la condición necesaria de extremo a la función $E(\mu_i)$, que para la función cuadrática *62* es

$$\frac{\partial E}{\partial \mu_i} = 0 \; ; \; \mu_i^{(t+1)} = \frac{1}{S_i^{(t)}} \sum_{x_j \,\epsilon\, S_i^t} x_j \tag{63}$$

Y se toma el promedio de los elementos pertenecientes a cada grupo como nuevo centroide (Gonzalo, 2019).

Como inconveniente de este algoritmo se puede encontrar entre otros la elección de k, que es el número de *clusters* en los que se dividirá el conjunto de datos inicial (Moya, 2016). Para la elección de este número no existe ningún criterio objetivo, no obstante, existen distintos métodos que serán de ayuda para tomar esta decisión. Entre ellos se encuentran el método del codo, el método de la silueta y el método de estadística de la brecha, como ya se explicó el método del codo para la elección de las componentes principales, para la elección de *clusters* se explicarán y se utilizarán los otros dos.

- *Método de la silueta*: En este análisis se mide la cantidad del agrupamiento o *clustering.* Se observa la separación entre los *clusters* y se indica cuál es la distancia de cada punto de un cluster a los puntos de los *clusters* vecinos. Esta medida se encuentra en el rango $[-1,1]$, un valor alto indica un *clustering* óptimo.

Los coeficientes que se encuentran más cercanos a $+1$ indican que la observación se encuentra lejos de los *clusters* vecinos, un valor de 0 nos muestra que esta está muy cerca o en la frontera entre los *clusters*, y, por último, los valores negativos indicarán que tal vez esas muestras estén en el cluster erróneo.

Este método calcula la esperanza de los coeficientes de silueta de todas las observaciones para varios valores de k, el número óptimo de *clusters* será aquel que consiga maximizar la media de los coeficientes de silueta para un rango en concreto de valores de k.

Este coeficiente de la silueta se calculará con la siguiente fórmula:

$$S = \frac{b - a}{\max(a, b)} \tag{64}$$

Siendo a el valor de a la distancia media *inter-cluster* y b la distancia media a las observaciones del cluster más próximo (Boutsidis & Magdon-Ismail, 2013).

- *Método estadístico de la brecha*: este método estadístico se encarga de comparar, para distintos valores de k, la varianza *intra-cluster* observada frente al valor esperado conforme a una distribución de referencia. El número óptimo de *clusters* que se estima será el valor de k con el que se maximiza el estadístico, es decir, se encarga de encontrar el valor de k con el que se logra conseguir una estructura de *clusters* que esté lo más alejada posible de una distribución uniforme. Este estadístico puede aplicarse a cualquier método de *clustering* (Tibshirani, Walther, & Hastie, 2001).

4.6.2. Agrupamiento jerárquico:

El agrupamiento jerárquico es conocido en la minería de datos como un método de análisis de grupos puntuales, el cual intenta generar una jerarquía de grupos. Este a diferencia del método de *k-mens*, no necesita de una especificación previa del número de *clusters* (Hastie, Tibshirani, & Friedman, 2009). Este método se puede realizar sobre los datos originales o sobre una matriz de distancia. Para el cálculo de esta matriz es necesario elegir una distancia, generalmente se utilizará la distancia euclídea y sólo se utilizará otra si se presenta una justificación teórica (Press, Teukolsky, Vetterling, & Flannery, 2007). La fórmula de la distancia euclídea es la siguiente:

$$d(G, H) = \sqrt{\sum (x_i - x_i')^2} \tag{65}$$

Según sea la dirección que tome este algoritmo para llevar a cabo el agrupamiento, encontramos dos tipos de *clustering* jerárquico (Espinel, 2015):

- *Aglomerativos*: esta agrupación parte desde cada elemento individual, de tal forma que al comienzo cada individuo se encuentra en un grupo independiente y se va van fusionando los k grupos más cercanos. Repetidamente van teniendo lugar estas fusiones de *clusters* originando una jerarquía de resultados hasta que finalmente sólo

quede un único grupo que comprenda a todos los objetivos del conjunto.

- *Divisivos:* al contrario que en los aglomerativos, se partirá de un único grupo que comprenda a todos los individuos para que posteriormente se vayan dividiendo en *clusters* de menor tamaño.

Para este método también es necesario elegir una manera de medir la distancia entre los grupos. Para hallar esta medida se pueden emplear 5 criterios de vinculación entre *clusters* (Vilà-Baños, Rubio-Hurtado, Berlanga, & Torrado-Fonseca, 2014) (Bridges, 1966):

- *Enlace completo*: en este método se eligen los dos puntos más dispares de cada *cluster* y se mide la distancia. Sean $G\ y\ H$ dos grupos distintos, la distancia entre estos será:

$$d_{CL}(A,B) = \max_{\substack{i\in A \\ i'\in B}} d_{ii}{}' \tag{66}$$

- *Enlace simple*: al contrario que en el *enlace completo*, esta vez los puntos que se toman son aquellos que tienen una distancia mínima.

$$d_{SL}(A,B) = \min_{\substack{i\in A \\ i'\in B}} d_{ii'} \tag{67}$$

- *Distancia entre medias*: la distancia entre los dos grupos se calculará como la distancia entre los centroides de cada grupo
- *Distancia promedio entre pares*: es el promedio de todas las distancias que existen entre todos los pares de puntos. Siendo $G\ y\ H$ dos *clusters* distintos y $N_G\ y\ N_H$ el número de observaciones que existen respectivamente en cada uno de ellos:

$$d_{GA}(A,B) = \frac{1}{N_A N_B}\sum\sum d_{ii'} \tag{68}$$

- *Método de Ward*: en este caso la selección del par de grupos a combinar en cada iteración se basará en el valor óptimo de una función objetivo determinada por el analista. Un caso particular de esta técnica es el método *Ward's minumum variance* en el que el objetivo es minimizar la suma total de la varianza entre los *cluster* (Ward Jr, 1963).

Para llevar a cabo la representación gráfica del análisis jerárquico lo más común es utilizar un diagrama en forma de árbol invertido denominado *dendrograma*, gráficos que representan a los individuos del conjunto anidados en jerarquías y cuyas líneas verticales o ramas representan las divisiones o fusiones realizadas en cada etapa del algoritmo, cuanto más cerca a la base del *dendrograma* ocurrirá una mayor homogeneidad entre individuos (Schonlau, 2004).

Una vez obtenido el *dendrograma*, es necesario establecer el número de grupos que existen y que individuos forman parte de ellos. Para esto se trazará una línea horizontal y discontinua a la altura en la que se desee hacer la poda, esta elección se realizará a criterio del investigador

y se elegirá la que se crea que hace una división más oportuna (Hand, Mannila, & Smyth, 2001). Un ejemplo de *dendograma* se puede observar en la *Ilustración 3*.

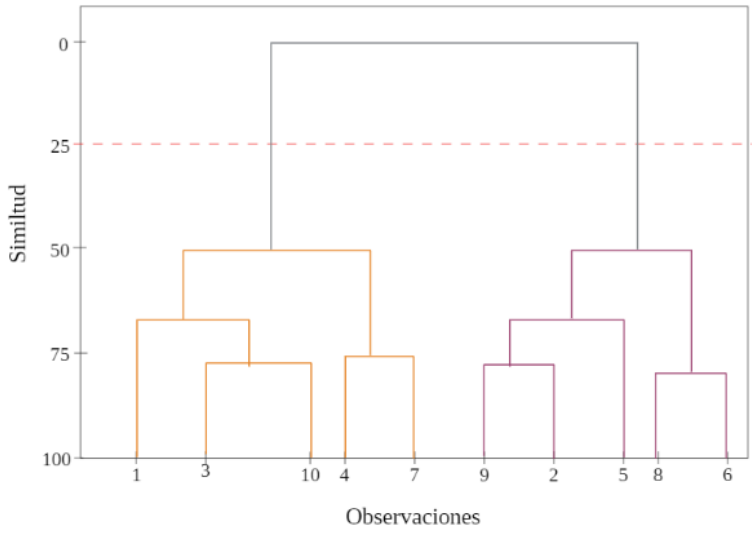


Ilustración 3. Ejemplo de dendrograma / Fuente: Elaboración propia con Biorender

4.7. Cálculo del IDH:

La construcción de un modelo es siempre una tarea compleja. Resulta muy difícil incluir todas las variables en un solo modelo, por lo que es necesario el poder elegir aquellas que son convenientes para plasmar la realidad de una forma precisa.

Los datos necesarios para calcular el IDH son los siguientes:

- La esperanza de vida en años que tendrá un valor mínimo de 20 años y un máximo de 85
- La esperanza de años de escolarización que irá de 0 a 18 años y la media de años deescolarización de 0 a 15 años
- El producto interior bruto per cápita que irá de un mínimo de 100 a un máximo de 75.000 dólares

Para calcular el valor de cada uno de los tres elementos se usará la siguiente fórmula (López, 2019):

$$Índice\ dimensión = (valor\ actual - valor\ mínimo)(valor\ máximo - valor\ mínimo) \quad (69)$$

Como apunte, en el caso del valor para la educación se realizará una media aritmética entre los componentes, por otro lado, para obtener el valor referente al ingreso, se tomarán las variables con logaritmos de base 10 para así ajustar los resultados.

Una vez obtenidos los valores de cada elemento, el IDH se calcula con la fórmula definida en (70):

$$IDH = (Isalud * Ieducación * Iingresos)^{1/3} \quad (70)$$

5. Resultados:

5.1. Análisis descriptivo:

En la *tabla 1* se encuentran los resultados de los descriptivos obtenidos de las variables. En ellos se observa que la variable con mayor inestabilidad es la x.3.3.2 *(objetivo 3)* correspondiente a la incidencia de tuberculosis siendo su media de 25,172 y su valor máximo 102 el cual difiere mucho del valor medio y por lo tanto es entendible que la varianza sea tan grande. Este valor de 102 corresponde a Moldavia siendo este el país europeo en el que más muertes por tuberculosis de producen, estando en el puesto 80 del mundo.

Otra variable cuya varianza es muy elevada, teniendo un valor de 152,252, es la x3.7.2 *(objetivo 3)* que hace referencia a la tasa de fecundidad en adolescentes, su media es de 14,68 y su valor máximo de 52,42. Este corresponde a Azerbaiyán, siendo la maternidad en adolescentes un gran problema tanto en ese país como en el resto de los países del este de Europa. Por otro lado, el valor mínimo de esta variable corresponde a Dinamarca, con un valor de 2,3, esto es debido a que la educación sexual en países del norte del Europa es mucho mayor que en los países del este.

También destaca la variable x.4.2.2 *(objetivo 4),* tasa de participación en el aprendizaje organizado, cuya varianza es de 197,88 y por lo tanto, también presenta una gran variabilidad. Esta vez no es debido a su valor máximo que teniendo un valor de 100, no se aleja tanto de la media cuyo valor es de 92,39, en esta ocasión el problema se encuentra en su valor mínimo que es de 27,60 y corresponde de nuevo a Azerbaiyán esto es debido a que en numerosos países del este la educación y la escolarización de la gran mayoría de su población es todavía una meta por cumplir.

Y por último, la variable donde la varianza tiene un valor más elevado es la x.8.10.2 *(objetivo 8)* que hace referencia a la proporción de adultos que poseen una cuenta en un banco o en otra institución financiera, su media es de 28,07 y su valor máximo de 67,51 correspondiente a Suiza y su valor mínimo de 0,56 perteneciente a Eslovenia.

En cuanto a la simetría de las variables es destacable observar cuales de ellas tienen una media y una mediana iguales, ya que esto indicaría que siguen una distribución simétrica y por lo tanto normal. En este caso ninguna de las variables cumple este criterio, aunque si hay algunas que están muy cerca de ello como pueden ser las variables x.3.8.1 *(objetivo 3),* x.4.5.1 *(objetivo 4)*, x.8.2.1 *(objetivo 8)* y la x.8.10.2 *(objetivo 8)* estas serían aquellas que más cerca estarían de seguir una distribución normal.

Por otro lado, y como variables no simétricas destaca la variable x.3.3.2 *(objetivo 3)* con una asimetría muy pronunciada hacia la derecha al ser la media notablemente superior a la mediana

Variables	*N*	*Mínimo*	*Máximo*	*Media*	*Varianza*	*Mediana*
x3.1.1	43	2,00	28,00	9,60	54,53	7
x3.2.1	43	2,20	26,30	6,00	20,44	4,3
x.3.2.2	43	1,10	14,20	3,55	7,71	2,6
x.3.3.2	43	2,40	102,00	25,17	794,79	10
x.3.4.1	43	8,70	27,50	16,23	35,09	15,50
x.3.4.2	43	4,10	34,90	14,05	44,27	13.30
x.3.5.2	43	0,46	15,40	10,41	7,67	11,125
x.3.7.2	43	2,30	52,40	14,69	152,52	10,5
x.3.8.1	43	58,00	86,00	74,47	62,06	75
x.3.9.3	43	0,10	2,70	0,51	0,35	0,3
x.4.1.2	43	90,46	100,00	97,52	6,56	98,58
x4.2.2	43	27,61	100,00	92,39	197,89	96,80
x.4.5.1	43	0,71	1,69	1,06	0,03	1,013
x.8.1.1	43	-9,39	24,49	2,69	20,17	2,330
x8.2.1	43	-9,80	22,10	1,38	19,86	1,1
x8.6.1	43	4,60	35,70	14,34	63,31	11,90
x.8.5.2	43	3,00	23,80	9,07	33,35	7,3
x.8.10.2	43	0,57	67,52	28,07	183,78	28,26
x8.10.1	43	28,57	99,92	81,37	349,90	86,14

Tabla 1. Estadísticos descriptivos / Fuente: Elaboración propia con SPSS

A continuación, se realiza un gráfico *boxplot* de las variables. Este tipo de gráficos, también denominados gráficos de cajas sirven para representar los valores entre los que se mueve una variable, y dan una representación gráfica de los resultados que se vió en la *Tabla 1*. Las partes que los componen son por una parte las cajas, donde viene representado el 50% de los datos delimitados por el cuartil 1 (Q1) y el cuartil 3 (Q3) que son los bordes extremos de la caja, dividiendo esta caja se encuentra una línea horizontal que representa la mediana. Por otro lado, se encuentran las líneas verticales que representan los valores extremos leves tanto inferior como superior, y por último, se encuentran los *outliers*, que son los puntos que están fuera de los límites superior e inferior de los datos.

Debido a que las variables están en diferentes escalas, para realizar el *boxplot* se procedió a escalarlas. En la *Ilustración 4* se observa como resultados más relevantes que las variables con mayor dispersión de los datos son la x.3.4.1 *(objetivo 3)* y la 3.8.1 *(objetivo 3),* mientras que las variables con menos son la x.4.2.2 *(objetivo 4)*, la x.3.9.3 *(objetivo 3)* y la x.8.2.1 *(objetivo 8).* Esto se conoce debido a que cuanto más grande sea la caja de la variable,significará que el abanico de datos perteneciente a ella será mayor, al contrario que aquellas en las que su caja es pequeña, que esto indicará que sus datos están cercanos todos a la mediana. En cuanto a los *outliers* se puede ver que la gran mayoría son datos que sobrepasan loslímites superiores, únicamente poseen *outliers* inferiores 7 variables que son las que son las que comprenden las tonalidades azules.

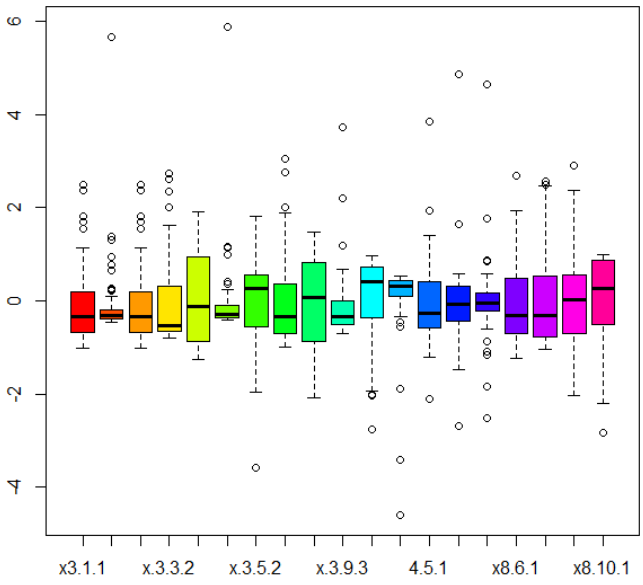


Ilustración 4. Boxplot variables escaladas / Fuente: elaboración propia con Rstudio

5.2. Clasificación en 2 grupos:

En esta sección se procederá a la agrupación de los países en dos grupos, para así posteriormente compararla con el *ranking* el *IDH.*

5.2.1. Análisis de componentes principales:

Antes de realizar el ACP es necesario observar si las variables tienen relación entre si no la tuvieran el ACP carecería de sentido como vimos en el apartado *4.5*. Para buscar esa posible relación se utilizará la matriz de correlaciones, obteniendo los resultados que se presentan en la *Tabla 2.*

En esta tabla se observa que la variable 3.1.1 *(objetivo 3)* referente a la tasa de mortalidad materna, tiene una correlación positiva con la variable 3.2.1 *(objetivo 3)* que mide la tasa de mortalidad infantil en menores de 5 años y con la variable 3.2.2 *(objetivo 3)* que se refiere a la tasa de mortalidad neonatal, esta correlación positiva indica que cuando una variable aumenta la otra también lo hace, lo que puede indicar que en los países en los que ambos valores sean altos, puede ser debido a que las condiciones sanitarias no sean las más óptimas y por ello haya más fallecimientos.

Esto se puede corroborar fijándose en la correlación entre la variable 3.1.1 *(objetivo 3),* mencionada anteriormente, y la 3.8.1 *(objetivo 3)* quehace referencia a la cobertura de los servicios de salud esenciales, estas dos variables tienen una correlación negativa por lo que cuando una aumenta de valor la otra disminuye, esto tiene sentido ya que cuanto más cubiertos estén los servicios de salud menos fallecimientos se producirán e hospitales y viceversa.

Variable 1	*Variable 2*	*Correlación*
3.1.1	3.2.1	0.7622
3.1.1	3.2.2	0.7080
3.1.1	3.4.1	0.7086
3.2.2	4.2.2	-0.7012
3.4.1	3.9.3	0.6696
3.1.1	3.8.1	-0.6254
3.8.1	8.10.1	0.7278

Tabla 2. Correlaciones más significativas / Elaboración propia con R

Una vez se ha visto que las variables guardan relación entre ellas a través de la matriz de correlaciones se puede proceder al cálculo de las componentes principales. En primer lugar, se deben escalar las variables ya que como se vio en el apartado *5.4.3*, cuando las variables están en distinta escala es necesario hacerlo si no aquellas con mayor varianza tendrán más influencia en el modelo.

Se obtiene un total de 15 componentes, como sólo es necesario quedarse con aquellas que mejor expliquen el modelo se procederá a realizar diferentes técnicas pala la selección del número óptimo de componentes.

En un primer lugar, se utilizará la técnica más común, que es la "*regla del codo",* cuya representación se puede observar en la *Ilustración 5* en el que se muestran los autovalores de la matriz de correlación. Para elegir el número de componentes es necesario encontrar el punto a partir del cual, la tendencia de la gráfica pasa a ser claramente descendente, en este caso, no se da una idea clara de donde está ese punto y por lo tanto sería preciso buscar otro método para la obtención de estas.

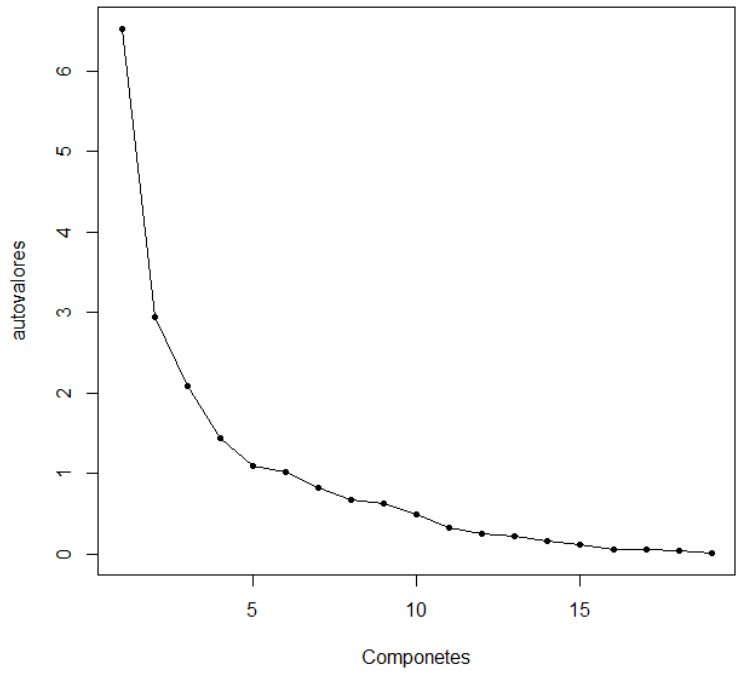


Ilustración 5. Gráfico de sedimentación / Fuente: Elaboración propia con Rstudio

Como método alternativo a la *regla del codo* se utilizará el criterio de *Kaiser*, el cual indica que se elegirán las componentes que tengan una varianza mayor que la unidad, este dato se

obtiene extrayendo los valores propios de la *matriz de correlación* ya que como se vio en apartado *4.5.2* estos coinciden. Una vez obtenidos estos valores se observa en la *Tabla 3* que el número óptimo de componentes que se deben de elegir es 6.

CP1	*CP2*	*CP3*	*CP4*	*CP5*	*CP6*
6.529	2.948	2.098	1.448	1.093	1.015

Tabla 3. Varianzas de las componentes elegidas / Fuente: Elaboración propia con Rstudio

Tras obtener el número de componentes que se utilizará en el modelo, es interesante estudiar la correlación lineal que existe entre las componentes elegidas y las variables iniciales, para ello se utilizará las saturaciones.

Estas se pueden observar de forma numérica mediante la matriz de saturaciones o de forma gráfica mediante un *heatmap* como se muestra en la *Ilustración 6.* En se puede observar ver que en la parte izquierda se representa el *dendrograma* de las variables iniciales, por otro lado, la relación entre las variables y las componentes viene dado por un abanico de colores que va desde el azul, para representar una fuerte relación negativa hasta el rojo para representar una fuerte relación positiva.

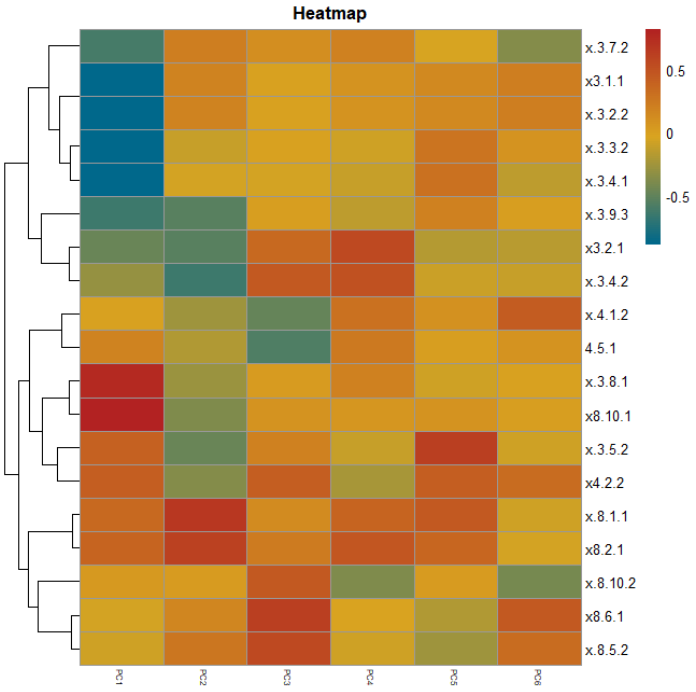


Ilustración 6. Heatmap de las componentes / Fuente: Elaboración propia con Rstudio

5.2.2. Análisis de cluster:

Una vez elegidas y visualizadas las componentes con las que se va a trabajar, es el momento de comenzar con los métodos de *clustering* para la clasificación de los países y así conseguir la clasificación en dos grupos que se busca.

5.2.2.1. Algoritmo k-means:

En un primer lugar, se utilizará el algoritmo *k-means* se vió en el apartado *4.6.1* es necesario elegir previamente el número *k* de *clusters*, para ello se utilizarán dos métodos. El primero será el *método de la silueta* cuyo resultado se muestra en la *Ilustración 7* e indica que el número óptimo de *clusters* es 2.

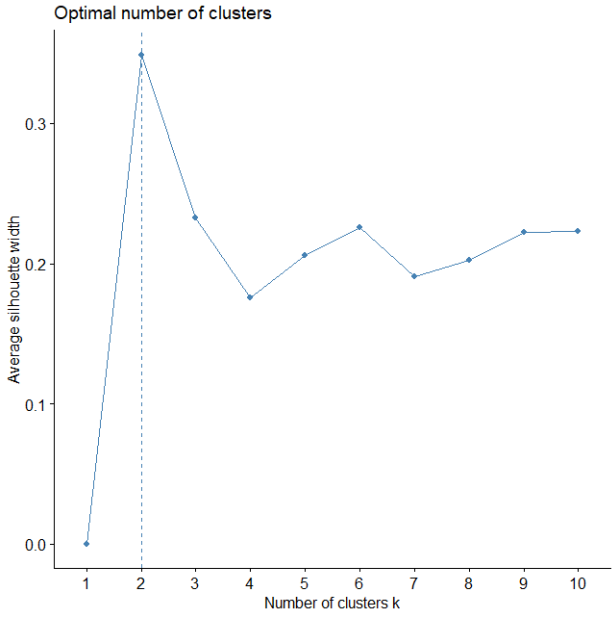


Ilustración 7. Método de la silueta / Fuente: Elaboración propia con Rstudio

Aunque el *método de la silueta* es fiable y seguramente sea 2 el número óptimo de *clusters*, es conveniente asegurarse utilizando otro método, en este caso el elegido será el método estadístico de la brecha y como se observa en la *ilustración 8* este vuelve a indicar como ya ocurría anteriormente que el número de *clusters* debe ser 2.

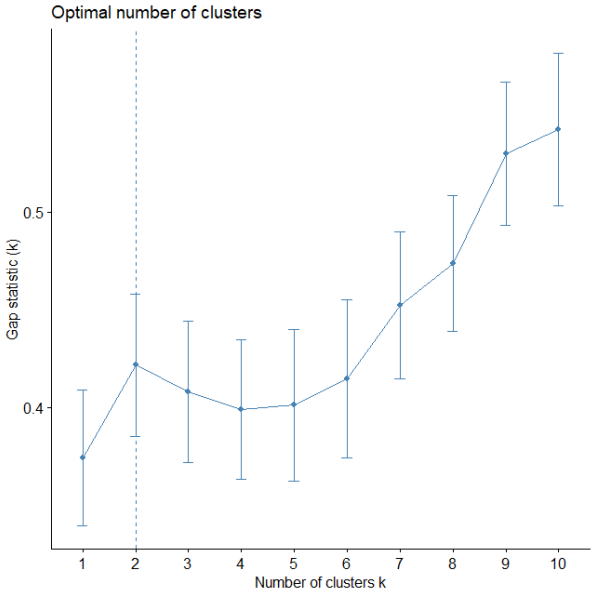


Ilustración 8. Método estadístico de la brecha / Fuente: Elaboración propia con Rstudio

Una vez se han obtenido el número de *clusters* se puede proceder a la creación y representación de los mismos. Esta se muestra en la *ilustración 9* en el que se observa que

divide los países en el *cluster* 1 (rojo) y el *cluster* 2 (azul), de esta división se puede obtener que en el *cluster* rojo se encuentran únicamente países pertenecientes al este de Europa lo que hace pensar que es lógico que estos países tengan un desarrollo humano similar, por el contrario en el *cluster* azul se encuentran en su mayoría países del centro sur de Europa estando estos mucho más agrupados que los del otro *cluster*, lo que hace ver que entre estos países hay un desarrollo humano más similar que en los otros. Se observa que por ejemplo los países nórdicos están muy juntos entre sí.

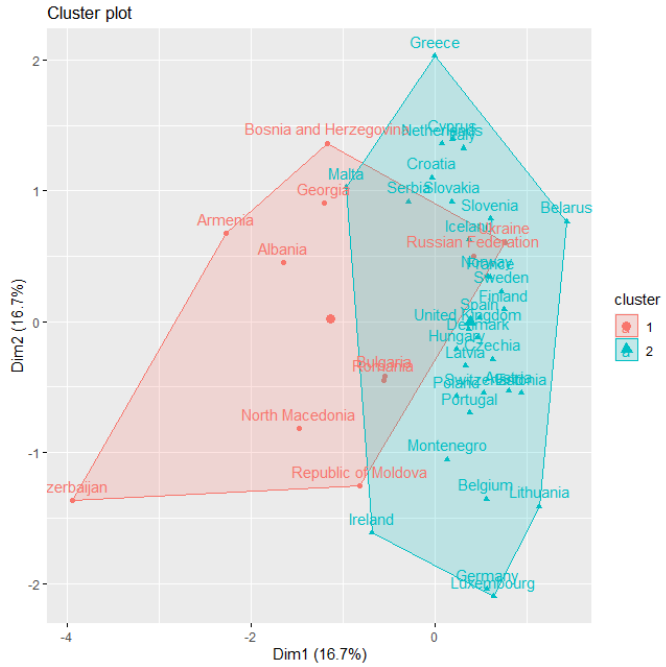


Ilustración 9. Algoritmo k-means con 2 clusters / Fuente: Elaboración propia con Rstudio

Antes de quedarse únicamente con esta clasificación se va a realizar otro método para así posteriormente compararlos y ver cuál de los dos es óptimo para después hacer la comparación con el *ranking* del *IDH*.

5.2.2.2. Agrupamiento jerárquico:

El método que se utilizará a continuación es el método de *agrupamiento jerárquico*, en este caso como se vio en el apartado *4.6.2,* no es necesario elegir un número de *clusters* previo ya que será tras la realización del agrupamiento, cuando se elegirá el quede la clasificación más apropiada de los elementos, para ello se utilizará un dendrograma y como medida de la distancia entre grupos se utilizará el criterio de *Ward.*

El resultado obtenido se muestra en la *ilustración 10* tras observarlo es óptimo realizar la poda del árbol a la altura del 15 ya que los dos *clusters* que deja son interesantes para poder compararlos posteriormente con los valores del *IDH*. En la división de la izquierda se puede observar que se encuentran los países pertenecientes al centro y sur de Europa, al igual que se tenía en el agrupamiento obtenido por el método de *k-means,* solo que esta vez la división es más clara ya que mientras en el otro método en el *cluster* de países del centro-sur encontrábamos varios pertenecientes al este de Europa, en este nuevo *cluster* se tiene ninguno y es en el segundo donde se puede observar que están todos los países pertenecientes a dicha zona geográfica.

Esta nueva agrupación tiene bastante sentido, ya que según la zona territorial en la que te encuentres es normal tener unas condiciones de vida y un desarrollo similar, siendo los países del este territorios que en su mayoría las condiciones de vida y de desarrollo son más bajas que en el resto de los países del continente europeo.

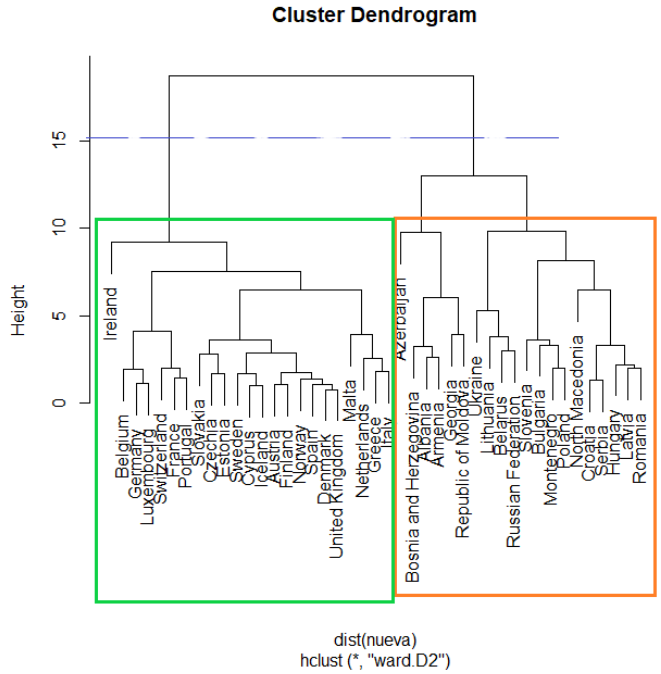


Ilustración 10. Clusters agrupamiento jerárquico / Fuente: Elaboración propia con Rstudio

5.2.3. Comparación con el IDH:

Una vez creados los dos grupos que se buscaban como objetivo principal del trabajo, se procederá a continuación a comprar la clasificación obtenida con el *ranking* oficial del *IDH*. Por un lado, en la *tabla 4* se muestran los países con un *IDH* más elevado mientras que en la *tabla 5* se tienen los países con un valor más bajo.

Países	*IDH*
Noruega	0.957
Irlanda	0.955
Suiza	0.955
Islandia	0.946
Alemania	0.947
Suecia	0.945
Países Bajos	0.944
Dinamarca	0.940
Finlandia	0.938
Reino Unido	0.932

Tabla 4. Países europeos con más IDH / Fuente: Ranking oficial del IDH

Países	*IDH*
Moldavia	0.750
Azerbaiyán	0.756
Macedonia del Norte	0.774
Armenia	0.776
Ucrania	0.779
Bosnia y Herzegovina	0.780
Albania	0.795
Serbia	0.806
Georgia	0.812
Bulgaria	0.816

Tabla 5. Países europeos con menos IDH / Fuente: Ranking oficial del IDH

Si se comparan estos valores con la clasificación obtenida en la *ilustración 10* se puede observar que los países pertenecientes al *cluster* rojo son aquellos que cuentan con un *IDH* más alto mientras que aquellos que pertenecen al *cluster* verde son los países que tienen un *IDH* más bajo.

5.3. Clasificación en varios grupos:

Una vez se ha obtenido el objetivo principal del trabajo que era la clasificación en dos grupos de los países para así posteriormente comparar los resultados con los valores del *IDH*, es interesante buscar una clasificación más precisa de estos países basándose en las variables originales.

Para ello se utilizará el algoritmo *k-means* creyendo oportuno, debido a la disparidad entre los países que se pudo ver anteriormente, utilizar una agrupación de 4 *clusters* como se muestra en la *ilustración 11.* En él se puede observar que la división que realiza el algoritmo es el *cluster* 1 (rojo), *cluster 2* (morado), *cluster 3* (azul) y *cluster 4* (verde). El en 1 se visualiza una diferencia clara entre los países a pesar de pertenecer al mismo grupo, Georgia se encuentra en el centro del *cluster* y el resto de los países en los extremos siendo Azerbaiyán y Moldavia los dos que más difieren entre sí.

En el número 2 se observa una mayor homogeneidad dentro del grupo en comparación con la que se tiene en el 1, este indica que países como Croacia, Montenegro, Macedonia etc. Tienen unas características muy similares debido a su cercanía en el *cluster*, por el contrario, Bosnia y Herzegovina y Bielorrusia con aquellos que muestran mayor diferencia.

El grupo 3 es aquel en el que más igualdad se encuentra entre los países, se puede observar una gran cantidad de países agrupados que pertenecen al norte, centro y sur de Europa, esto es debido a que tienen un estilo de vida y una cultura muy similares, lo que hace que sus características de desarrollo también lo sean. Los únicos dos países que distan del resto son Malta y de una forma más llamativa Irlanda.

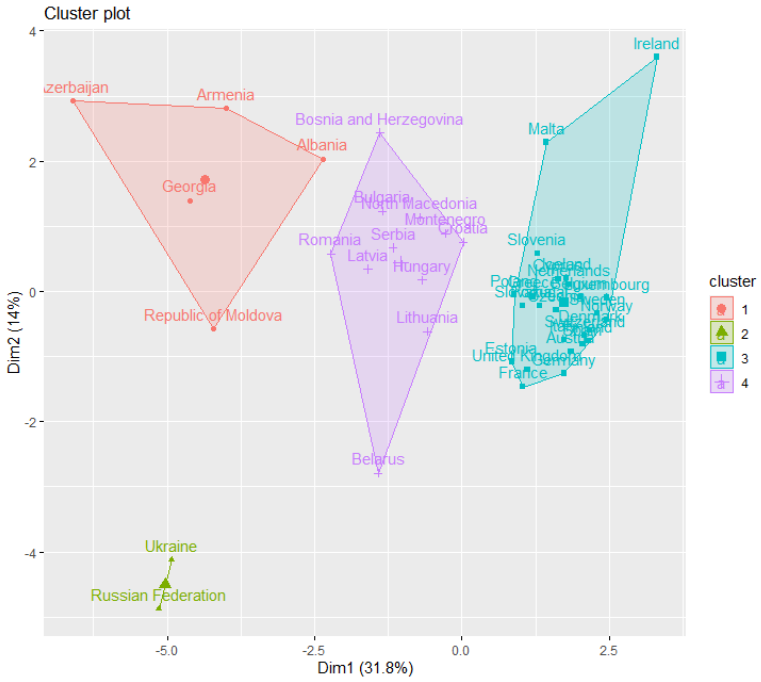


Ilustración 11. Gráfico algoritmo k-means con 4 clusters / Fuente: Elaboración propia con Rstudio

El último *cluster* a observar es el 4, en el únicamente se agrupan los países de Ucrania y Rusia, esto indica que son los dos que más difieren con el resto de Europa, ambos países pertenecientes a la Unión soviética, siendo Rusia la principal potencia tanto antes como ahora. También se sabe que es un país con una política muy restrictiva que hace que las condiciones de vida no sean las óptimas para un desarrollo próspero

6. Conclusiones:

A nivel concluyente, centrándose en primer lugar en el principal objetivo del trabajo que era buscar un modelo más completo para poder medir el desarrollo humano se extrae de conclusión que el nuevo método es efectivo ya que como se pudo ver en el apartado *5* y tras elegir el método de clasificación más apropiado que en este caso fue el Agrupamiento jerárquico, se pudo observar que la división que realizó este método en los dos grupos que se buscaba encontrar, clasifica a los países de tal forma que junta a aquellos con mayor *IDH* en uno y en el otro a los países con menor.

Esto indica que a pesar de que para el modelo del *IDH* los investigadores buscaban hacerlo de la forma más simplificada posible y por ello eligieron únicamente una variable para cada uno de los elementos, que eran: salud, educación y nivel de vida, también es posible introducir nuevas variables a cada elemento para así crear un modelo más complejo y que como se ha visto nos facilita una clasificación óptima de los países.

Como conclusión a la agrupación más especifica que se realizó para así dar una clasificación más concreta de los países y no sólo en los dos grupos iniciales. Se observó que esta clasificación en 4 *clusters* agrupa a los países según su distribución geográfica, lo cual es coherente debido a que son países con una calidad de vida muy similar la cual influye en su desarrollo.

7. Bibliografía:

Aravena-Jara, L. I. (2021). *La agenda 2030 de Naciones Unidas: historia de los objetivos de desarrollo sostenible y el objetivo de desarrollo sostenible N° 13 en Chile.*

Barrero-Barrero, D., & Baquero-Valdés, F. (2020). Objetivos de Desarrollo Sostenible: un contrato social posmoderno para la justicia, el desarrollo y la seguridad. *Revista Científica General José María Córdova*, *18*(29), 113–137.

Beattie, R. M., Brown, N. J., & Cass, H. (2015, February 1). Millennium development goals progress report. *Archives of Disease in Childhood*, Vol. 100, p. S1. https://doi.org/10.1136/archdischild-2014-307933

Blashfield, R. K., & Albenderfer, M. S. (1978). The literature on cluster analysis. *Multivariate Behavioral Research*, *13*(3), 271–295. https://doi.org/10.1207/s15327906mbr1303_2

Boutsidis, C., & Magdon-Ismail, M. (2013). Deterministic feature selection for k-means clustering. *IEEE Transactions on Information Theory*, *59*(9), 6099–6110. https://doi.org/10.1109/TIT.2013.2255021

Bridges, C. C. (1966). Hierarchical Cluster Analysis. *Psychological Reports*, *18*(3), 851–854. https://doi.org/10.2466/pr0.1966.18.3.851

Carmona, F. (2014). *Un ejemplo de ACP paso a paso.*

Celebi, M. E., & Aydin, K. (2016). *Unsupervised learning algorithms.* Berlin: Springer International Publishing.

Clapham, C. (2004). *Diccionario de Matemáticas (1ª edición).* Madrid: Editorial Complutense.

Conceição, P. (2019). *Informe sobre Desarrollo Humano*.

Cosme-Casulo, J. (2018). Los Objetivos de Desarrollo Sostenible y la academia. *Medisan*, *22*(8), 838–848.

Delval, J. (1994). *El desarrollo humano. Siglo XXI de España Editores.*

Espinel, P. (2015). Procedimiento para efectuar una Clasificación Ascendente Jerárquica de un Conjunto de Puntos utilizando el Método de Ward. *Infociencia*, *9*(1), 13–18.

Fisher, R. A. (1919). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. In *Transactions of the Royal Society of Edinburgh* (pp. 399–433).

García, C., & Gómez, I. (2006). Algoritmos de aprendizaje: knn & kmeans. *Universidad Carlos III de Madrid*. Retrieved from http://www.it.uc3m.es/jvillena/irc/practicas/08-09/06.pdf

Gil, C. G. (2017). Objetivos de Desarrollo Sostenible (ODS): una revisión crítica. In *Nº* (Vol. 140).

Gonzalo, Á. (2019). Segmentación utilizando K-means en Python. Retrieved July 3, 2021, from Machine Learning para todos website: https://machinelearningparatodos.com/segmentacion-utilizando-k-means-en-python/

Griggs, D. (2013). Sustainable development goals for people and planet. *Nature*, *495*, 305–307.

Gurrea, M. (2000). *Análisis de componentes principales.*

Hand, D., Mannila, H., & Smyth, P. (2001). Principles of Data Mining Cambridge. In *MIT Press* (Vol. 2001). Retrieved from http://link.springer.com/10.1007/978-1-4471-4884-5

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Hierarchical clustering. In *The Elements of Statistical Learning* (pp. 520–528). Nueva York.

Hulme, D. (2009). *The Millennium Development Goals (MDGs): A Short History of the World's Biggest Promise Creating and sharing knowledge to help end poverty*. Retrieved from www.manchester.ac.uk/bwpi

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151.

Kenney, J. F., & Keeping, E. S. (1951). *Mathematics of Statistics*. Princeton, NJ: Van Nostrand.

Kettenring, J. R. (2006). The Practice of Cluster Analysis. *Journal of Classification*, *23*(1), 3–30. https://doi.org/10.1007/s00357-006-0002-6

López, J. F. (2019). Guía para calcular e interpretar el IDH. Retrieved July 3, 2021, from Economipedia website: https://economipedia.com/guia/guia-para-calcular-e-interpretar-el-idh.html

Lozares-Colina, C., & López-Roldán, P. (1991). El análisis de componentes principales: aplicación al análisis de datos secundarios. *Papers: Revista de Sociologia*, (37), 31–63.

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.

Marín, C. (2015, April 22). ¿Se han cumplido realmente los Objetivos de Desarrollo del Milenio? Retrieved July 2, 2021, from https://www.elmundo.es/salud/2015/04/22/5536600422601d16058b457e.html

Marín, J. M. (2015). *Tema 3: Análisis de Componentes Principales*.

Méndez-Peña, D. P., & Sánchez-Gutiérrez, R. (2014). *Análisis de Componentes Principales en la estimación de índices de empoderamiento en mujeres en Colombia*.

Molina-Salazar, R. E., & Pascual-García, J. M. J. (2015). El Índice de Desarrollo Humano como indicador social. *Nómadas. Revista Crítica de Ciencias Sociales y Jurídicas*, *44*(4), 127–143. https://doi.org/10.5209/rev_noma.2014.v44.n4.49298

Moya, R. (2016). Jarroba. Retrieved from Selección del número óptimo de Clusters website: https://jarroba.com/seleccion-del-numero-optimo-clusters/

Navarro-Céspedes, J. M., Casas-Cardoso, G. M., & González-Rodríguez, E. (2010). Análisis de Componentes Principales y Análisis de Regresión para Datos Categóricos. Aplicación en la Hipertensión Arterial. *Revista de Matemática: Teoría y Aplicaciones*, *17*(2), 199–230. https://doi.org/10.15517/rmta.v17i2.2128

Ochoa-Reyes, A. J., Orellana-García, A., Sánchez-Corales, Y., & Davila-Hernández, F. (2014). Componente web para el análisis de información clínica usando la técnica de Minería de Datos por agrupamiento. *Rev. Cuba. Inform. Méd*, *6*(1), 5–16.

Papalia, D. E. (2009). *Desarrollo humano.*

Perales, J. A. S. (2014). De los Objetivos del Milenio al desarrollo sostenible: Naciones Unidas y las metas globales post-2015. *Anuario Ceipaz*, (7), 49–84.

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, *49*(4), 974–997.

Prendergast, A. J., Essajee, S., & Penazzato, M. (2015). HIV and the millennium development goals. *Archives of Disease in Childhood*, *100*(Suppl 1), S48–S52. https://doi.org/10.1136/archdischild-2013-305548

Press, W., Teukolsky, S., Vetterling, W., & Flannery, B. (2007). Hierarchical Clustering by Phylogenetic Trees. In *Numerical Recipes: The Art of Scientific Computing (3rd edición)*. New York.

Robaina-Romero, R. (2019). *Prólogo: Transformar nuestro mundo: la agenda 2030 para el desarrollo sostenible.*

Robles-Llamazares, M. (2006). *Objetivos de desarrollo del milenio.*

Rosenberg, H. (1994). El índice de desarrollo humano. *Boletín de La Oficina Sanitaria Panamericana*, *117*(2).

Schonlau, M. (2004). Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams. *Computational Statistics*, *19*(1), 95–111.

Shlens, J. (2005). *A Tutorial on Principal Component Analysis*.

Significado de Desarrollo Humano. (2019). Retrieved from https://www.significados.com/desarrollo-humano/

Snedecor, G. W., & Cochran, W. G. (1980). Statistical Methods, 7th ed. *IA: Iowa State Press*, 342.

Spiegel, M. R. (1992). *Theory and Problems of Probability and Statistics, 2nd ed.* New York: McGraw-Hill.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conference Series: Materials Science and Engineering*, *336*(1), 012017.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293

Torres, C., & Mújica, O. J. (2004). Salud, equidad y los Objetivos de Desarrollo del Milenio. *Revista Panamericana de Salud Pública*, *15*, 430–439.

Torres, V. E. R., Bertone, C. L., & Andrada, M. J. (2018). Brechas en la mortalidad infantil según nivel educativo de las madres en la provincia de Córdoba. Estimación indirecta a partir de datos censales 2010. *Revista de Salud Pública*, *22*(3), 37–47.

Vilà-Baños, R., Rubio-Hurtado, M. J., Berlanga, V., & Torrado-Fonseca, M. (2014). Cómo aplicar un cluster jerárquico en SPSS. *REIRE. Revista d'Innovació i Recerca En Educació*, *7*(1), 113–127.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244.

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2008). Top 10 algorithms in data mining. In *Knowledge and Information Systems* (Vol. 14). https://doi.org/10.1007/s10115-007-0114-2

Xu, R., & Wunsch, D. (2008). *Clustering (Vol. 10).* John Wiley & Sons.

8. Abstract:

The Sustainable Development Goals (SDGs) are a necessary asset in order to being able to achieve global sustainability that allows us to advance as a society in the future and for the benefit of all humanity. The SDGs are direct heirs of the Millennium Development Goals (MDGs) and exist to support the achievement of them through a series of short-term proposals.

Before continuing to explain the SDGs, it is necessary to speak beforehand about their predecessors; the Millennium Development Goals (MDGs). These objectives reflect and summarize the development goals that were set out at the international conferences and world summits that took place during the 1990s. It was on September 8, 2000, when the United Nations General Assembly summarized the main objectives and goals in what was called the Millennium Declaration.

The Millennium Declaration is divided into eight objectives, each objective was divided into a series of goals, there were a total of 28 goals and 48 specific indicators. The objectives are the following:

- Eradicate extreme poverty and hunger
- Achieve universal primary education
- Promote gender equality and empower women
- Reduce infant mortality
- Fight HIV, malaria, and other diseases
- Guarantee the sustainability of the environment
- Foster the global partnership for development

Despite the progress got by the MDGs, for the international community these were not enough, and they wanted to look for another long-term project that would guarantee the sustainable development of the planet, that is why in 2015 the "Sustainable Development Goals were approved.

This is considered the continuation of the MDGs, in order to address the problems that could not be solved and incorporating 17 objectives and 169 goals. These objectives are:

- *Goal 1:* End poverty
- *Goal 2:* Zero hunger
- *Goal 3:* Health and well-being
- *Goal 4:* Quality education
- *Goal 5*: Gender equality
- *Goal 6*: Clean water and sanitation
- *Goal 7:* Affordable and clean energy
- *Goal 8:* Decent work and economic growth
- *Goal 9:* Water industry innovation and infrastructures
- *Goal 10:* Reduction of inequalities
- *Goal 12:* Sustainable cities and communities
- *Goal 13:* Responsible consumption and production
- *Goal 14:* Climate action
- *Goal 15:* Underwater life
- *Goal 16:* Life of terrestrial ecosystems

- *Goal 17:* Peace, justice, and strong institutions

Human development is the process by which a society, based on a good economic evolvement, improves the living conditions of his members.
Therefore, human development not only means that people choose the minimum resources to cover their basic needs, but also that they have access to health and education systems, relevant levels of personal security, full political freedoms. and cultural.

This progress is one of the goals of the United Nations, more specifically, the organism whose mission is to coordinate all policies on human development is the United Nations Development Program (UNDP). They present periodically the Annual World Program on Human Development, this program shows the statistical data that calculate, according to a series of indicators, the level of human development. The most common way and the one which I will focus is the Human Development Index (HDI).

The aim of this project is to find another alternative method to the HDI to be able to classify a series of countries according to their Human Development in two groups. For this I will be taking a sample of 43 European countries and, as variables, 19 indicators belonging to three objectives of the SDGs.

It knows as mean the following formula.

$$E(X) = \sum_{i=1}^{+\infty} x_i p(x_i)$$

And if it is a continuous function, you obtain.

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

If are X e Y two random variables and their means are $\mu x \; y \; \mu y$. The covariance between $X \; e \; Y$ is:

$$Cov(X, Y) \; = \; \sigma x, y \; = \; E((X \; - \; \mu x)(Y \; - \; \mu y))$$

Having a random vector $X = (X_1, X_2, \dots, X_k)^t$ and his mean vector is $\mu x = (\mu_1, \mu_2, \dots \mu_k)^t \; E(X^2) \; < \; \infty$ for $i \; = \; 1,2, \dots, k$ his *covariance matrix* will be:

$$S = Cov(X) = E\big((X - \mu_x)(X - \mu_x)^t\big) = \begin{pmatrix} \sigma_1^2 & \sigma_1^2 & \dots & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \cdots & \sigma_1^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1^2 & \sigma_1^2 & \cdots & \sigma_1^2 \end{pmatrix}$$

Having two random variables *X* e *Y*. it knows as *coefficient of correlation* between two variables to:

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sqrt{Var(X) \cdot Var(Y)}}$$

And it knows as *correlation matrix* as:

$$R = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1k} \\ \rho_{21} & 1 & \dots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \dots & 1 \end{pmatrix}$$

To calculate the components there is a vector $X = (X_1, X_2, \dots, X_k)$ with dimension k con *covariance matrix* S. The components are defining as:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{pmatrix} = P^t X = \begin{pmatrix} p_{11} & p_{21} & \dots & p_{k1} \\ p_{12} & p_{21} & \dots & p_{k2} \\ \vdots & \ddots & \ddots & \vdots \\ p_{1k} & p_{2k} & \dots & p_{kk} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$$

To calculate the components through the *correlation matrix*, the variables must be scaled.

$$X_i^* = \frac{X_i - \mu_{xi}}{\sigma_i}$$

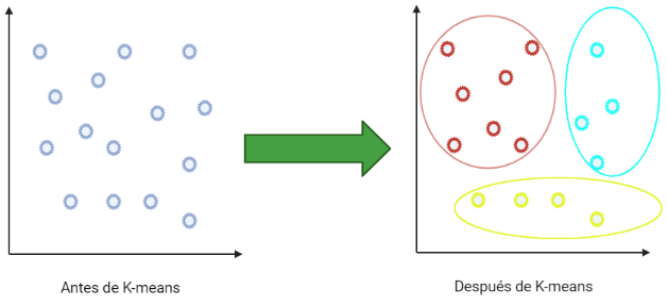
Then the *correlation matrix* will be:

$$\tilde{Y} = \begin{pmatrix} \tilde{Y_1} \\ \tilde{Y_2} \\ \vdots \\ \tilde{Y_k} \end{pmatrix} = \tilde{P}^t X = \begin{pmatrix} \tilde{p}_{11} & \tilde{p}_{21} & \dots & \tilde{p}_{k1} \\ \tilde{p}_{12} & \tilde{p}_{22} & \dots & \tilde{p}_{k2} \\ \vdots & \ddots & \ddots & \vdots \\ \tilde{p}_{1k} & \tilde{p}_{2k} & \dots & \tilde{p}_{kk} \end{pmatrix} = \begin{pmatrix} X_1^* \\ X_2^* \\ \vdots \\ X_k^* \end{pmatrix}$$

After calculating all the components, I had to choose with how many it will be going to work. To make this decision, there are too many methods, in this project, both the *elbow method and Kaiser's rule* will be used.

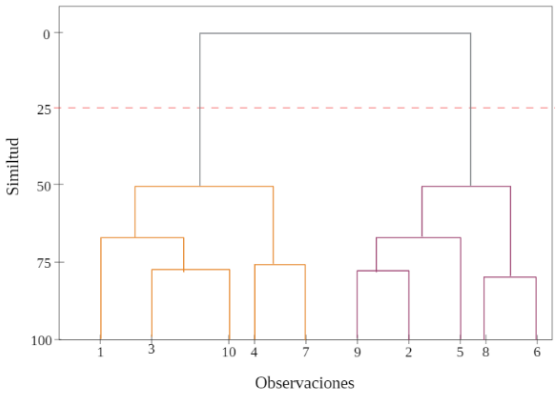
Clustering methods (or cluster analysis) are Exploratory Data Analysis techniques used to solve classification problems. Their main objective consists of ordering objects (things, animals, people, variables, etc.) in different groups (clusters or conglomerates) in such a way that the degree of homogeneity between the members of the same clusters is greater than that which exists between the groups. members of different clusters. Each cluster is defined as the class to which its members belong on it.

The *k-means* algorithm, which was created by MacQueen in 1967, is the most popular and widely used clustering algorithm, because its application is very simple and efficient. It is a simple classification process of a set of objects in a determined k number of clusters, which are determined a priori.



Hierarchical grouping is known in data mining as a point group analysis method, which attempts to fabricate a hierarchy of groups. For this hierarchical grouping, there are different strategies that are generally grouped into two types: agglomerative or divisive.

To make the graphic representation of the hierarchical analysis, the most common method is to use an inverted tree diagram called a dendrogram, a graphic in which the successive mergers of the branches at the different levels give us the information of the successive mergers of the groups. in higher level groups.



The data that is necessary to calculate the HDI are obtained as follows:

- The life expectancy in years will have a minimum value of 20 years and a maximum of 85
- The expected years of schooling will range from 0 to 18 years and the average years of schooling from 0 to 15 years
- The per capita gross domestic product will range from a minimum of 100 to a maximum of 75,000 dollars

To calculate the dimension index of each specific case, the formula that will be used:

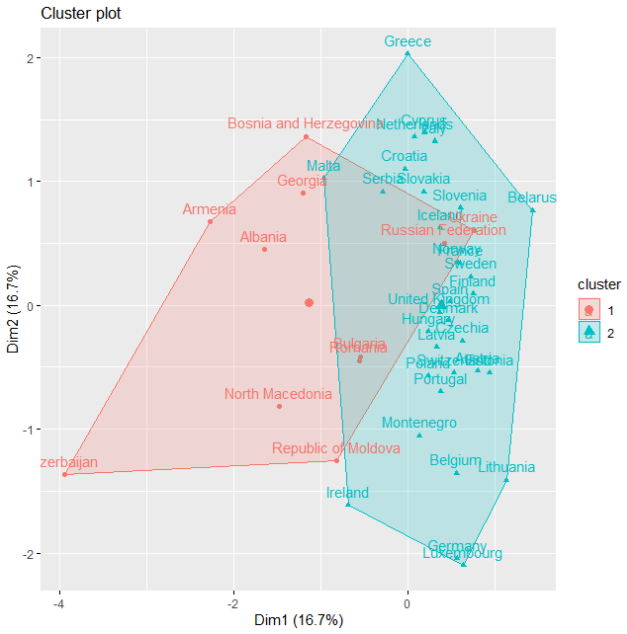
$$dimension\ index = (current\ value\ - minimum\ value)(maximum\ value - minimum\ value)$$

Then, after obtaining the components, the HDI is calculated with the following formula:

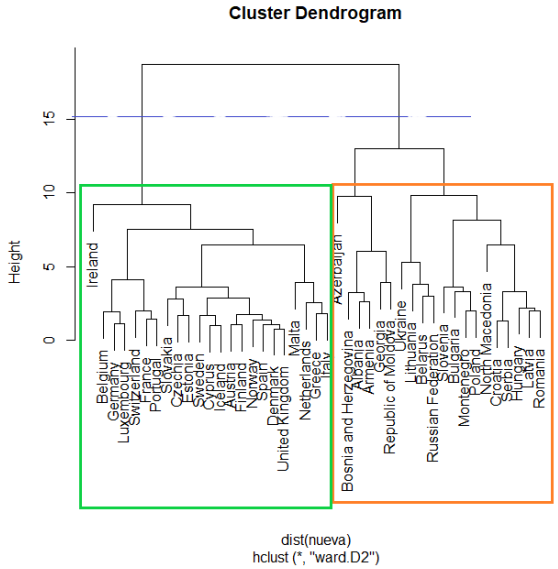
$$HDI = (Ihealth * Ieducation * Iincome)^{1/3}$$

With the obtention of 15 components, it is necessary to make the choice of which ones to consider. To make this decision, the elbow method will be applied, but the result is confusing, so it is necessary to choose another one, in this case, the Kaiser method. This one show that you must choose 6 components, after that will be started to make the clustering analysis.

To begin with, the k-means method will be executed. In order to choose the number of clusters, both the silhouette method and the statistical gap method will be employed, getting as a result the need to use 2 clusters, finally obtaining this next result:



It can be observed that it divides the countries into two clusters: 1 (red) and 2 (blue). In this division, one can see that the red cluster only shows countries from Eastern Europe, this is logical because these countries have a similar human development; while on the contrary, in the blue cluster it shows mostly South and Central European countries. After that, a different practice will be used so it is possible to compare both the methods.

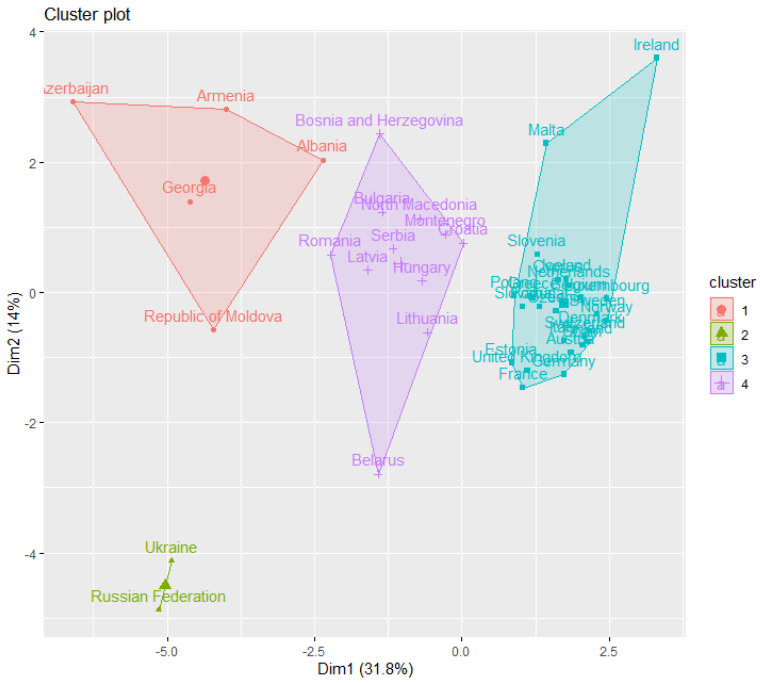


The method that is going to be worked with for this next step is the hierarchical clustering, which results in a dendrogram that shows the following classification.

First of all, it can be observed that it makes a division into two large groups, these will be the clusters in which we are going to group the countries. In the left splitting it can be noticed the appearance of countries both of central and southern Europe, the same as in the k-means method, while in the other cluster we have the countries belonging to the east of Europe.

This aggrupation is better than the other one, and if we compared this result with the value of the HDI of the European countries, it is possible to notice that the countries with more HDI are in the first cluster, conversely the ones with the les IDH are in the second one.

Now it is interesting to make a new aggrupation in more groups to have more than the previous 2, so now the k-means algorithm will be with 4 clusters, the result is the following:



The number 1 and 2 are the groups of the east countries, the third the one of the North, south and Central Europe, and the last one only has the countries of Ukraine and Russia these countries are the more different ones.