UNIVERSIDAD DE SALAMANCA

Departamento de Estadística

Doctorado en Estadística Multivariante Aplicada

Contribuciones a la minería de datos basadas en herramientas multivariantes de Cluster para el tratamiento de Big Data

Autor:

Jorge Fabricio Guevara Viejó

Directoras:

Dra. M. Purificación Galindo Villardón

Dra. M. Purificación Vicente Galindo

2022

Contribuciones a la minería de datos basadas en herramientas multivariantes de Cluster para el tratamiento de Big Data

DEPARTAMENTO DE ESTADÍSTICA
UNIVERSIDAD DE SALAMANCA

Memoria que para optar al Grado de Doctor,
por el Departamento de Estadística de la
Universidad de Salamanca, presenta:

Jorge Fabricio Guevara Viejó

Salamanca
2022

DEPARTAMENTO DE ESTADÍSTICA

UNIVERSIDAD DE SALAMANCA

Purificación Galindo Villardón

Catedrática del Departamento de Estadística de la Universidad de Salamanca

Purificación Vicente Galindo

Profesora Titular del Departamento de Estadística de la Universidad de Salamanca

CERTIFICAN:

Que **Don Jorge Fabricio Guevara Viejó** ha realizado, en la Universidad de Salamanca, bajo su dirección, el trabajo **Contribuciones a la minería de datos basadas en herramientas multivariantes de Cluster para el tratamiento de Big Data,** para optar al título de Doctor en Estadística Multivariante Aplicada, autorizando expresamente su lectura y defensa.

Y para que conste, firman el presente certificado en Salamanca a 30 de Mayo de 2022.

Dra. Purificación Galindo Villardón Dra. Purificación Vicente Galindo

Agradecimientos

Mi más sincero agradecimiento a mi Directora la Dra. Purificación Galindo Villardón, promotora de la enseñanza de la Estadística a nivel mundial, extraordinaria investigadora, profesora y persona, que con sus enseñanzas me llevaron a concluir con éxito mis estudios.

A la Dra. Purificación Vicente Galindo, Directora del Programa de Doctorado de Estadística Multivariante, por su aporte a la construcción de esta investigación y por sus valiosos consejos.

A mi UMEMI querida, proyecto de vida y de la sociedad.

A mi familia, mi esposa, mis hijas y mi hijo.

Dedicatoria

A ti, mi jibarito,
la carga fue apreciada,
el éxito es todo tuyo.

Resumen

Esta investigación se centra en dos pilares esenciales en la economía del Ecuador, los camarones juveniles *Litopenaeus vannamei* y los hongos comestibles del género de *Pleurotus.* Estos hongos son cultivados en mezclas de residuos agrícolas lo cual supone también una solución a este gran problema.

La contribución más importante de esta tesis se enmarca en la transferencia del conocimiento a los sectores agrícolas y acuícolas de Ecuador, y su impacto en la economía circular de estos sectores. También sigue el clásico formato de las publicaciones científicas como producto más relevante de la investigación ya que esta tesis se presenta como un compendio de tres artículos de investigación publicados en revistas de alto impacto.

Utiliza como herramientas de Análisis, Técnicas de Minería de Datos cuyo uso ha permitido conocer qué tipo de dieta se asocia con los mejores parámetros biológicos y comerciales relacionados con el crecimiento de los camarones y sus implicaciones en la composición nutricional, así como cuáles son las mezclas de residuos agrícolas en las que los hongos comestibles del género de *Pleurotus*, muestran las mejores características miceliales, culturales, de productividad y nutricionales.

Palabras claves: Economía; parámetros biológicos y comerciales, técnicas de minería de datos.

Índice General

Índice de Figuras

Índice de Ecuaciones

CAPÍTULO I

INTRODUCCIÓN Y OBJETIVOS

CAPÍTULO I

1. INTRODUCCIÓN Y OBJETIVOS

1.1. Introducción

Desde un punto de vista metodológico, uno de los objetivos de esta investigación es hacer una exhaustiva revisión bibliográfica de las herramientas de clusters más utilizadas con especial énfasis en los algoritmos diseñados para el manejo de grandes bases de datos. Desde el punto de vista del impacto social y económico de esta investigación, estas herramientas multivariantes nos permitirán elaborar propuestas de mejora productos de gran importancia económica en los sectores de agricultura y acuicultura en el Ecuador.

La provincia de Guayas, que comprende 25 cantones, incluido el cantón de Guayaquil, que consta del Municipio de Guayaquil y dos áreas de expansión urbana, junto con cinco cantones rurales: Morro, Juan Gómez Rendón, Posorja, Puna y Tenguel. La provincia de Guayas es parte de la Zona 5, según SENPLADES, Secretaría Nacional de Planificación y Desarrollo, que subdivide el país en nueve zonas. El principal uso de la tierra de la provincia de Guayas es la agricultura (27%), seguida de la ganadería y la acuicultura. Esta provincia ha atravesado tres grandes auges económicos (Delgado, 2013). El primer boom económico se produjo en 1880 cuando fue la principal ciudad exportadora de cacao (entre el 20 y el 25% del mundo), convirtiéndose en un importante centro comercial y financiero, lo que generó un aumento de la población de la ciudad (Mora, 1988), los terratenientes monopolizaron las mejores tierras y el acceso al transporte, así como el control de las principales fuentes de crédito y vínculos comerciales (Pineo, 2008). El segundo boom económico fue en 1950 con el llamado "boom bananero", y llegaron empresas bananeras extranjeras, como la United Fruit Company en Tenguel, una de las mayores plantaciones bananeras (localizada a 100 millas al sur de Guayaquil) (Striffer, 2008). El tercer boom fue el boom petrolero en 1972 que trajo nuevos desarrollos, principalmente en forma de invasiones de tierras en las afueras de la ciudad, provocando un inmenso deterioro del sector agrícola. La falta de una política nacional para la

agricultura rural a pequeña escala llevó a muchos de los pequeños agricultores rurales (principalmente indígenas de las áreas centrales) a abandonar sus parcelas y participar en actividades no agrícolas, con mayor frecuencia en el sector informal urbano (Swanson, 2007). Con los antecedentes descritos anteriormente, es importante seguir utilizando los residuos agrícolas que se siguen generando de toda la provincia del Guayas, teniendo como idea innovadora la producción de hongos comestibles y la elaboración de insumos acuícolas para la acuicultura. (Crisan & Sands, 1978).

Las primeras investigaciones se han realizado enfocadas en el desarrollo biotecnológico de hongos comestibles a nivel de cepas e inóculo tales como: el uso de medios de cultivo sólidos con suplementación de productos agrícolas para mejorar la velocidad de crecimiento, y la influencia de nutrientes del inóculo en medio líquido para aumentar la producción de biomasa y exopolisacáridos (Chegwin & Nieto, 2013; Díaz-Talamantes et al., 2017; Economou et al., 2017). Los requisitos nutricionales y ambientales de las especies biológicas de hongos comestibles presentan una relación directa con los parámetros de productividad después del cultivo (Arana-Gabriel et al., 2014). El cultivo de setas comestibles ha ido creciendo paulatinamente utilizando técnicas caseras, hasta convertirse en una industria altamente técnica (Jong & Peng, 1975; Farr, 1983; Kaul, 1983; Kaul & Kapur, 1987). La producción mundial de hongos comestibles cultivados en los últimos tres años, ha supuesto un incremento anual del 24,5%. El valor nutritivo de los hongos comestibles es alto en comparación con otros tipos de alimentos. Según estudios realizados por especialistas en alimentación, los hongos tienen un contenido proteico entre el 19 y el 35%, en comparación con las verduras (verduras y frutas) que solo tienen proteínas del 7,3 al 13,2%; por otro lado, la leche, la carne y los huevos tienen un contenido de proteínas entre el 25 y el 90%. Sin embargo, a nivel de aminoácidos, las sustancias precursoras de proteínas, como la lisina y el triptófano, alcanzan niveles entre 1,1 y 2,09 g. Por otro lado, el bajo contenido en carbohidratos hace que los hongos sean un alimento poco energético y se recomiendan como dietético. Además, el contenido de ácidos grasos esenciales como el oleico y linoleico se encuentra en cantidades apreciables (Chang & Miles, 2008). Los hongos comestibles son plantas nutritivas que contienen riboflavina, ácido nicotínico, pantotenato y biotina, que reducen la presión arterial, previenen la aterosclerosis y refuerzan el sistema inmunológico contra las enfermedades (Singh et al., 2010).

Uno de los principales géneros de hongos comestibles con mayor producción en

todo el mundo es *Pleurotus* spp. (Valenzuela-Cobos et al., 2017; Sánchez-Hernández et al., 2019). Estos hongos se caracterizan por su valor nutricional y son una fuente importante de proteínas, vitaminas y minerales (Manzi et al., 2001; Reis et al., 2012). Esta especie requiere climas tropicales o subtropicales similares a los de la provincia del Guayas para el cultivo y producción de cuerpos fructíferos (Mori et al., 1974; Fultz, 1988; Kashangura et.al., 2006). Además, estos hongos se utilizan activamente en tratamientos médicos con propiedades antioxidantes y antimicrobianas que protegen la salud al amortiguar el oxígeno activo y los radicales libres (Bakir et al., 2018). Sin embargo, el desconocimiento de las propiedades nutricionales y farmacéuticas de *Pleurotus* spp. no ha permitido que este sector agrícola en Ecuador sea explotado económicamente. En el sector agrícola, donde la agroindustria tiene que tomar innumerables decisiones todos los días y las intrincadas complejidades que involucran los diversos factores que influyen en ellas, es necesaria la estimación precisa del rendimiento de los muchos cultivos involucrados en la planificación. Las técnicas de minería de datos son un enfoque necesario para lograr soluciones prácticas y efectivas a este problema. La agricultura ha sido un objetivo obvio para el big data (Majumdar et al., 2017).

Por otro lado, Ecuador es considerado uno de los países más importantes para la producción de camarón. Existe un interés creciente en expandir la acuicultura utilizando especies y tecnologías alternativas (Naylor et al., 1998; Naylor et al., 2000). La acuicultura es el segundo componente más grande de la economía ecuatoriana, después de los combustibles fósiles (Naylor et al., 2011). Esta expansión es atribuible, casi en su totalidad, a la acuicultura camaronera y ha llevado a transiciones en el uso de la tierra o la cobertura terrestre en los estuarios ecuatorianos, con manglares históricos y otras tierras estuarinas que se utilizan como estanques camaroneros (Hamilton & Stankwitz, 2012). La producción más alta de camarón en Ecuador se debe a dos factores: la producción de camarón en Ecuador ha sido tradicionalmente semi-intensiva, usando intercambio de alimento y agua pero sin aireación, y hay mucha tierra deshabitada adecuada para grandes estanques y granjas en Ecuador (Boyd et al., 2021). A fines de 2009, el país contaba con 175 mil hectáreas de camaroneras activas que representan a 2578 empresas acuícolas, con una producción de exportación de 450 millones de libras que representan el 34% del total de productos elaborados (PROECUADOR, 2013). El incremento de la actividad camaronera también ha representado una gran fuente de trabajo en el país (Carrillo, 2009). Con el fin de diversificar la producción acuícola en el Ecuador, se han llevado a cabo

varios proyectos para la producción de camarón, *Litopenaeus stylirostris* (Rivera et al., 2018), *Sciaenops ocelatus* (Guartatanga et al., 1993) y *Seriola rivoliana* (Blacio et al., 2003). Ecuador produjo 510 000 toneladas métricas de camarón blanco (*Litopenaeus vannamei*) en 2018 (Boyd et al., 2001). En 2002, como consecuencia del síndrome de la mancha blanca, se desarrollaron métodos alternativos de producción de camarón, como el cultivo en estanques cubiertos que permitía un menor intercambio de agua y un nivel de temperatura más constante y el llamado sistema "onshore", que consistía en cultivar camarón a muy baja salinidad utilizando agua de pozos y ríos en zonas agrícolas de las provincias de Manabí y Guayas. La alimentación suplementaria está asociada con esta actividad, lo que influye en el costo de producción del camarón (Lawrence et al., 1997). La estrategia y la optimización de la alimentación son aspectos de importancia en acuicultura que implican la formulación de diferentes dietas (pellets o gránulos). El contenido de nutrientes presente en los gránulos influirá en el crecimiento, la supervivencia y los productos de desecho excretados por los camarones (Smith et al., 2002). En la formulación de gránulos, es necesario mantener los valiosos nutrientes dietéticos utilizando aglutinantes (Partridge et al., 1999). Los aglutinantes afectan la estabilidad de los gránulos de tres maneras: al reducir los vacíos, lo que da como resultado un gránulo más compacto y duradero que actúa como adhesivo, une las partículas y ejerce una acción química sobre los ingredientes y altera la naturaleza del alimento, obteniendo un gránulo más duradero (Palma et al., 2008). Los aglutinantes se utilizan para reducir la lixiviación de medicamentos aplicados a alimentos balanceados y medicamentos como antibióticos, vitaminas, ácidos orgánicos. Este tipo de producto como mezcla de gluten en la dieta se puede utilizar para obtener los valores más altos de digestibilidad aparente de proteínas (ADP) y digestibilidad aparente de materia seca (ADMD) (Arguello-Guevara et al., 2013). El uso de aglutinantes como atrayentes en alimentos para camarones no es común.

La Minería de Datos es el proceso de extracción de big data útiles siendo su principal objetivo encontrar información oculta o implícita, que no es posible obtener por métodos estadísticos convencionales. La entrada al proceso de minería está formada generalmente por registros de bases de datos operativas o almacenes de datos (Febles Rodríguez & González, 2002) La minería de datos es un proceso de extracción de información y búsqueda de patrones de comportamiento que se ocultan a primera vista entre grandes cantidades de información (Rodríguez Suárez & Díaz Amador, 2011),

existen varios algoritmos y técnicas que son utilizados en el sector agrícola como: Algoritmos de K-means, K-medoids (PAM). Meter citas de los algoritmos y de su uso en el sector agrícola Los algoritmos de particionamiento de agrupamiento, como K-means y K-medoids (PAM) asigna objetos en K (número de cluster predefinido) clusters, y de forma iterativa reasignar objetos para mejorar la calidad de los resultados de los clusters. Los algoritmos de agrupación jerárquica asignan objetos en grupos estructurados en árbol, es decir, un grupo puede tener puntos de datos representantes de clusters de bajo nivel (Han & Kamber, 2001). La idea de los métodos de agrupación en clusters basados en densidad consiste en que para cada punto de un grupo, la densidad en el vecindario debe alcanzar algún umbral. La idea del algoritmo de agrupamiento basado en densidad es que, para cada punto de un grupo, la vecindad de una unidad de distancia dada, debe contener al menos un número mínimo de puntos (Ester et al., 1996).

En este proyecto de investigación, el énfasis se pone en los métodos de cluster. Los clusters se han aplicado en muchas áreas de investigación como matemáticas, ingeniería, economía, marketing, aprendizaje automático, reconocimiento de patrones, genética, bioinformática, psicología, biología, compresión de datos y recuperación de información (Guler et al., 2002). Hay muchos enfoques diferentes para el análisis de cluster, muchos de ellos son simplemente técnicas de análisis de datos sin un modelo probabilístico claro y bien fundado detrás (Flury, 1997). Por esta razón, muchos profesionales de la estadística han percibido los métodos de agrupamiento como una colección de técnicas principalmente heurísticas. Sin embargo, también existen enfoques de clusters basados en modelos probabilísticos bien fundamentados (Bock, 1996). En este trabajo, nos enfocamos en este modelo basado enfoque de agrupamiento. Los valores iniciales de los "centroides" del cluster de datos que se repite hasta que se alcanza la convergencia o para un número definido de iteraciones. Un nuevo centroide para un cluster se calcula en función de cada muestra de datos que pertenece a ese grupo, y los centroides iniciales generalmente se eligen al azar dependiendo de la aplicación de algoritmos de tipo K-means, y K-medoids (PAM= partitioning around medoids) (Hot & Popović-Bugarin, 2016; Bezdek, 1984; Atkinson, & Riani, 2007; García-Escudero & Gordaliza, 1999).

Se aplican técnicas de minería de datos como el PCA-Biplot y el algoritmo de K-means para verificar la influencia del uso de residuos agrícolas en la producción de hongos comestible y de producción de insumos acuícolas. El uso de estas técnicas

permitirá elegir la especie biológica con la que se obtendrán los mejores parámetros comerciales.

1.2. Formulación del problema

¿Es posible mejorar la elección de productos de gran importancia económica en el Ecuador con altos parámetros biológicos y comerciales usando herramientas cuantitativas de análisis de datos multivariantes?

1.2.1. Sistematización:

- ¿Es la gramática K-means la mejor alternativa para estudiar los datos biológicos y comerciales de productos de gran importancia económica en el Ecuador?
- ¿Se pueden integrar las metodologías K-means para el análisis de datos biológicos y comerciales de productos de gran importancia económica en el Ecuador?
- ¿Al utilizar el Análisis de Componentes Principales (PCA-Biplot) se logra ofrecer una explicación más detallada acerca de los parámetros biológicos y comerciales de productos de gran importancia económica en el Ecuador?

1.3. Objetivos de la investigación

1.3.1. Objetivo general

Evaluar la gramática de PCA-Biplot y del algoritmo de K-means para inspeccionar y clasificar datos multivariantes y obtener una mejor correlación entre parámetros biológicos y comerciales de diferentes variedades de productos de gran importancia económica en el Ecuador.

1.3.2. Objetivos específicos

- Realizar una exhaustiva revisión bibliográfica sobre técnicas de Big data: K-means desde el punto de vista del algebra asociada y desde el punto de vista computacional.
- Realizar una exhaustiva revisión bibliográfica sobre el estado del arte de las técnicas multivariantes: Análisis de Componentes Principales (PCA-Biplot).

- Generar las matrices de información de características biológicas y comerciales de diferentes variedades de productos de gran importancia económica en el Ecuador.
- Analizar la interrelación entre los diferentes parámetros biológicos y comerciales de diferentes variedades de productos de gran importancia económica en el Ecuador.

1.4. Justificación de la investigación

La Ciencia de Datos es un término que se ha popularizado en los últimos años, para describir todo esfuerzo para extraer conocimiento a partir de datos, por ello es similar a otros términos como minería de datos, aprendizaje automático o análisis predictivo.

Es un término equivalente a "Minería de Datos entendida como el proceso de extracción de conocimiento (relaciones, patrones, anomalías) en un conjunto de datos, usando técnicas de modelización provenientes de otras ramas como la estadística y aprendizaje automático (Witten et al., 2011), de tal manera que convierten los datos en conocimiento e información disponible (Tsiptsis & Chorianopoulos, 2009) y permite a las personas tomar decisiones basadas en información.

Los algoritmos, se clasifican en supervisados, no supervisados y de reforzamiento. Los primeros orientados a resolver problemas para "predecir" eventos, los segundos buscan encontrar asociaciones o relaciones existentes entre los objetos de estudio, y los últimos a aprender por si mismos a tomar decisiones para conseguir un objetivo. Además, están desarrollándose algoritmos Semi-Supervisados que se sitúan entre los dos primeros. (Bucheli & Thompson, 2014). Entre los más populares, por ser los más utilizados en Minería de Datos, están C4.5, k-medias, SVM, EM, KNN, Naive Bayes y CAR. Estos algoritmos se caracterizan por haber sido probados por la comunidad científica, proporcionando resultados satisfactorios y muchos de ellos por la facilidad para entenderlos, implementarlos e interpretarlos. .

Esta tesis se centra en el uso del PCA-Biplot y del algoritmo de K-means. Estos métodos se utilizan para categorizar los diferentes parámetros biológicos y comerciales de diferentes variedades de productos de gran importancia económica en el Ecuador.

1.5. Hipótesis de trabajo

Si las técnicas multivariantes han servido para agrupar datos de diversos campos como: social, administrativo y financiero, es posible utilizar estas técnicas estadísticas para poder agrupar y seleccionar especies biológicas con los mejores parámetros comerciales.

CAPÍTULO II

MARCO TEÓRICO

CAPÍTULO II

2. MARCO TEÓRICO

2.1. Técnicas de Minería de Datos

Tal como ya hemos señalado en la introducción, Data Mining engloba un conjunto de técnicas que buscan descubrir patrones, tendencias o relaciones existentes en una gran cantidad de datos. En los últimos años se ha producido un interés generalizado y creciente, motivado no sólo por la mayor competitividad en el mercado sino también, por la explosión de datos disponibles (Big Data). Predecir el comportamiento humano permite, aumentar la satisfacción de sus clientes, combatir el riesgo financiero, descubrir fraudes, aumentar las ventas, curar enfermedades, mejorar la calidad de vida, conseguir un entorno más sostenible, etc. No obstante, hay que tener en cuenta que tener más datos no implica tener mayor información o mayor conocimiento.

La tarea de minería de datos es el análisis automático o semi-automático de grandes cantidades de datos para extraer patrones interesantes desconocidos, como grupos de registros de datos (análisis clúster), registros poco usuales (la detección de anomalías) y dependencias (minería por reglas de asociación). Estos patrones pueden ser utilizados en análisis posteriores o, por ejemplo, en máquinas de aprendizaje y/o análisis predictivo.

Existen varios modelos utilizados en Data Mining, entre ellos CRISP-DM (Cross Industry Standard Process for Data Mining), que consta de 6 fases: Comprensión del área de trabajo, Comprensión de los datos, Preparación de los datos, Modelización, Evaluación y Despliegue. El modelo CRISP-DM es el más utilizado por expertos en minería de datos (Kurgan & Musilek, 2006), puede ser aplicado en todo proyecto de minería de datos, independientemente del área o de herramientas informáticas específicas.

En la práctica un proyecto de minería de datos implica la ejecución de varias fases, donde la modelización es una de ellas. Se han desarrollado varias propuestas de este proceso desde los años 90, varios de ellos fueron construidos desde una perspectiva orientada a la aplicación académica y la mayoría con aplicación comercial (Kurgan & Musilek, 2006) y tiene 6 fases:

1) Comprensión del problema: Incluye la comprensión de los objetivos, el conocimiento del estado del arte, la alineación de los objetivos de minería con los organizacionales, y la planificación del proyecto.

2) Selección del conjunto de datos y las variables relevantes: Se refiere tanto a las variables objetivo que se quiere predecir o inferir, como a las variables independientes que sirven para explicar la respuesta. Estos datos pueden ser: estructurados (tradicionales bases de datos relacionales), no-estructurados (audio, video, texto, el cuerpo de texto de documentos de emails o de páginas web), y semi-estructurados (emails, o los que sirven para intercambiar información vía web).

3) Preprocesamiento: Filtrado de Datos. Búsqueda de valores no válidos, atípicos o incorrectos, discretización, etc.

4) Selección y aplicación de la técnica de minería de datos para construir el modelo predictivo, de clasificación o de segmentación: Es la implementación de los algoritmos de aprendizaje automático, de acuerdo a los objetivos buscados. La fase incluye el examen de las diferentes alternativas de modelización y su validación.

La modelización comienza con la construcción con "datos de entrenamiento" y la parametrización de los algoritmos hasta obtener un desempeño satisfactorio (exactitud, nivel de error, costos de clasificación). La validación del modelo implica la revisión de su desempeño sobre "datos de prueba" para garantizar su estabilidad y aplicabilidad a futuros eventos. Entre los métodos de validación tenemos el uso de datos de prueba externos, validación cruzada, validación por división y el muestreo con reemplazamiento (bootstrap). Así, obtenemos el mejor modelo para nuestros datos.

5) Extracción de conocimiento: A partir de la minería de datos se obtiene un modelo de conocimiento que captura y representa patrones de comportamiento ocultos generalmente en los datos, bien sea asociaciones de sujetos con perfiles similares, o grupos de variables con altas covariaciones, que nos van a permitir simplificar el problema y presentar los resultados al usuario de una forma comprensible para ellos.

6) Interpretación, evaluación y puesta en funcionamiento: Los resultados deben validarse y las conclusiones deben contrastarse con la realidad antes de su explotación.

Las técnicas de la minería de datos proceden de la estadística y de la inteligencia artificial. En la minería de datos es difícil establecer una separación de ambos, pero es necesario destacar la forma como ambas realizan las generalizaciones, la estadística sobre una muestra representativa de la población, mientras que el aprendizaje automático sobre las bases de datos disponibles (Jordan, 2014).

Las técnicas más utilizadas son la regresión lineal (simple, múltiple, logística, ridge,...), los árboles de decisión (Algoritmos ID3, C4.5, CHAID, THAID, CART,...), análisis de cluster (K-means, Kmedoids, ...), reglas de asociación, redes neuronales (perceptrón, perceptrón multicapa, mapas de Kohonen, ...), máquinas de vectores de soporte (SVM) y técnicas factoriales de reducción de la dimensión (ACP, BIPLOT, MDS, ...).

Es importante tener en cuenta que la terminología usada en el aprendizaje automático es diferente a la usada en estadística aunque traduzcan ideas similares. En Aprendizaje automático, hablamos de si Aprendizaje Supervisado si el objetivo es predecir un evento (valor específico de una variable categórica) o estimar valores de una variable continua. Una variable tendrá un rol como objetivo y las demás serán las predictoras (de entrada). Los modelos construidos están "supervisados" por la relaciones evaluadas entre la variable objetivo y las predictoras.

La bibliografía sobre el análisis de clúster y árboles de decisión, pone de manifiesto terminología, métodos y aproximaciones contradictorias, dependiendo de las distintas ramas de la ciencia en las que se aplican, ya que cada rama de la ciencia tiene sus preferencias para la construcción de grupos. Existen muchos algoritmos (supervisados y no supervisados) para la creación de cluster y árboles de decisión. La mayoría de los artículos están enfocados al estudio del coste computacional pero la es muy escasa a la hora de analizar las propiedades algebraicas de los algoritmos y el uso de software para su aplicación, desde un punto de vista crítico comparado.

2.1.1. Algoritmo de K-means

El K-means es una técnica de minería de datos para la agrupación (Hartigan, 1975; Yang et al., 2000). Dado un conjunto de datos con clasificación desconocida, el objetivo es encontrar una partición del conjunto en la que los datos similares se agrupan en el mismo grupo. El algoritmo clásico fue propuesto por Forgy en 1965, el cual fue modificado por McQueen en 1967.

En la propuesta de Forgy primero se hacen todas las asignaciones y luego se recalculan los centroides; en la de McQueen inmediatamente después de haber asignado un individuo a un clúster, el centro de gravedad es recalculado. Esto es mucho más costoso computacionalmente.

La medida de similitudes entre las muestras de datos se proporcionan utilizando una distancia adecuada: muestras que están cerca de cada otros se consideran similares. El parámetro K en el algoritmo de K-means juega un papel importante ya que especifica el número de clusters en los que los datos deben ser particionados.

La idea detrás del en la que está basado el algoritmo K-means es bastante simple. Dada una determinada partición de los datos en K clusters, los centros de los clusters se pueden calcular como la media de todas las muestras observaciones pertenecientes a un cluster. El centro del cluster puede considerarse como el representativo del cluster, porque el centro está bastante cerca de a todas las observaciones en el cluster, y por lo tanto es similar a todos ellas. De ello se deduce que un cluster contiene datos similares si todas las unidades a clasificar están más cerca de su centro y no del centro de algún otro grupo. Por tanto, cuando las muestras pertenecientes a un cluster están más cerca del centro de un grupo diferente, el algoritmo K-means mueve las muestras de datos correspondientes desde su grupo original al nuevo grupo.

La idea principal del algoritmo k-medias consiste en los siguientes pasos:

- ✓ *Inicialización:* Consiste en definir los objetos que se van a particionar, el número de grupos y el centroide para cada grupo. Para ello:
  - Se colocan k puntos en el espacio representado por los objetos que se están agrupando. Estos puntos representan los centroides del grupo inicial. Aunque existen varios métodos para definir los centroides iniciales, el más utilizado es la selección aleatoria.
  - Clasificación: Se calcula la distancia hacia todos los centroides para cada objeto y se asigna cada objeto al grupo que tiene el centroide más cercano.
  - Cálculo de centroides: cuando se hayan asignado todos los objetos, se vuelven a calcular las posiciones de los k centroides.
- ✓ *Proceso iterativo:* Mientras los centroides no cambien se procede a calcular la distancia del centroide y volver a distribuir todos los objetos según el centroide más cercano. El proceso se repite hasta que ya no hay cambio en los grupos, es decir, los k centroides no cambian después de una iteración, lo cual es equivalente a decir que el valor de la función utilizada como criterio de optimización no varía.
- ✓ *Criterio de convergencia:* Consiste en establecer el criterio de paro del algoritmo. Es posible converger cuando alcanza un número de iteraciones dado, cuando no exista un intercambio de objetos entre los grupos o converger cuando la diferencia entre los centroides de dos iteraciones consecutivas es más pequeña que un umbral dado. Si la condición de convergencia no se satisface, se repiten los pasos anteriores del algoritmo.

En este algoritmo se usa como una métrica la distancia euclídea y la varianza como medida de dispersión entre los grupos. El algoritmo tiene como objetivo maximizar una función objetivo, denominada suma de errores cuadráticos (SSE) o también conocida como

sumas residuales de cuadrados (RRS). La función objetivo se define como:

$$SSE = J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2$$

Dónde:

$\left\| x_i^j - c_j \right\|^2$ es una medida de distancia elegida entre un punto de datos x^j y el centro de clúster c_j, es un indicador de la distancia de los n puntos de datos de sus respectivos centros de clusters.

Entre las ventajas del método destacan:

- Es eficiente
- La implementación es sencilla

Como desventajas se pueden señalar:

- Se necesita conocer k de antemano, es decir, requiere que se indique previamente el número de clusters a crear.
- Las agrupaciones resultantes pueden variar dependiendo de la asignación aleatoria inicial de los centroides. El algoritmo es significativamente más sensible a los centros de agrupamiento seleccionados al azar inicialmente, por tanto, el resultado obtenido es dependiente de la selección inicial de los centroides de los clústeres y puede converger a óptimos locales. Se recomienda repetir el proceso de clustering entre 25-50 veces y seleccionar como resultado definitivo el que tenga menor suma total de varianza interna.
- Presenta problemas de robustez frente a outliers.
- No trata datos nominales.

Un esquema del algoritmo de K-medias de McQueen es dado en la Figura 1.

```
randomly assign each sample to one of the k clusters S(j), 1 ≤ j ≤ k
compute the center c(j) for each cluster S(j)
while (clusters are not stable)
   for each sample Sample(i)
      compute the distances between Sample(i) and all centers c(j)
      find j* such that c(j*) is the closer to Sample(i)
      assign Sample(i) to the cluster S(j*)
      recompute the centers of changed clusters
   end for
end while
```

Figura 1. El Algoritmo K-means

El algoritmo de K-means puede verse como un algoritmo de optimización, en que la función f que se minimizará es la suma de todas las distancias al cuadrado entre cada muestra y el centro de su grupo. La función f no es convexa en general. El algoritmo K-means es un algoritmo para la optimización local, porque identifica una secuencia de particiones en grupos que tienen estrictos valores de función decrecientes. Por lo tanto, el algoritmo de K-means es capaz de encontrar solo uno de los mínimos locales de la función f, que puede o no corresponder al mínimo global. Por esta razón, el algoritmo K-means generalmente se realiza varias veces utilizando diferentes particiones iniciales. La partición correspondiente al valor más pequeño de la función f se considera la solución óptima. Vale la pena señalar que el algoritmo K-means pertenece a la categoría de algoritmos de maximización de expectativas (EM), que son elegantes y métodos poderosos para encontrar soluciones de máxima verosimilitud para modelos con variables latentes (Dempster et al. 1977).

Hay muchas aportaciones posteriores que modifican estos algoritmos de k-medias; citamos los más importantes: Fuzzy C-means (Dunn, 1973 ; Bezdek, Ehrlich & Full, 1984); K-medoids (PAM) (Kaufman & Rousseeuw, 1990); K-means genético (Krishna & Murty, 1999); K-modes (Chaturvedi et al., 2001); J-means (Hansen & Mladenovic, 2001); Y-means (Guan et al., 2003); Kernel K-Means (Dhillon, Guan & Kulis, 2004); K-means ++ (Arthur & Vassilvitskii, 2007); MINI-BATCH K-means (Sculley, 2010); Spherical K-Means (Hornik et al., 2012); Distributed K-means (Oliva et al., 2013); Sequential K-Means (Liberty et al., 2014). Sin embargo, la más utilizada sigue siendo la propuesta inicial de Forgy, salvo en casos específicos de tratamiento de Big Data.

El algoritmo de K-means no estima las covarianzas de los clusters; solo considera la media de los clusters (Bishop, 2006). Existe una versión de asignación dura del modelo de mezcla gaussiana con matrices de covarianzas generales, conocido como algoritmo elíptico de K-means (Sung & Poggio, 2009).

2.1.2. PCA Biplot

PCA es una técnica estadística que transforma linealmente un conjunto de p variables en un conjunto con un número menor **(k)** de variables no correlacionadas que explican una parte sustancial de la información del conjunto original. Las p variables originales **(X1,...,Xp)** se transforman en variables **p (Y1,...,Yp),** de modo que Y1 es la que explica la mayor parte de la variabilidad total de los datos, Y2 explica la segunda parcela más grande y pronto.

Los principales objetivos del análisis de componentes principales son: reducir la

dimensionalidad de los datos, obtener combinaciones de variables interpretables y, finalmente, discriminar y comprender la estructura de correlación de las variables.

El análisis se realiza con el fin de resumir el patrón de correlación entre variables y en algunos casos es posible llegar a conjuntos de variables que no están correlacionadas, dando lugar así a una agrupación de las mismas. Algebraicamente, los componentes principales son combinaciones lineales de variables originales. Geométricamente, los componentes principales son las coordenadas de los puntos de muestreo en un sistema de ejes obtenido al rotar el sistema original de ejes en la dirección de máxima variabilidad de los datos.

El PCA depende únicamente de la covarianza **(Σ)** o de la matriz de correlación **(ρ)** de X1,...,Xp. No requiere ninguna suposición acerca de la forma de distribución multivariada de estas variables.

Según Zou & Hastie (2005) El éxito del PCA se debe a las siguientes propiedades óptimas:

1) Las componentes principales capturan secuencialmente la máxima variabilidad entre las columnas de X, lo que garantiza que haya una mínima pérdida de información.

2) Las componentes principales son no correlacionadas, por lo que se puede hablar de una componente principal sin hacer referencia a otras. El PCA permite transformar las variables originales, en general correlacionadas, en nuevas variables no correlacionadas, facilitando la interpretación de los datos Sin embargo, el PCA sufre del hecho de que cada componente principal es una combinación lineal de todas las variables originales, por lo que es a menudo difícil interpretar los resultados.

2.1.2.1.Interpolación de muestras

En el PCA Biplot, la interpolación se logra proyectando ortogonalmente cada punto de muestra en el plano biplot (Gabriel, 1971; Galindo, 1986; Gower et al., 2011). Supongamos $\mathbf{d_o}$ que es una muestra centrada. Con $\mathbf{V_{[r]}^T V_{[r]} = I_r Z V_{[r]}^T}$ se puede escribir como:

$$\mathbf{Z V_{[r]}^T = D_o V_{[r]} V_{[r]}^T = D_o V_{[r]} \left(V_{[r]}^T V_{[r]}\right)^{-1} V_{[r]}^T}$$

La representación de la muestra $\mathbf{d_o}$ proyectada en el plano Biplot es dado por:

$$\mathbf{d_{oproj}^T = d_o^T V_{[r]} \left(V_{[r]}^T V_{[r]}\right)^{-1} V_{[r]}^T = z^T V_{[r]}^T}$$

Como resultado, las coordenadas de las proyecciones de la muestra $\mathbf{d_o}$ en el plano biplot están dadas por $\mathbf{z^T}$**.** Es decir, la muestra $\mathbf{d_o}$ se interpola en el plano biplot por:

$$\mathbf{z^T} = \mathbf{d_o^T V_{[r]}}$$

Considere un punto **z* (r x 1)** descrito en términos del sistema de coordenadas del plano biplot. Los puntos representan **z*** en el plano biplot también tiene una representación de coordenadas $\mathbf{d_o^*}$ relativo a la ejes del espacio P-dimensional. Esto es cierto porque el plano biplot es un subespacio del espacio P-dimensional. Las coordenadas del punto en el espacio P–dimensional esta dado por $\mathbf{d_{oproj}}$ y las coordenadas del punto en el espacio r-dimensional están dadas por **z***.

Así que cualquiera punto **z* (r x 1)** en términos de la base del plano biplot también es un punto $\mathbf{d^*_{oproj}}$ **(P x 1)** en términos de la base para el espacio P-dimensional de $\mathbf{D_o}$ y tal punto se proyectará sobre sí mismo. Para ser preciso,

$$\mathbf{d^T_{oproj}} = \mathbf{d^T_{oproj}}\ \mathbf{V_{[r]}}\ \mathbf{V^T_{[r]}}$$

Con la interpolación de un punto $\mathbf{d_o^*}$ dada por $\mathbf{z*^T} = \mathbf{d_o^{*T} V_{[r]}}$, $\mathbf{d^T_{oproj}} = \mathbf{z*^T}\ \mathbf{V^T_{[r]}}$. . Por lo tanto, muestra $\mathbf{d_o^*}$ es predicho por

$$\mathbf{d_o^{*T}} = \mathbf{z*^T}\ \mathbf{V^T_{[r]}}$$

2.1.2.2. Predicción de ejes

Los marcadores de columna para el PCA Biplot están definidos por las filas de la matriz $\mathbf{V_{[r]}}$. El eje del factor de calibración reemplazando $\mathbf{b_j}$ por $\mathbf{V^T_{[r]}}\ \mathbf{e_k}$ dando el factor de calibración para el eje $\mathbf{k^{th}}$ como:

$$\propto = \frac{\mathbf{u^*}}{\mathbf{e_k^T}\ \mathbf{V_{[r]}}\ \mathbf{V^T_{[r]}}\mathbf{e_k}}$$

Donde $\mathbf{e_k}$ es el vector unitario con ceros excepto uno en la posición $\mathbf{k^{th}}$. Por lo tanto, el marcador **u*** sobre el eje de predicción Biplot se obtiene mediante la expresión:

$$\frac{\mathbf{u}^*}{\mathbf{e}_\mathbf{k}^\mathbf{T}\,\mathbf{V}_{[\mathbf{r}]}\,\mathbf{V}_{[\mathbf{r}]}^\mathbf{T}\mathbf{e}_\mathbf{k}}\,\mathbf{V}_{[\mathbf{r}]}^\mathbf{T}\mathbf{e}_\mathbf{k}$$

2.1.2.3. Medida de ajuste para PCA Biplots

Dado que se ha encontrado el mejor plano bidimensional biplot en el que proyectar los puntos, tienen las coordenadas de los puntos cuando se proyectan en el plano, la calidad de la representación proporcionada por estas proyecciones es necesaria para determinar la idoneidad de la representación de los datos de la matriz original $\mathbf{D_o}$. En otras palabras, ¿qué tan cerca está $\widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}$ de $\mathbf{D_o}$? Para evaluar la calidad de la representación, considerar $\mathbf{D_o}$ dividida en una parte ajustada $\widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}$ y una parte residual $\left(\mathbf{D_o} - \widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}\right)$ (Oyedele & Lubbe, 2015). Es decir,

$$\mathbf{D_o} = \widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]} + \left(\mathbf{D_o} - \widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}\right)$$

La ecuación antes mencionada se puede considerar como una descomposición ortogonal $\mathbf{D_o}$, en que

$$\|\mathbf{D_o}\|^2 = \left\|\widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}\right\|^2 + \left\|\mathbf{D_o} - \widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}\right\|^2$$

.

Ésta es la condición de ortogonalidad de la descomposición por suma de cuadrados, los dos tipos de ortogonalidad. Primero es

$$\mathbf{D_o}\mathbf{D_o^T} = \widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}\ \widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}^\mathbf{T} + \left(\mathbf{D_o} - \widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}\right)\left(\mathbf{D_o} - \widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}\right)^\mathbf{T}$$

Esto se da por:

$$\mathbf{D_o}\left(\mathbf{D_o} - \widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}\right)^\mathbf{T} = \mathbf{D_o}\mathbf{D_o^T} - \mathbf{D_o}\,\mathbf{V}_{[\mathbf{r}]}\,\mathbf{V}_{[\mathbf{r}]}^\mathbf{T}\mathbf{D}_{\mathbf{o}[\mathbf{r}]}^\mathbf{T}$$

$$= \mathbf{D_o}\mathbf{D_o^T} - \mathbf{D_o}\mathbf{D_o^T}$$

$$= \mathbf{0}$$

y

$$\mathbf{D_o}\left(\mathbf{D_o} - \widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}\right)^\mathbf{T} = \mathbf{D_o}\,\mathbf{V}_{[\mathbf{r}]}\,\mathbf{V}_{[\mathbf{r}]}^\mathbf{T}\mathbf{D}_{\mathbf{o}[\mathbf{r}]}^\mathbf{T} - \mathbf{D_o}\,\mathbf{V}_{[\mathbf{r}]}\,\mathbf{V}_{[\mathbf{r}]}^\mathbf{T}\mathbf{V}_{[\mathbf{r}]}\,\mathbf{V}_{[\mathbf{r}]}^\mathbf{T}\mathbf{D}_{\mathbf{o}[\mathbf{r}]}^\mathbf{T}$$

$$= \mathbf{D_o D_o^T} - \mathbf{D_o}\, \mathbf{V_{[r]}}\ \mathbf{V_{[r]}^T D_{o\,[r]}^T}$$

$$= \mathbf{D_o D_o^T} - \mathbf{D_o D_o^T}$$

$$= \mathbf{0}$$

Para $\mathbf{V_{[r]}^T\, V_{[r]} = I_r}$ **y** $\mathbf{V_{[r]} V_{[r]}^T = I_p}$ **.** El siguiente tipo ortogonal es:

$$\mathbf{D_o^T D_o} = \widehat{\mathbf{D_{o\,[r]}^T}}\widehat{\mathbf{D}}_{\mathbf{o[r]}} + \left(\mathbf{D_o} - \widehat{\mathbf{D}}_{\mathbf{o[r]}}\right)^{\mathbf{T}} \left(\mathbf{D_o} - \widehat{\mathbf{D}}_{\mathbf{o[r]}}\right)$$

donde

$$\mathbf{D_o^T}\left(\mathbf{D_o} - \widehat{\mathbf{D}}_{\mathbf{o[r]}}\right) = \mathbf{D_o^T D_o} - \mathbf{D_o^T D_o V_{[r]} V_{[r]}^T}$$

$$= \mathbf{D_o^T D_o} - \mathbf{D_o^T D_o}$$

$$= \mathbf{0}$$

y

$$\widehat{\mathbf{D_{o\,[r]}^T}}\left(\mathbf{D_o} - \widehat{\mathbf{D}}_{\mathbf{o[r]}}\right) = \mathbf{V_{[r]} V_{[r]}^T D_o^T D_o} - \mathbf{V_{[r]} V_{[r]}^T D_o^T D_o V_{[r]} V_{[r]}^T}$$

$$= \mathbf{D_o^T D_o} - \mathbf{V_{[r]} V_{[r]}^T D_o^T D_o}$$

$$= \mathbf{D_o^T D_o} - \mathbf{D_o^T D_o}$$

$$= \mathbf{0}$$

2.1.2.3.1. Predictividad de muestras

El grado en que las filas de $\widehat{\mathbf{D}}_{\mathbf{o[r]}}$ concuerdan con las filas correspondientes de medidas $\mathbf{D_o}$ a la distancia lejana cada muestra está de su aproximación dimensional (Gardner-Lubbe et al., 2008). Con la suma de cuadrados de los valores aproximados para cada muestra dados por los elementos diagonales de $\left(\widehat{\mathbf{D}}_{\mathbf{o[r]}}\ \ \widehat{\mathbf{D_{o\,[r]}^T}}\right)$, expresando estas sumas de cuadrados como una proporción de sus respectivas sumas de cuadrados totales, se obtiene el poder predictivo de cada muestra. Para ser preciso,

$$\textbf{Predictividad de muestras} = \textbf{diag}\left(\widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}\ \ \widehat{\mathbf{D}_{\mathbf{o}[\mathbf{r}]}^{\mathbf{T}}}\right)\ [\textbf{diag}\,(\mathbf{D_o D_o^T})]^{-1}$$

Los valores de predicción de la muestra se encuentran entre 0 y 1, indicando que la muestra es ortogonal al plano de aproximación bidimensional biplot e implicando que la muestra está en el plano.

2.1.2.3.2. Predictividad de ejes

Se puede evaluar qué tan bien los ejes biplot individuales reproducen las variables de $\mathbf{D_o}$midiendo el grado en que las columnas de $\widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}$ concuerdan con las columnas correspondientes de $\mathbf{D_o}$. Expresando la suma de cuadrados de los valores aproximados para cada variable, dada por $\textbf{diag}\left(\widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}\ \ \widehat{\mathbf{D}_{\mathbf{o}[\mathbf{r}]}^{\mathbf{T}}}\right)$, como una proporción de su respectiva suma de cuadrados total produce la potencia predictiva de cada eje. Más precisamente,

$$\textbf{Predictividad de ejes} = \textbf{diag}\left(\widehat{\mathbf{D}}_{\mathbf{o}[\mathbf{r}]}\ \ \widehat{\mathbf{D}_{\mathbf{o}[\mathbf{r}]}^{\mathbf{T}}}\right)\ [\textbf{diag}\,(\mathbf{D_o D_o^T})]^{-1}$$

Los valores de predictividad se encuentran entre 0 y 1. Un eje de predictividad de 1 significa que todos los valores pueden leerse exactamente fuera del eje. Cuanto menor sea el eje valor predictivo, con menor precisión el eje se aproxima a los valores observados bajo esa variable.

2.1.2.3.3. Calidad de la aproximación

En general, la calidad de la aproximación se puede medir en términos del porcentaje de variación en $\mathbf{D_o}$ explicado por $\mathbf{Z = D_o V_{[r]}}$. Desde el **SVD** de $\mathbf{D_o}$, con $\mathbf{U^T U = I_p}$,

$$\mathbf{D_o^T D_o} = (\mathbf{U\Lambda V^T})^{\mathbf{T}}\ (\mathbf{U\Lambda V^T}) = \mathbf{V\Lambda U^T\ U\Lambda V^T} = \mathbf{V\Lambda^2 V^T}$$

Desde $\mathbf{V^T V = Ip} = \mathbf{VV^T}$

$$\textbf{tr}\,\{\mathbf{D_o^T D_o}\} = \textbf{tr}\,\{\mathbf{V\Lambda^2 V^T}\} = \textbf{tr}\,\{\mathbf{\Lambda^2 VV^T}\} = \sum_{\mathbf{j=1}}^{\mathbf{P}} \boldsymbol{\lambda}_{\mathbf{j}}^{\mathbf{2}} = \sum_{\mathbf{j=1}}^{\mathbf{P}} \boldsymbol{\sigma}_{\mathbf{j}}^{\mathbf{2}}$$

dónde, $\boldsymbol{\Lambda = diag\left(\lambda_1, \lambda_2, \ldots\ldots, \lambda_p\right)}$, $\boldsymbol{\lambda_j}$ es el $\mathbf{j^{th}}$ valor singular de $\mathbf{D_o}$ y $\boldsymbol{\sigma_j^2 = \lambda_j^2}$ es el $\mathbf{j^{th}}$ valor propio (y valor singular) de cov ($\mathbf{D_o}$). Por lo tanto,

$$\textbf{Calidad} = \frac{\mathbf{tr}\left\{\mathbf{D}_{\mathbf{o}\,[r]}^{\mathbf{T}}\,\mathbf{D}_{\mathbf{O}\,[r]}\right\}}{\mathbf{tr}\left\{\mathbf{D}_{\mathbf{o}}^{\mathbf{T}}\mathbf{D}_{\mathbf{o}}\right\}}$$

$$= \frac{\sum_{j=1}^{r} \lambda_j^2}{\sum_{j=1}^{r} \lambda_j^2}$$

$$= \frac{\sum_{j=1}^{r} \sigma_j^2}{\sum_{j=1}^{r} \sigma_j^2}$$

2.1.3. Reglas de asociación

Las reglas de asociación se han utilizado ampliamente desde aplicaciones comerciales tradicionales como el marketing cruzado, envío de correo adjunto, diseño de catálogo, análisis de líder de pérdidas, tienda diseño y segmentación de clientes (Agrawal et al., 1993; Srikant & Agrawal, 1995) a aplicaciones de comercio electrónico como la renovación de las páginas web (Cooley et al., 1999) y personalización web (Mobasher et al., 2000) y agricultura.

Dado un conjunto de transacciones donde cada una de las transacciones es un conjunto de elementos, una regla de asociación implica la forma X⇒Y, donde X e Y son conjuntos de elementos; X e Y se denominan cuerpo y cabeza, respectivamente. Una regla puede ser evaluada por dos medidas, llamadas confianza y apoyo. Una medida, el apoyo a la regla de asociación X⇒Y es el porcentaje de transacciones que contienen tanto el conjunto de elementos X como Y entre todas las transacciones. La confianza de la regla X⇒Y es el porcentaje de transacciones que contienen un conjunto de elementos Y entre las transacciones que contienen un conjunto de elementos X. El soporte representa la utilidad de las reglas descubiertas y la confianza representa la certeza de las reglas.

Muchos algoritmos pueden ser usados para descubrir reglas de asociación a partir de datos para extraer patrones útiles. Algoritmo a priori es una de las técnicas más utilizadas y famosas para encontrar reglas de asociación (Agrawal & Srikant, 1994). Apriori opera en dos fases. En la primera fase, se generan todos los conjuntos de elementos con soporte mínimo (conjuntos de elementos frecuentes). Esta fase utiliza la propiedad de cierre de apoyo. En otras palabras, si un conjunto de elementos del tamaño K es un conjunto de elementos frecuente,

luego todos los conjuntos de elementos a continuación (K-1) el tamaño también debe ser conjuntos de elementos frecuentes. Usando esta propiedad, los conjuntos de elementos candidatos de tamaño K se generan a partir del conjunto de elementos frecuentes de tamaño (K-1) imponiendo la restricción de que todos los subconjuntos de tamaño (K-1) de cualquier conjunto debe estar presente en el conjunto de elementos frecuentes de tamaño. (K-1). La segunda fase del algoritmo genera reglas para el conjunto de todos los elementos frecuentes.

Las reglas de asociación es una solución poderosa para extracción de reglas alternativas, porque tiene como objetivo descubrir todas las reglas en los datos y, por lo tanto, es capaz de proporcionar una imagen completa de asociaciones en un gran conjunto de datos. Sin embargo, hay dos problemas importantes con respecto a la generación de la regla de asociación. El primer problema surge de la regla cantidad y problemas de calidad. Si el soporte mínimo se establece demasiado alto, las reglas que involucran elementos raros que podrían ser de interés para no se encontrarán tomadores de decisiones. La configuración mínima del soporte bajo, sin embargo, puede causar una explosión combinatoria. En otras palabras, se generan demasiadas reglas independientemente de su interés (Tan & Kumar, 2000). Para lidiar con este problema, técnicas que permiten al usuario especificar múltiples soportes mínimos para reflejar las frecuencias de cada uno de los elementos de las bases de datos (Liu et al., 1999) o que explotan restricciones que especifican el soporte mínimo necesario para que puedan generarse los elementos de cada conjunto de datos (Wang et al., 2000). Esos enfoques, sin embargo, no consideran valores comerciales heterogéneos de las reglas de asociación. Por tanto, es necesario sugerir una aproximación para reglamentación de cantidad y calidad entre los criterios de valores comerciales que no son claros. El segundo, no es independiente del primero, si un conjunto de reglas que coincide con un contexto, algunos enfoques pueden ser aplicados para resolver esos conflictos (Barr & Feigenbaum, 1981; Hayes-Roth et al., 1983). Una forma de abordar ese problema es depender de la intervención de la inteligencia humana. Por tanto, se puede obtener un resultado satisfactorio mediante el uso de juicios de preferencia humana en la resolución de reglas en conflicto que se caracterizan por criterios conflictivos multifacéticos. En resumen, ambos problemas pueden manejarse priorizando las reglas que resultaron de la minería de datos de acuerdo con la consideración explícita de los valores comerciales.

2.1.4. Aplicaciones de técnicas de minerías de datos

Existen varias aplicaciones de las técnicas de minería de datos en el campo de la agricultura, a continuación se presentarán las aplicaciones más utilizadas.

Algunas de las técnicas de minería de datos están relacionadas con estudios sobre el clima, condiciones y previsiones. Por ejemplo, el método K-means se ha utilizado para realizar pronósticos de la contaminación en la atmósfera (Jorquera et al., 2001), el punto más cercano se aplica para simular precipitaciones diarias y otras variables meteorológicas (Rajagopalan & Lall, 1999), y diferentes posibles cambios del clima se analizan utilizando SVMs (Tripathi et al., 2006). Las técnicas de minería de datos se utilizan a menudo para estudiar las características del suelo. Como un ejemplo, el enfoque de K-means se utiliza para clasificar suelos en combinación con tecnologías basadas en GPS (Verheyen et al., 2001). Meyer et al., (2004) utilizaron el logaritmo de K-means para clasificar suelos y plantas (Camps-Valls et al. 2003) utiliza SVMs para clasificar cultivos.

Leemans & Destain, (2004) usaron el logaritmo de K-means para analizar el color imágenes de frutas mientras corren sobre cintas transportadoras. Shahin et al., (2001) utilizaron imágenes de rayos X de manzanas para monitorear la presencia de núcleos de agua, y una red neuronal es entrenada para discriminar entre manzanas buenas y malas.

El proceso de fermentación del vino se puede monitorear utilizando técnicas de minería de datos. Los sensores de gusto se utilizan para obtener datos del proceso de fermentación que se clasificarán utilizando ANNs (Riul Jr et al., 2004). Del mismo modo, los sensores se utilizan para oler la leche, es decir clasificados utilizando SVMs (Brudzewski et al., 2004). Las técnicas de minería de datos se aplican para estudiar problemas de reconocimiento de sonido. Por ejemplo, (Fagerlund, 2007) utilizó SVMs para clasificar el sonido de los pájaros y otros diferentes sonidos. Holmgren & Thuresson, (1998) usaron el logaritmo del punto más cercano para evaluar inventarios forestales y estimar variables forestales analizando imágenes de satélite. Das & Evans (1992) utilizaron ANNs para clasificar los huevos como fértiles o infértiles y (Patel et al., 1994) utilizaron la visión por computadora para reconocer las grietas en los huevos. Du & Sun, (2005) utilizaron SVMs para clasificar la salsa de pizza para untar, y (Karimi et al., 2006) utilizaron SVMs para detectar el estrés por malezas y nitrógeno en el maíz. Sin embargos, estos últimos métodos expuestos no fueron utilizados en esta investigación debido que no presentan una adecuada visualización de los clusters debido a que son eficientes con parámetros biológicos y químicos en comparación con las técnicas de minería de datos como el PCA Biplot y el logaritmo de K-means son más eficientes .

CAPÍTULO III

PRIMER CASO DE CONTRIBUCIÓN A LA MINERÍA DE DATOS

CAPÍTULO III

3. PRIMER CASO DE CONSTRIBUCIÓN A LA MINERÍA DE DATOS

3.1. Metodología

Los datos de características miceliales y culturales de las cepas de *Pleurotus* spp., parámetros de productividad, composición nutricional, actividades antioxidante y antimicrobiana de los hongos comestibles de *Pleurotu*s spp. fueron obtenidos de forma experimental por el autor de esta tesis, con la colaboración el Ing. Cristian Vargas (Gerente General de Ecuahidrolizados S.A.).

3.1.1. Material biológico

En este estudio, se utilizaron 50 cepas de *Pleurotus ostreatus* (PO) y 50 cepas de *Pleurotus djamor* (PD). Estas cepas fueron recolectadas de productores de la provincia de Guayas. Las cepas se mantuvieron en placas MEA y se depositaron en la colección de hongos del Laboratorio de Investigación y Desarrollo de Ecuahidrolizados.

3.1.2. Sustrato y suplementación

Las cepas se cultivaron utilizando dos mezclas de desechos agrícolas: 80% de bagazo de caña de azúcar y 20% de paja de trigo (S1), y 60% de paja de trigo y 40% de bagazo de caña de azúcar (S2). Las mezclas de desechos agrícolas se humedecieron durante 1 día. Posteriormente, la mezcla se colocó (1 kg de peso húmedo) en bolsas de plástico y se pasteurizó durante 10 ha 80 °C.

Después de la pasteurización y acondicionamiento, con el sustrato a temperatura ambiente, las bolsas con el sustrato se inocularon con 150 g de grano de trigo previamente colonizado con las cepas de *Pleurotus ostreatus* (PO) y las cepas de *Pleurotus djamor* (PD). Posteriormente, las bolsas con el sustrato se incubaron en una habitación oscura a una temperatura de 30 ± 1 °C.

Finalmente, una vez que el micelio de la cepa colonizo el sustrato, las bolsas con el sustrato se trasladaron a una habitación con condiciones favorables para la fructificación: se mantuvo la humedad relativa entre 85% y 90%, una temperatura de 25

± 1 °C, recirculación de aire y período de iluminación de 12 h (Valenzuela-Cobos et al., 2020).

3.1.3. Parámetros de productividad

Eficiencia biológica

La eficiencia biológica (BE) es un parámetro de productividad que explica la capacidad del sustrato para producir cuerpos fructíferos y se calculó mediante la siguiente ecuación (Thongsook & Kongbangkerd, 2011):

$$BE(\%) = \frac{peso\ fresco\ de\ hongos\ comestibles\ (g)}{peso\ seco\ del\ sustrato(g)} \times 100 \quad (1)$$

Ecuación 1. Eficiencia biológica de los hongos comestibles.

Rendimiento

El rendimiento es una variable analizada ampliamente adoptada para cultivos industriales y se calculó con la siguiente ecuación (Salmones et al., 1997):

$$Y(\%) = \frac{peso\ fresco\ de\ hongos\ comestibles\ (g)}{peso\ fresco\ del\ sustrato\ (g)} \times 100 \quad (2)$$

Ecuación 2. Rendimiento de los hongos comestibles.

Tasa de productividad (PR)

La tasa de productividad es la relación entre BE y la precocidad (es decir, días entre la inoculación y la cosecha) y se calculó utilizando la siguiente ecuación (Cardoso et al., 2020):

$$PR(\%\ por\ día) = \frac{biological\ efficiency\ (\%)}{precocity\ (days)} \quad (3)$$

Ecuación 3. Tasa de productividad de los hongos comestibles.

3.1.4. Composición nutricional

El valor nutricional de la muestra de hongos se analizó utilizando procedimientos de la AOAC sobre la composición de proteínas, grasas, carbohidratos y cenizas (AOAC, 2016). Para la estimación del contenido de proteína cruda (N × 4,38) se utilizó el método macro-Kjeldahl; el contenido de grasa bruta se determinó extrayendo un peso conocido de muestra con hexano, utilizando un aparato Soxhlet, mientras que el contenido de cenizas se determinó mediante calcinación a 600 °C (Mocan et al., 2018). El contenido total de carbohidratos (% C) se calculó utilizando la siguiente ecuación:

$$C(\%) = 100 - (\%humedad + \%proteina + \%grasa + \%cenizas) \quad (4)$$

Ecuación 4. Porcentaje de carbohidratos de los hongos comestibles.

3.1.5. Actividad antioxidante

Para evaluar la actividad antioxidante, se utilizó el ensayo de captación de radicales DPPH. Al principio, se pipetearon 30 µL del extracto y 270 µL de metanol que contenía radicales DPPH (6 x 10^{-5} mol L^{-1}) y se mezclaron en una placa de 96 pocillos. La mezcla de reacción se incubó en la oscuridad durante 30 min y la absorción se midió a 515 nm utilizando un lector de microplacas (Kostic et al., 2017). La actividad de eliminación de radicales de DPPH (RSA) se calculó como un porcentaje de decoloración de DPPH usando la siguiente ecuación:

$$RSA(\%) = \frac{ADPPH - AS}{ADPPH} \times 100 \quad (5)$$

Ecuación 5. Actividad captadora de radicales DPPH (RSA)

3.1.6. Actividad antimicrobiana

La actividad antimicrobiana se analizó utilizando las siguientes bacterias Gram negativas: *Pseudomonas aeruginosa* (ABN 187) y *Salmonella typhimurium* (ABN 572); y las siguientes bacterias Gram positivas: *Micrococcus flavus* (ABP 147) y *Staphylococcus aureus* (ABP 784). Los microorganismos se depositan en el Laboratorio de Investigación y Desarrollo de Ecuahidrolizados. Las suspensiones bacterianas se ajustaron con solución salina estéril a una concentración de 1.0 x 10^6 UFC / mL. Los extractos de hongos de *Pleurotus* spp. se disolvieron en etanol al 30%, se mezclaron con medio nutriente para bacterias (caldo de soja tríptico) que contenía inóculo bacteriano (1,0 × 10^5 UFC por pocillo) con un volumen final de 100 µL (Tsukatani et al., 2012).

3.1.7. Técnicas Multivariantes

Las herramientas de análisis de agrupamientos utilizadas fueron PCA-Biplot y el algoritmo de K-means.

3.1.8. K-means

El método de agrupación de K-means es una técnica no jerárquica que se utiliza para agrupar observaciones en K grupos. Cada elemento se asigna a un grupo con el centro más cercano. El algoritmo actualiza iterativamente los grupos para minimizar la variación de sus elementos. El algoritmo básico de K-means, que se utilizó en este estudio, se refiere a la métrica euclidiana para definir la distancia entre los elementos y los centros de los clusters (Stolz & Huertas, 2020). La distancia euclidiana se selecciona como índice de similitud y los objetivos de agrupamiento minimizan la suma de los cuadrados de los distintos tipos; es decir, minimiza (Wang et al., 2012):

$$d = \sum_{k=1}^{k} \sum_{i=1}^{k} \|(x_i - u_k)\|^2$$

donde k representa K centros de clusters, u_k representa el k-ésimo centro y x_i representa el i-ésimo punto en el conjunto de datos. La solución al centroide u_k es la siguiente:

$$\frac{\partial}{\partial u_k} = \frac{\partial}{\partial u_k} \sum_{k=1}^{k} \sum_{i=1}^{n} (x_i - u_k)^2$$

$$= \sum_{k=1}^{k} \sum_{i=1}^{n} \frac{\partial}{\partial u_k} (x_i - u_k)^2$$

$$= \sum_{i=1}^{n} 2\,(x_i - u_k)$$

$$u_k = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Además, el resultado del método de K-means depende en gran medida del número de grupos definidos de antemano. En general, el método de agrupamiento iterativo de K-means se implementa de la siguiente manera: Paso 1: Se elige un valor de K. Se utiliza como el conjunto inicial de K centroides. Paso 2: Cada uno de los objetos se asigna al grupo con el centroide más cercano. Paso 3: Se determinan los nuevos centroides de los K grupos, calculando la media de los miembros del grupo. Paso 4: Los pasos 3 y 4 se repiten hasta que no haya cambios en la función de criterio después de una iteración (Govender & Sivakumar, 2020).

Las principales ventajas del algoritmo K-means son su baja complejidad, es computacionalmente rápido, la capacidad de manejar grandes conjuntos de datos y la flexibilidad para ajustar el número de cluster. La agrupación en clusters de K-medias se

utilizó para extraer clusters del conjunto de datos que se había optimizado mediante la selección de características.

La matriz de datos se transforma a formato .txt y luego se procede a cargar la matriz en el software estadístico R, con la ayuda de Rstudio, utilizando la siguiente instrucción:

```
>DATA<-KMEDIODSCINPD2
>DATA
>rownames(DATA)<-DATA$Strains
>COMPOACTM1BIPLOT<-data.frame(KMEDIODSCINPD2)
>COMPOACTM1BIPLOT
>rownames(COMPOACTM1BIPLOT)<-COMPOACTM1BIPLOT$Strains
>COMPOACTM1BIPLOT
>COMPOACTM1BIPLOT<-COMPOACTM1BIPLOT[,2:5]
>COMPOACTM1BIPLOT
>nutri<-COMPOACTM1BIPLOT
>dim(nutri)
>class(nutri)
>row.names(nutri)
>library(gplots)
>clas<-c(rep("M1",50) , rep("M2",50))
>pca1<-prcomp(nutri)
>pca1$x
>plot(pca1$x[,1:2], pch=19, col=as.factor(clas) )
>legend("bottomleft", pch=19, legend=unique(as.factor(clas)), col=unique(as.factor(clas) , cex=1.7)
>kmeans1<-kmeans(nutri, 4, algorithm="Forgy")
>kmeans1$cluster
>library(gplots)
>balloonplot(table(kmeans1$cluster, clas))
>plot(pca1$x[,1:2], pch=19, col=as.factor(kmeans1$cluster) )
```

3.1.9. PCA-Biplot

Biplot es una aproximación de una matriz realizada sin hacer suposiciones sobre distribuciones probabilísticas subyacentes que proporciona la estructura geométrica de los datos gráficamente, mostrando la variabilidad del conjunto de individuos y variables. El prefijo bi se refiere a la representación de filas y columnas simultáneas de la matriz.

Teóricamente, en una Biplot una matriz rectangular Y de orden (nxp) y rango r, por otra de rango q (q <r), mediante su descomposición en valores singulares (DVS), es decir, $Y \cong U\Sigma V'$

donde U y V son matrices de vectores singulares ortonormales tales que $U'U = V'V = I$ (donde I es la matriz identidad) y $\sum$ es una matriz diagonal que contiene los α_k

valores singulares más grandes.

Para garantizar la representación es necesaria una factorización como: Y ≅ (UΣS) (Σ1-SV´) = AB ', siendo A y B las matrices que contienen las coordenadas de los (n + p) vectores o marcadores filas ai y columnas bj para usar sobre el gráfico (i = 1,…, n; j = 1,…, p) (Cárdenas et al., 2007).

La matriz de datos se transforma a formato .txt , el software utilizado para realizar estos análisis fue MULTIBIPLOT, desarrollado por (Vicente-Villardón, 2010a), disponible en la página web: http://biplot.usal.es/multbiplot y el MultbiplotR en R (Vicente-Villardón, 2010b), disponible en el sitio web: https://CRAN.R-project.org/package=MultBiplotR. Este programa fue escrito en lenguaje R.

```
>library(MultBiplotR)
>data<-KMEDIODSCINPD2
>data
>X= data[,2:5]
>X
>bipUNEMI=PCA.Biplot(X, Scaling = 5)
>bipUNEMI
>summary(bipUNEMI)
>Inercias=data.frame(paste("Eje",1:length(bipUNEMI$EigenValues)),bipUNEMI$EigenValue
s, >bipUNEMI$Inertia, bipUNEMI$CumInertia)
>colnames(Inercias)=c("Eje", "Valor Propio", "Inercia", "Inercia acumulada")
>library(knitr)
>kable(Inercias)
>kable(bipUNEMI$ColContributions)
>plot(bipUNEMI, mode="ah", margin=0.05, ShowBox=TRUE)
>bipUNEMI=AddCluster2Biplot(bipUNEMI, NGroups=4, ClusterType="hi",
method="ward.D", Original=TRUE)
>plot(bipUNEMI, PlotClus=TRUE,ShowAxis=TRUE)
```

3.2. Resultados y discusiones

Se utilizaron técnicas de minería de datos y métodos de agrupación para realizar el estudio de determinar la viabilidad del uso de residuos agrícolas de la provincia del Guayas en el cultivo de las cepas *Pleurotus ostreatus* y *Pleurotus djamor* y evaluar su influencia en parámetros comerciales: eficiencia biológica, rendimiento del cultivo, tasa de producción, composición nutricional, actividades antioxidantes y antimicrobianas.

La numeración de las cepas cultivadas en las dos mezclas de agricultura se realizó utilizando la siguiente distribución:

1-50: Cepas de *Pleurotus ostreatus* o *Pleurotus djamor* cultivadas en la mezcla S1.

51-100: Cepas de *Pleurotus ostreatus* o *Pleurotus djamor* cultivadas en la mezcla S2.

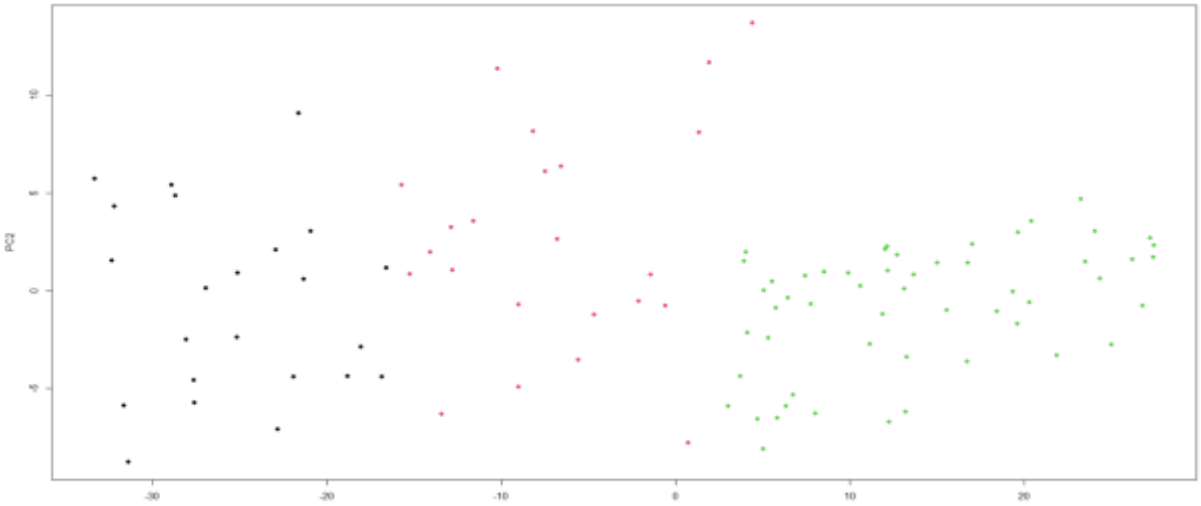
3.2.1. Parámetros de productividad

La Figura 3 muestra la aplicación del método del algoritmo de agrupación de K-means a 100 objetos que tienen tres variables, cada una de las cuales utiliza el software RStudio.

El gráfico (a) presenta el uso de tres clusters para los parámetros de productividad de cepas de *Pleurotus ostreatus* cultivadas sobre residuos agrícolas de la provincia de Guayas, mientras que en el gráfico (b) indica el uso de tres clusters para los parámetros de productividad de cepas de *Pleurotus djamor* cultivadas en las dos mezclas de sustratos. Los resultados muestran la distribución normal de 100 puntos de datos alrededor de tres grupos en cada gráfico. El tamaño de cada grupo está relacionado con la cantidad de puntos de datos, en el gráfico (a): el tamaño del Grupo 1 (color rojo) es 34, el tamaño del Grupo 2 (color negro) es 27 y el tamaño del Grupo 3 (color verde) es 39. Las cepas de *Pleurotus ostreatus* cultivadas en las dos mezclas pertenecientes al Cluster 3 no mostraron una relación con las cepas de *Pleurotus ostreatus* cultivadas en las dos mezclas pertenecientes al Cluster 1 y Cluster 2. Este resultado indica que las cepas que pertenecen al Cluster 1 y Cluster 2 mostraron valores más altos de los parámetros de productividad en comparación con las otras cepas de *Pleurotus ostreatus* (Cluster 3). Por otro lado, en el gráfico (b): el tamaño del Grupo 1 (color rojo) es 34, el tamaño del Grupo 2 (color rojo) es 27 y el tamaño del Grupo 3 (color negro) es 39. Las cepas de *Pleurotus djamor* producidas en las dos mezclas pertenecientes al Cluster 3 no mostraron una relación con las cepas de *Pleurotus djamor* cultivadas en los dos sustratos pertenecientes al Cluster 1 y Cluster 2. Este resultado indica que las cepas pertenecientes al Cluster 1 y Cluster 2 presentaron valores más altos de parámetros de productividad en comparación con las cepas de *Pleurotus djamor* pertenecientes al Cluster 3. Dado que los puntos de datos están normalmente distribuidos, los clusters varían en tamaño con los puntos de datos máximos y los puntos de datos mínimos. La suplementación del sustrato en el cultivo de hongos se ha llevado a cabo con relativo éxito, con el objetivo de controlar plagas o incrementar el rendimiento de los cultivos (Cardoso et al., 2021). Los resultados de los parámetros de productividad obtenidos fueron influenciados por las diferentes cepas y las mezclas

utilizadas en la investigación.

(**a**)



(**b**)

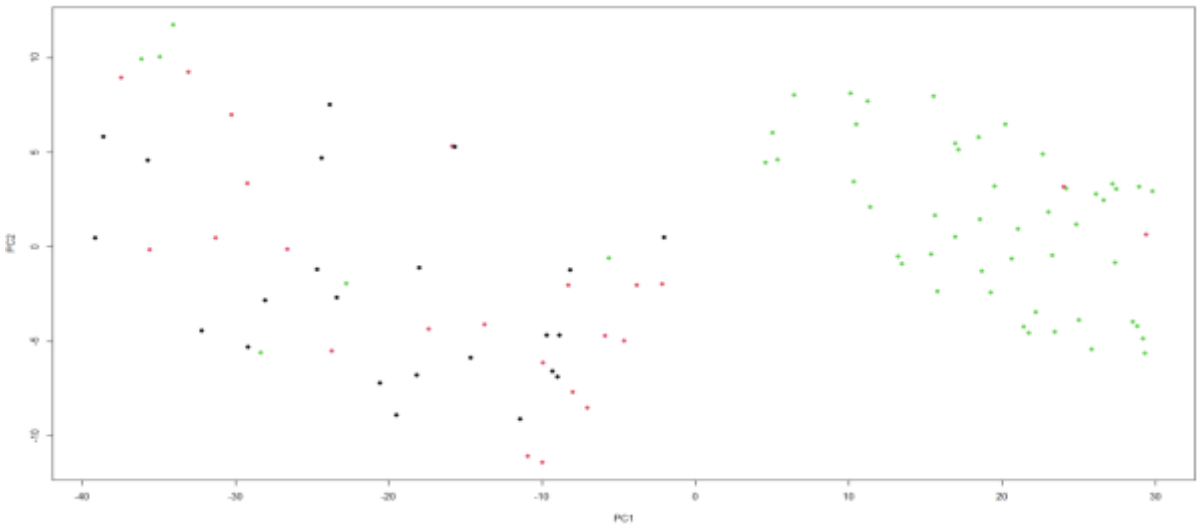


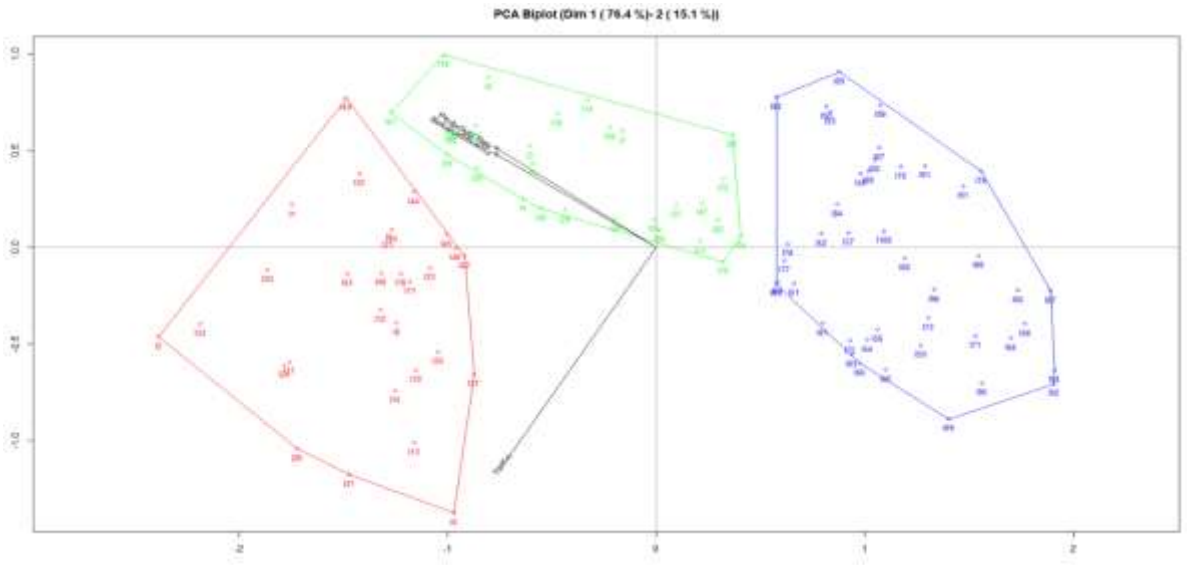
Figura 2. (a) K-means usando 3 clusters para los parámetros de productividad de hongos de Pleurotus ostreatus obtenidos en dos mezclas de desechos agrícolas, (b) K-means usando 3 grupos para los parámetros de productividad de hongos de Pleurotus djamor usando dos mezclas de desechos agrícolas.

La Figura 4 muestra el gráfico factorial del plano 1-2 (PCA Biplot). El gráfico (a) presenta la inercia acumulada asciende al 91,5%, mientras que el gráfico (b) presenta la inercia acumulada asciende a 93,0%. Además, los clusters se han calculado utilizando las coordenadas Biplot; la descripción general de los clusters se basa en tres variables. En el

gráfico (a), se observan diferencias importantes entre los clusters, el Cluster 2 (color verde) indica la presencia de 29 cepas de *Pleurotus ostreatus* cultivadas en las dos mezclas de residuos agrícolas con mayor relación con las eficiencias biológicas y las tasas de producción, mientras que el Cluster 1 (color rojo) indica la presencia de 28 cepas de *Pleurotus ostreatus* cultivadas sobre las dos mezclas de sustratos con mayor relación a los rendimientos, y el Cluster 3 (color azul) indica la presencia de 43 cepas de *Pleurotus ostreatus* cultivadas sobre las dos mezclas de residuos agrícolas. Por otro lado, en el gráfico (b) también existen diferencias entre los clusters, el Cluster 1 (color rojo) indica la presencia de 22 cepas de *Pleurotus djamor* crecimiento en las dos mezclas de residuos agrícolas con una mayor relación con las eficiencias biológicas y los rendimientos, mientras que el Cluster 2 (color verde) indica la presencia de 28 cepas de *Pleurotus djamor* cultivadas en las dos mezclas de sustratos con mayor relación con las tasas de producción, y el Cluster 3 (color azul) indica la presencia de 50 cepas de *Pleurotus djamor* en crecimiento en las dos mezclas de sustratos.

La producción comercial de hongos está determinada en gran medida por la disponibilidad y utilización de materiales baratos de desechos agrícolas que representan los sustratos ideales y más prometedores para el cultivo (Abrar et al., 2009; Da Silva et al., 2012). El aprovechamiento de estos residuos agrícolas de la provincia de Guayas se puede aprovechar para obtener la mayor productividad de cuerpos frutales proporcionando una alternativa para el mercado de hongos.

(**a**)



(**b**)

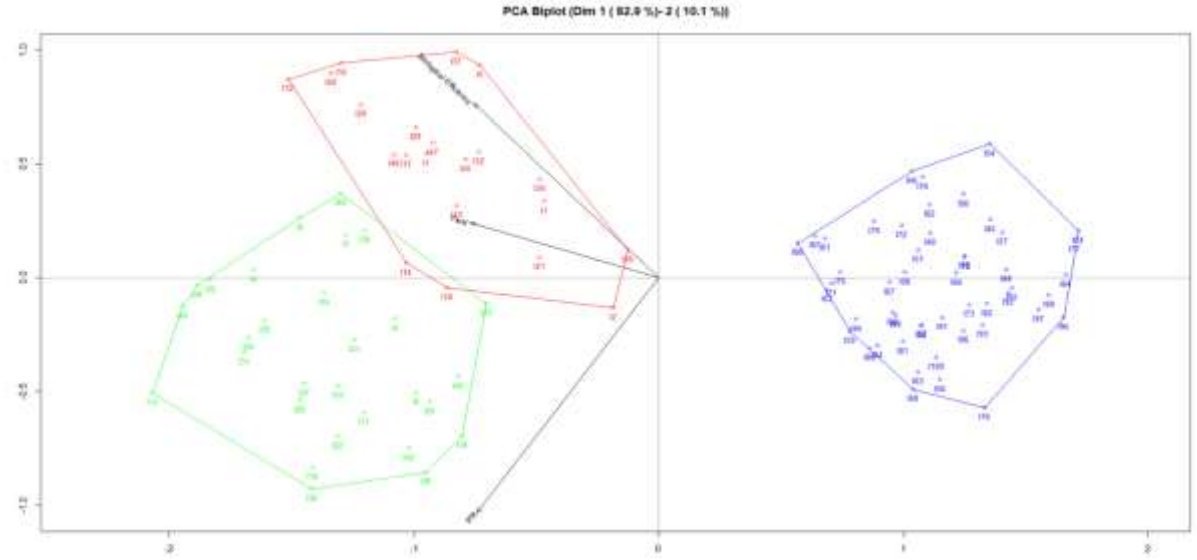
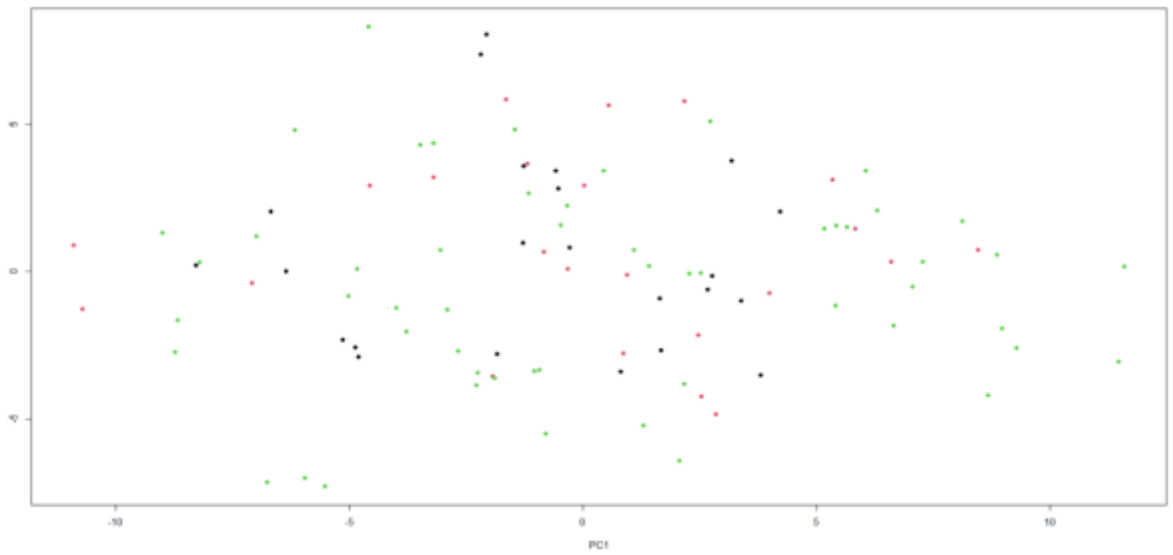


Figura 3. (a) PCA Biplot para parámetros de productividad de Pleurotus ostreatus, (b) PCA Biplot para parámetros de productividad de Pleurotus djamor.

3.2.2. Composición nutricional y propiedades biológicas

La Figura 5 presenta el uso del algoritmo de agrupamiento del método K-means para 100 objetos que tienen siete variables, cada una usando el software RStudio. El gráfico (a) muestra la aplicación de tres clusters para la composición nutricional y propiedades biológicas de los cuerpos fructíferos de *Pleurotus ostreatus* producidos sobre residuos agrícolas de la provincia del Guayas, mientras que en el gráfico (b) el uso de tres clusters para la composición nutricional muestra las propiedades biológicas de los hongos *Pleurotus djamor* cultivados en las dos mezclas de sustratos. El tamaño de cada grupo está relacionado con el número de puntos de datos, en el gráfico (a): el tamaño del Grupo 1 (color rojo) es 23, el tamaño del Grupo 2 (color negro) es 23 y el tamaño del Grupo 3 (color verde) es 54. Los tres clusters presentan los cuerpos fructíferos de *Pleurotus ostreatus* con los valores más altos de composición nutricional y propiedades biológicas. Por otro lado, en el gráfico (b): el tamaño del Grupo 1 (color rojo) es 23, el tamaño del Grupo 2 (color negro) es 23 y el tamaño del Grupo 3 (color verde) es 54. Los hongos de *Pleurotus djamor* con los valores más altos de composición nutricional y propiedades biológicas se muestran en los tres grupos. Es importante indicar que los clusters varían en tamaño con puntos de datos máximos y puntos de datos mínimos.

(**a**)



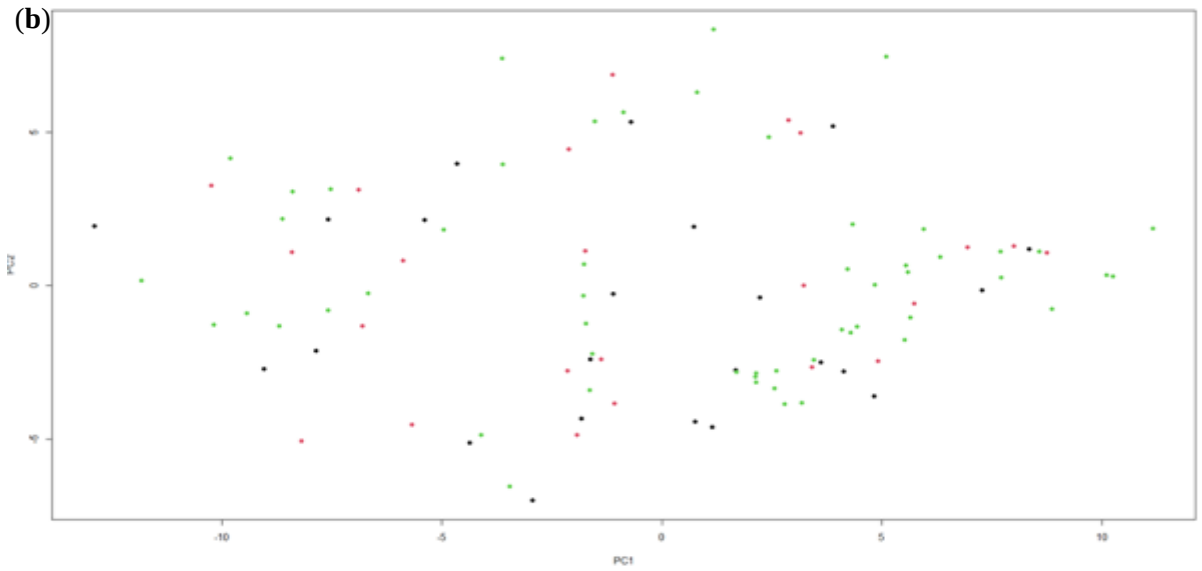


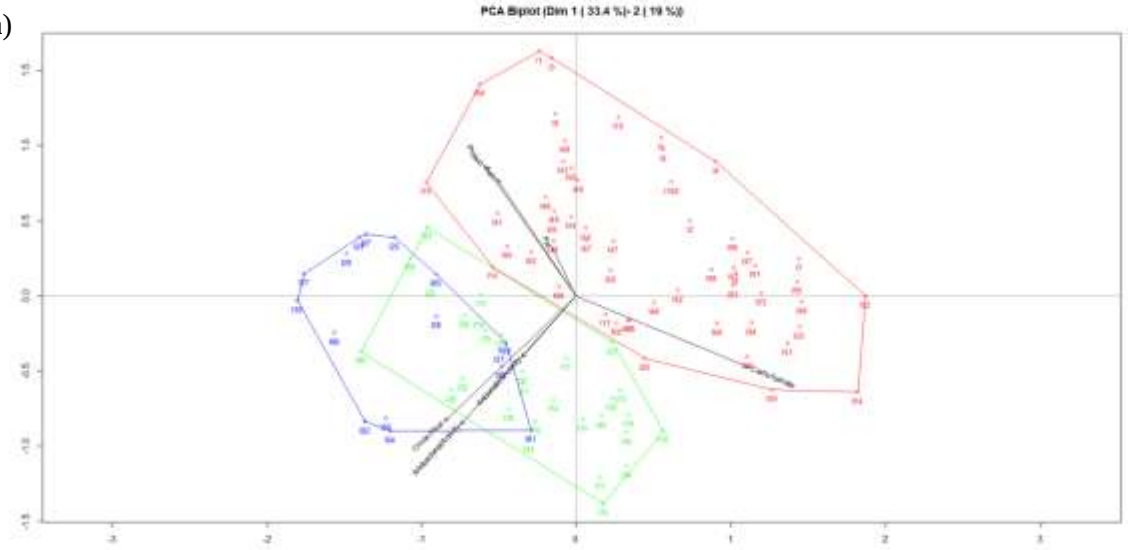
Figura 4. (a) K-means usando 3 grupos para las propiedades biológicas de Pleurotus ostreatus cultivado en dos mezclas de desechos agrícolas, (b) K-means usando 3 grupos para las propiedades biológicas de Pleurotus djamor cultivado en dos mezclas de desechos agrícolas.

La Figura 6 muestra el PCA Biplot del plano 1-2, el gráfico (a) indica que la inercia acumulada asciende al 52,4%, mientras que el gráfico (b) presenta la inercia acumulada asciende al 62,6%. Los tres clusters se han calculado utilizando las coordenadas Biplot, la descripción general de los clusters se basa en siete variables. El gráfico (a) muestra diferencias importantes entre los clusters, el Cluster 1 (color azul) indica la presencia de cuerpos fructíferos de 16 cepas de *Pleurotus ostreatus* cultivadas en las dos mezclas de desechos agrícolas con una mayor relación con el contenido de fibra cruda y las actividades antibacterianas , mientras que el Grupo 2 (color verde) indica la presencia de hongos de 28 cepas de *Pleurotus ostreatus* cultivadas en las dos mezclas de desechos alimenticios con una relación más alta con las actividades antioxidantes, y el Grupo 3 (color rojo) indica la presencia de cuerpos fructíferos de 56 cepas de *Pleurotus ostreatus* cultivadas en las dos mezclas de desechos agrícolas con mayor relación con el contenido de proteínas, cenizas, grasas y carbohidratos. Por otro lado, en el gráfico (b), también existen diferencias entre los clusters, Cluster 1 (color rojo) indica la presencia de 19 cepas de hongos de *Pleurotus djamor* cultivadas en las dos mezclas de residuos agrícolas con una mayor relación relación con el contenido de proteínas y las actividades antibacterianas, mientras que el Grupo 2 (color verde) indica la presencia de hongos de 34 cepas de *Pleurotus djamor* cultivadas en las dos mezclas de sustratos con una mayor

relación con el contenido de cenizas y grasas y también con actividades antioxidantes. , y el Grupo 3 (color azul) indica la presencia de hongos de 47 cepas de *Pleurotus djamor* crecidos en las dos mezclas de sustratos con una mayor relación con el contenido de carbohidratos y fibra cruda.

El contenido de humedad y grasa de los hongos está influenciado por la composición de los desechos agrícolas utilizados en el cultivo de hongos comestibles (Liu et al., 2017; Valencia del Toro et al., 2018). La composición nutricional de los hongos está influenciada por las cepas de los hongos comestibles y también por los residuos agrícolas utilizados en el cultivo, por lo que indicamos en base a los resultados que los residuos alimenticios de la provincia del Guayas se pueden utilizar para producir cuerpos fructíferos con las más altas propiedades biológicas.

(**a**)



(**b**)

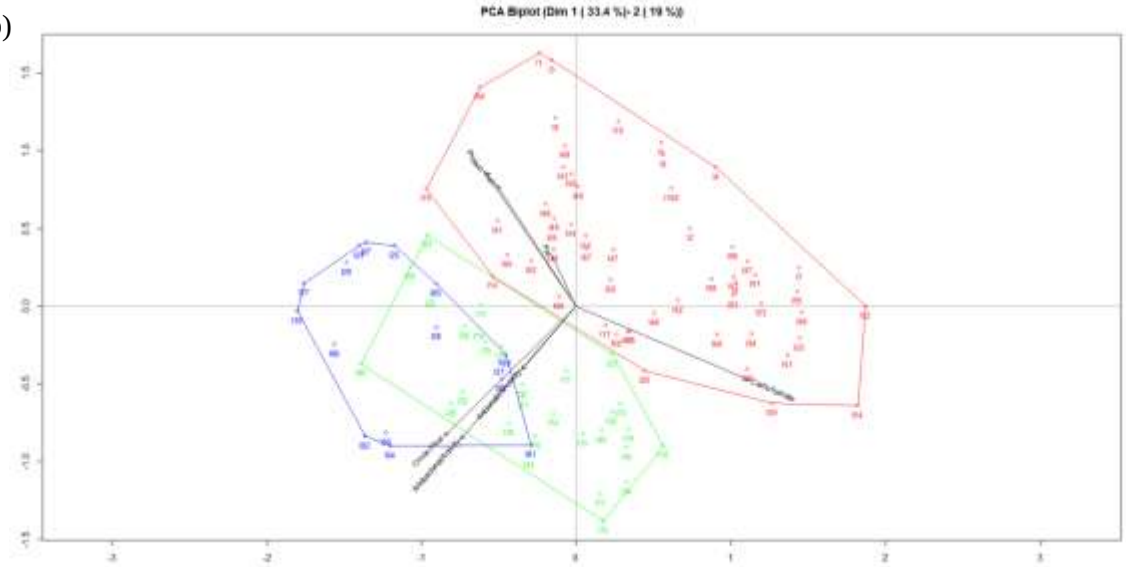


Figura 5. (a) Biplot de PCA para las propiedades biológicas de Pleurotus ostreatus, (b) Biplot de PCA para las propiedades biológicas de Pleurotus djamor.

Características de la Revista en que se publicó el artículo

Nombre de Revista: Journal of Fungi

Nivel de Cuartil: JCR – Q1

Factor de Impacto: 5.816

Journal of *Fungi*

MDPI

Article

Application of K-Means Clustering Algorithm to Commercial Parameters of *Pleurotus* spp. Cultivated on Representative Agricultural Wastes from Province of Guayas

Fabricio Guevara-Viejó [1,2], Juan Diego Valenzuela-Cobos [1,2], Purificación Vicente-Galindo [2] and Purificación Galindo-Villardón [2,*]

[1] Facultad de Ciencias e Ingeniería, Universidad Estatal de Milagro (UNEMI), 091050 Milagro, Ecuador; jguevarav@unemi.edu.ec (F.G.-V.); juan_diegova@hotmail.com (J.D.V.-C.)
[2] Department of Statistics, University of Salamanca, 37008 Salamanca, Spain; purivic@yahoo.com
* Correspondence: pgalindo@usal.es; Tel.: +34-646665034

Abstract: Data of the commercial parameters of *Pleurotus ostreatus* and *Pleurotus djamor* were analyzed using the data mining technique: K-means clustering algorithm. The parameters evaluated were: biological efficiency, crop yield ratio, productivity rate, nutritional composition, antioxidant and antimicrobial activities in the production of fruit bodies of 50 strains of *Pleurotus ostreatus* and 50 strains of *Pleurotus djamor*, cultivated on the most representative agricultural wastes from the province of Guayas: 80% sugarcane bagasse and 20% wheat straw (M1), and 60% wheat straw and 40% sugarcane bagasse (M2). The database of the parameters obtained in experimental procedures was grouped into three clusters, providing a visualization of the strains with a higher relation to each parameter (vector) measured.

Keywords: K-means clustering algorithm; *Pleurotus ostreatus*; *Pleurotus djamor*; commercial parameters; visualization

check for updates

Citation: Guevara-Viejó, F.; Valenzuela-Cobos, J.D.; Vicente-Galindo, P.; Galindo-Villardón, P. Application of K-Means Clustering Algorithm to Commercial Parameters of *Pleurotus* spp. Cultivated on Representative Agricultural Wastes from Province of Guayas. *J. Fungi* **2021**, *7*, 537. https://doi.org/10.3390/jof7070537

Academic Editors: Monika Gąsecka and Zuzanna Magdziak

Received: 1 June 2021
Accepted: 2 July 2021
Published: 4 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The province of Guayas, comprising 25 cantons, including the canton of Guayaquil, which consists of the Municipality of Guayaquil and two areas of urban expansion, along with five rural cantons: Morro, Juan Gómez Rendón, Posorja, Puna and Tenguel. The province of Guayas is part of Zone 5, according to SENPLADES, National Secretariat for Planning and Development, which subdivides the country into nine zones. The main land-use of the Guayas province is agriculture (27%), followed by livestock production and aquaculture. This province has gone through three main economic booms [1].

The first economic boom occurred in 1880 when it was the main cocoa exporting city (between 20% and 25% of the world), becoming an important commercial and financial center, which generated an increase in the city's population [2], the landowners monopolized the best land and access to transportation, as well as controlling key sources of credit and commercial links [3]. The second economic boom was in 1950 with the so-called "banana boom", and foreign banana-producing companies arrived, such as the United Fruit Company in Tenguel, one of the largest banana plantations (localized 100 miles south of Guayaquil) [4]. The third boom was the oil boom in 1972 that brought new developments, mainly in the form of land invasions on the outskirts of the city, causing an immense deterioration of the agricultural sector. The lack of a national policy for small-scale rural agriculture led many of the rural small farmers (mainly indigenous from the central areas) to abandon their plots and engage in non-agricultural activities, most often in the urban informal sector [5]. With the previously described background, it is important to continue using the agricultural wastes that continue to be generated from all of the provinces of Guayas, having an innovative idea, such as the production of edible mushrooms, also called "vegetable steak" [6].

The cultivation of edible mushrooms has gradually grown using homemade techniques until becoming a highly technical industry [7–10]. The world production of edible mushrooms has grown in the last three years, with an annual increase of 24.5%. The nutritional value of edible mushrooms is high in comparison with other kinds of food. According to studies carried out by food specialists, mushrooms have a protein content between 19% and 35%, compared to vegetables (vegetables and fruits) that only have protein between 7.3% from 13.2%; on the other hand, milk, meat and eggs have a protein content between 25% and 90%. However, at the amino acid level, protein precursor substances, such as lysine and tryptophan, reach levels between 1.1 and 2.09 g. On the other hand, the low carbohydrate content makes mushrooms a low-energy food and is recommended as a dietary one. In addition, the content of essential fatty acids such as oleic and linoleic is found in appreciable quantities [11]. Edible mushrooms are nutritious plants that contain riboflavin, nicotinic acid, pantothenate and biotin, which lower blood pressure, prevent atherosclerosis and boost the immune system (immune system) against disease [12].

One of the principal genera of edible mushrooms with the highest production around the world is *Pleurotus* spp. [13,14]. These mushrooms are characterized by their nutritional value and are an important source of proteins, vitamins and minerals [15,16]. These species require tropical or subtropical climates similar to the Province of Guayas for the cultivation and production of fruiting bodies [17–19]. Additionally, these mushrooms are actively used in medical treatments with antioxidant and antimicrobial properties protecting health by damping active oxygen and free radicals [20]. However, the lack of knowledge of the nutritional and pharmaceutical properties of *Pleurotus* spp. has not been allowed to be used for the benefit of human health in Ecuador. In order to illustrate adequate visualization, the use of data mining tools, such as the K-means clustering algorithm, upon big data, about the commercial parameters of *Pleurotus* spp. is important [21].

In this paper, the focus is on clustering. The clusters have been applied in many research areas such as mathematics, engineering, economics, marketing, machine learning, pattern recognition, genetics, bioinformatics, psychology, biology, data compression and information retrieval [22]. The initial values of cluster "centroids" are randomly selected from the available data. Updating centroids and clustering of data is then repeated until convergence is reached or for a defined number of iterations. A new centroid for a cluster is calculated based on each data sample that belongs to that cluster, and the initial centroids are usually chosen randomly for the application of K-means-type algorithms [23].

The main goal was to use a data mining technique, the K-means clustering algorithm, to verify the influence of two mixtures of agricultural wastes obtained from the province of Guayas on the viability of the mushrooms production, the nutritional profile and also in the antioxidant and antimicrobial properties. The use of the K-means clustering algorithm allowed the indication of the strain cultivated on a specific mixture of agricultural wastes that obtained the highest values in commercial parameters.

2. Materials and Methods

2.1. Mushroom Strains

In this study, 50 strains of *Pleurotus ostreatus* (PO) and 50 strains of *Pleurotus djamor* (PD) were used. These strains were collected from producers in the province of Guayas. The strains were maintained on MEA dishes and are deposited at the fungal collection of the Research and Development Laboratory of Ecuahidrolizados.

2.2. Substrate and Supplementation

Strains were cultivated using two mixtures of agricultural wastes: 80% sugarcane bagasse and 20% wheat straw (M1), and 60% wheat straw and 40% sugarcane bagasse (M2). The mixtures of agricultural wastes were moistened for 1 day. Subsequently, the mixture was placed (1 kg wet weight) in plastic bags and pasteurized for 10 h at 80 °C.

After pasteurization and conditioning, with the substrate at ambient temperature, the bags with the substrate were inoculated with 150 g of wheat grain previously colonized

with the strains of *Pleurotus ostreatus* (PO) and the strains of *Pleurotus djamor* (PD). Thereafter, the bags with the substrate, were incubated in a dark room at a temperature of 30 ± 1 °C.

Finally, once the mycelium of the strain had colonized the substrate, the bags with the substrate were transferred to a room with favorable conditions for the fructification: relative humidity was maintained between 85% and 90%, a temperature of 25 ± 1 °C, air recirculation and period of illumination of 12 h [24].

2.3. Productivity Parameters

2.3.1. Biological Efficiency

The biological efficiency (BE) is a productivity parameter that explains the capacity of the substrate to produce fruit bodies and was calculated using the following equation [25]:

$$BE(\%) = \frac{fresh\ weight\ of\ mushrooms\ (g)}{weight\ of\ dry\ substrate\ (g)} \times 100 \quad (1)$$

Equation (1). Biological efficiency of the mushrooms.

2.3.2. Yield Ratio

The yield ratio is an analyzed variable widely adopted for industrial crops and was calculated with the following equation [26]:

$$Y(\%) = \frac{fresh\ weight\ of\ mushrooms\ (g)}{fresh\ weight\ of\ substrate\ (g)} \times 100 \quad (2)$$

Equation (2). Yield ratio of the mushrooms.

2.3.3. Productivity Rate (PR)

The productivity rate is the relation between BE and the precocity (namely days between inoculation and harvest) and was calculated using the following equation [27]:

$$PR(\%\ per\ day) = \frac{biological\ efficiency\ (\%)}{precocity\ (days)} \quad (3)$$

Equation (3). Productivity rate of the mushrooms.

2.4. Nutritional Composition

The nutritional value of the mushroom sample was analyzed using AOAC procedures concerning the composition of proteins, fat, carbohydrates and ash [28]. For the estimation of the crude protein content (N × 4.38), the macro-Kjeldahl method was used; the crude fat content was determined by extracting a known weight of sample with hexano, using a Soxhlet apparatus while the ash content was determined by calcination at 600 °C [29]. The total carbohydrate content (%C) was calculated by using the following equation:

$$C(\%) = 100 - (\%moisture + \%protein + \%fat + \%ash\ \ contents) \quad (4)$$

Equation (4). Percentage of carbohydrates of the mushrooms.

2.5. Antioxidant Activity

To evaluate the antioxidant activity, the DPPH radical-scavenging assay was used. In the beginning, 30 µL of the extract and 270 µL of methanol containing DPPH radicals (6×10^{-5} mol L^{-1}) were pipetted and mixed in a 96 well plate. The reaction mixture was incubated in the dark for 30 min, and the absorption was measured at 515 nm using a microplate reader [30]. The DPPH radical scavenging activity (RSA) was calculated as a percentage of DPPH discoloration using the following equation:

$$RSA(\%) = \frac{ADPPH - AS}{ADPPH} \times 100 \quad (5)$$

Equation (5). DPPH radical scavenging activity (RSA).

2.6. Antimicrobial Activity

The antimicrobial activity was analyzed using the following Gram-negative bacteria: *Pseudomonas aeruginosa* (ABN 187) and *Salmonella typhimurium* (ABN 572); and the following Gram-positive bacteria: *Micrococcus flavus* (ABP 147) and *Staphylococcus aureus* (ABP 784). The microorganisms are deposited at the Research and Development Laboratory of Ecuahidrolizados.

Bacterial suspensions were adjusted with sterile saline to a concentration of 1.0×10^6 CFU/mL. The mushroom extracts of *Pleurotus* spp. were dissolved in 30% ethanol, mixed with nutrient media for bacteria (Tryptic Soy Broth) containing bacterial inoculum (1.0×10^5 CFU per well) with a final volume of 100 µL [31].

2.7. Statistical Analysis

K-Means Clustering

The K-means grouping method is a non-hierarchical technique used to group observations into K groups. Each item is assigned to a group with the closest center. The algorithm iteratively updates the groups to minimize the variation of their elements. The basic K-means algorithm, which was used in this article, refers to the Euclidean metric to define the distance between the elements and the centers of the clusters [32]. The Euclidean distance is selected as the similarity index, and the clustering targets minimize the sum of the squares of the various types; that is, it minimizes [33]:

$$d = \sum_{k=1}^{k} \sum_{i=1}^{k} \| \left(x_i - u_k\right) \|^2$$

where k represents K cluster centers, u_k represents the kth center, and x_i represents the ith point in the data set. The solution to the centroid u_k is as follows:

$$\frac{\partial}{\partial u_k} = \frac{\partial}{\partial u_k} \sum_{k=1}^{k} \sum_{i=1}^{n} (x_i - u_k)^2$$

$$= \sum_{k=1}^{k} \sum_{i=1}^{n} \frac{\partial}{\partial u_k} (x_i - u_k)^2$$

$$= \sum_{i=1}^{n} 2(x_i - u_k)$$

$$u_k = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Furthermore, the result of the K-means method is highly dependent on the number of clusters defined beforehand. In general, the iterative clustering method of K-means is implemented as follows: Step 1: A value of K is chosen. It is used as the initial set of K centroids. Step 2: Each of the objects is assigned to the group with the closest centroid. Step 3: The new centroids of the K groups are determined, calculating the mean of the group members. Step 4: Steps 3 and 4 are repeated until there are no changes in the criterion function after one iteration [34].

The main advantages of the K-means algorithm are its low complexity, it is computationally fast, the ability to handle large data sets and the flexibility to adjust the cluster number. K-means clustering was used to extract clusters from the dataset that had been optimized by feature selection.

Additionally, a PCA biplot [35] was applied to explore and visualize the different parameters and the most relevant responses.

3. Results and Discussion

The focus of this work was to determine the viability of the use of agricultural wastes from the province of Guayas on the cultivation of the strains *Pleurotus ostreatus* and *Pleurotus djamor* and assess its influence on commercial parameters: biological efficiency, crop yield ratio, productivity rate, nutritional composition, antioxidant and antimicrobial activities.

The numeration of the strains cultivated on the two mixtures of agriculture was made using the following distribution:

1–50: Strains of *Pleurotus ostreatus* or *Pleurotus djamor* cultivated on the mixture M1.

51–100: Strains of *Pleurotus ostreatus* or *Pleurotus djamor* cultivated on the mixture M2.

3.1. Productivity Parameters

Figure 1 shows the application of the K-means clustering algorithm method to 100 objects having three variables, with each one using the software RStudio. The graphic (a) presents the use of three clusters for the productivity parameters of *Pleurotus ostreatus* strains cultivated on agricultural wastes from the province of Guayas, while in the graphic (b) the use of three clusters for the productivity parameters of *Pleurotus djamor* strains grown on the two mixtures of substrates was shown. The results show the normal distribution of 100 data points around three clusters in each graphic. The size of each cluster is related to the number of data points, in graphic (a): the size of Cluster 1 (color red) is 34, the size of Cluster 2 (color black) is 27, and the size of Cluster 3 (color green) is 39. *Pleurotus ostreatus* strains grown on the two mixtures belonging to Cluster 3 did not show a relationship with the *Pleurotus ostreatus* strains grown on the two mixtures belonging to Cluster 1 and Cluster 2. This result indicates that the strains that belong to Cluster 1 and Cluster 2 showed higher values of the productivity parameters in comparison to the other *Pleurotus ostreatus* strains (Cluster 3). On the other hand, in graphic (b): the size of Cluster 1 (color red) is 34, the size of Cluster 2 (color red) is 27, and the size of Cluster 3 (color black) is 39. *Pleurotus djamor* strains produced on the two mixtures belonging to Cluster 3 did not show a relationship with the *Pleurotus djamor* strains cultivated on the two substrates belonging to Cluster 1 and Cluster 2. This result indicates that the strains belonging to Cluster 1 and Cluster 2 presented higher values of productivity parameters in comparison to *Pleurotus djamor* strains belonging to Cluster 3. Since the data points are normally distributed, the clusters vary in size with the maximum data points and minimum data points. The supplementation of the substrate on mushroom cultivation has been carried out with relative success, aiming at controlling pests or increasing crop yields [36]. The results of the productivity parameters obtained were influenced by the different strains and the mixtures used in the research.

Figure 2 shows the factorial graph of the plane 1–2 (PCA Biplot). Graphic (a) presents the accumulated inertia amounts to 91.5%, while graphic (b) presents the accumulated inertia amounts to 93.0. In addition, clusters have been calculated using the Biplot coordinates; the overview of clusters is based on three variables. In graphic (a), we observe important differences between clusters, Cluster 2 (color green) indicates the presence of 29 strains of *Pleurotus ostreatus* cultivated on the two mixtures of agricultural wastes with a higher relation to biological efficiencies and production rates, while Cluster 1 (color red) indicates the presence of 28 strains of *Pleurotus ostreatus* cultivated on the two mixtures of substrates with a higher relation to the yields, and Cluster 3 (color blue) indicates the presence of 43 strains of *Pleurotus ostreatus* cultivated on the two mixtures of agricultural wastes. On the other hand, in graphic (b) also there are differences between the clusters, Cluster 1 (color red) indicates the presence of 22 strains of *Pleurotus djamor* growth on the two mixtures of agricultural wastes with a higher relation to biological efficiencies and yields, whereas Cluster 2 (color green) indicates the presence of 28 strains of *Pleurotus djamor* cultivated on the two mixtures of substrates with a higher relation to production rates, and Cluster 3 (color blue) indicates the presence of 50 strains of *Pleurotus djamor* growth on the two mixtures of substrates.

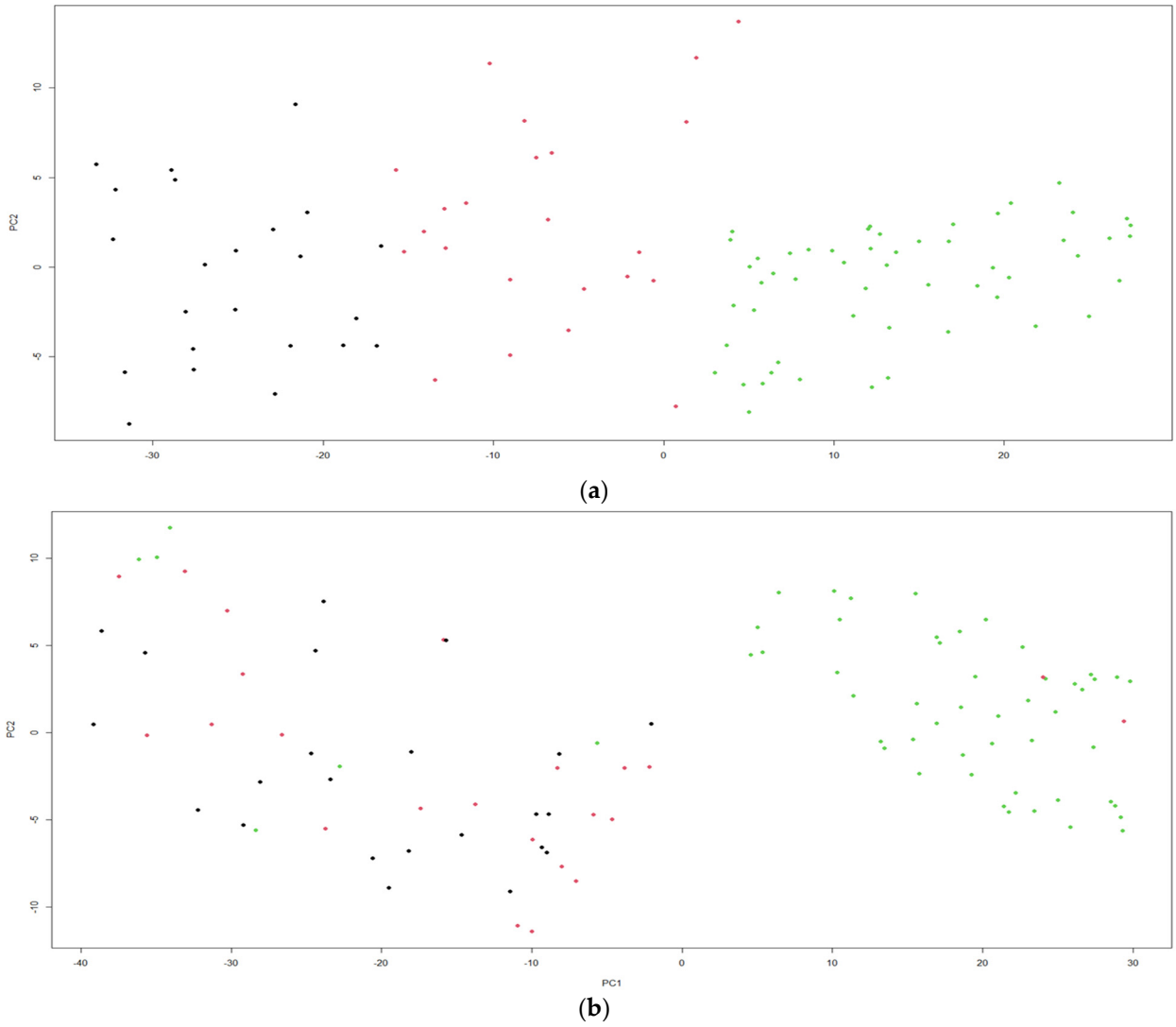


Figure 1. (**a**) K-means using 3 clusters for productivity parameters of *Pleurotus ostreatus* cultivated on two mixtures of agricultural wastes, (**b**) K-means using 3 clusters for productivity parameters of *Pleurotus djamor* cultivated on two mixtures of agricultural wastes.

The commercial production of mushrooms is largely determined by the availability and utilization of cheap materials of agricultural wastes that represent the ideal and most promising substrates for cultivation [37,38]. The use of these agricultural wastes from the province of Guayas can be used to obtain the highest productivity of fruit bodies providing an alternative for the mushroom market.

3.2. Nutritional Composition and Biological Properties

Figure 3 presents the use of method K-means clustering algorithm to 100 objects having seven variables, each one using the software RStudio. Graphic (a) shows the application of three clusters for the nutritional composition and biological properties of *Pleurotus ostreatus* fruit bodies produced on agricultural wastes from the province of Guayas, while in graphic (b) the use of three clusters for the nutritional composition and biological properties of *Pleurotus djamor* mushrooms cultivated on the two mixtures of substrates is shown. A normal distribution of 100 data points around three clusters in each graphic was presented. The size of each cluster is related to the number of data points, in graphic (a): the size of Cluster 1 (color red) is 23, the size of Cluster 2 (color black) is 23, and the size of Cluster 3 (color green) is 54. The three clusters present the *Pleurotus ostreatus* fruit bodies with the highest values of nutritional composition and biological properties. On the other hand, in

graphic (b): the size of Cluster 1 (color red) is 23, the size of Cluster 2 (color black) is 23, and the size of Cluster 3 (color green) is 54. *Pleurotus djamor* mushrooms with the highest values of nutritional composition and biological properties are shown by the three clusters. It is important to indicate that the clusters vary in size with maximum data points and minimum data points.

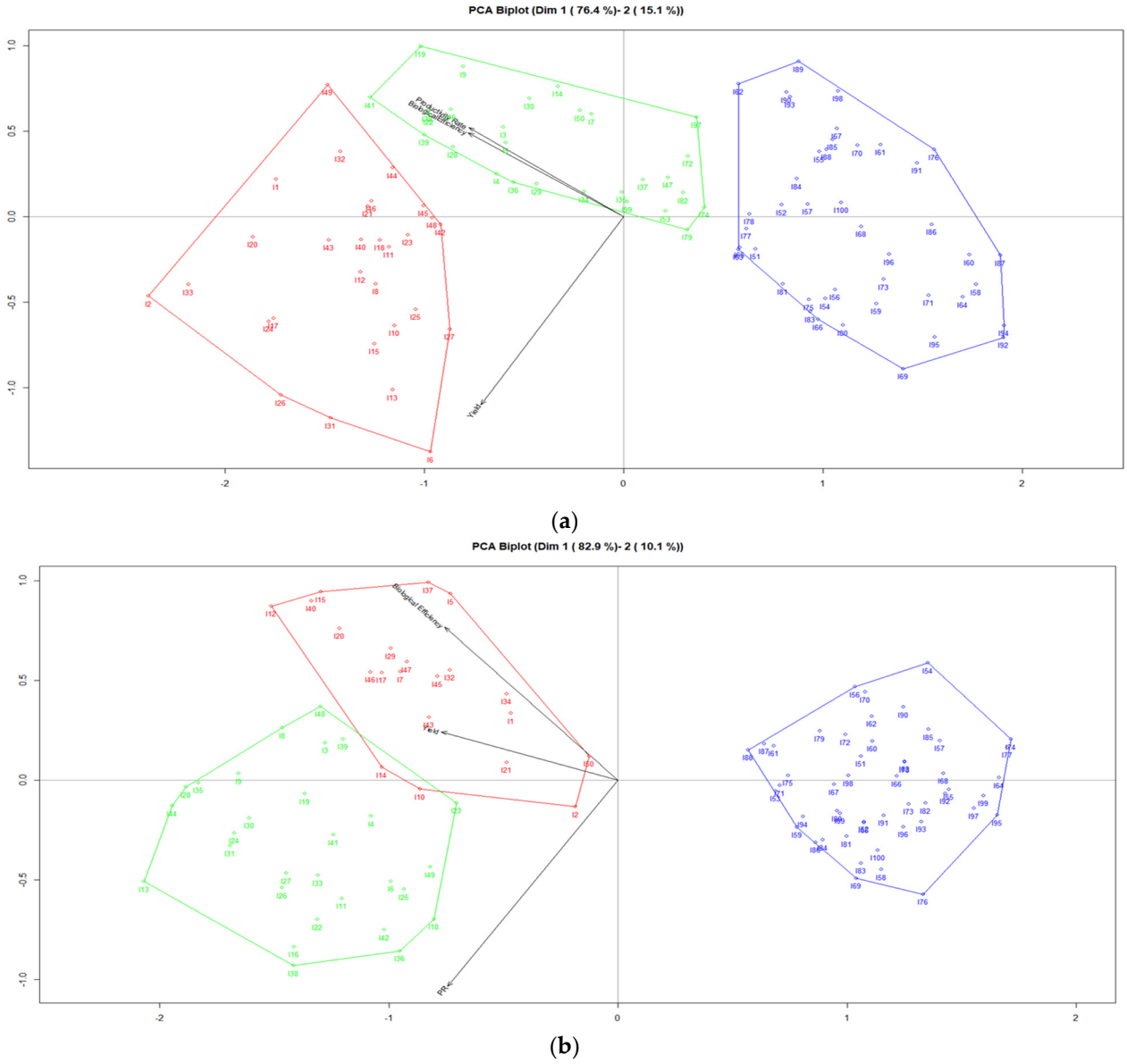
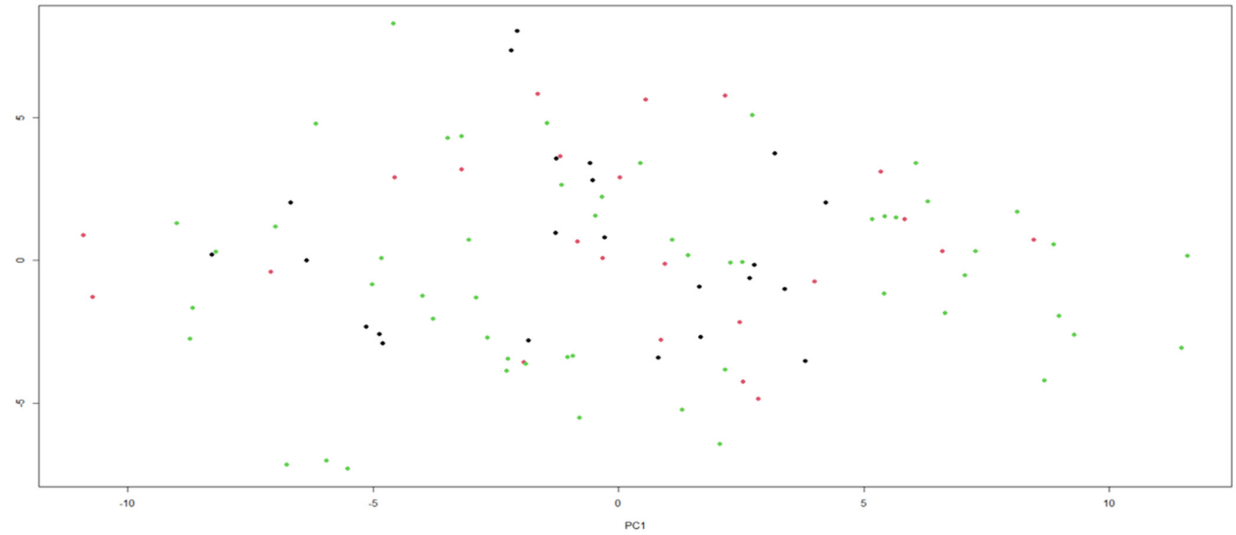


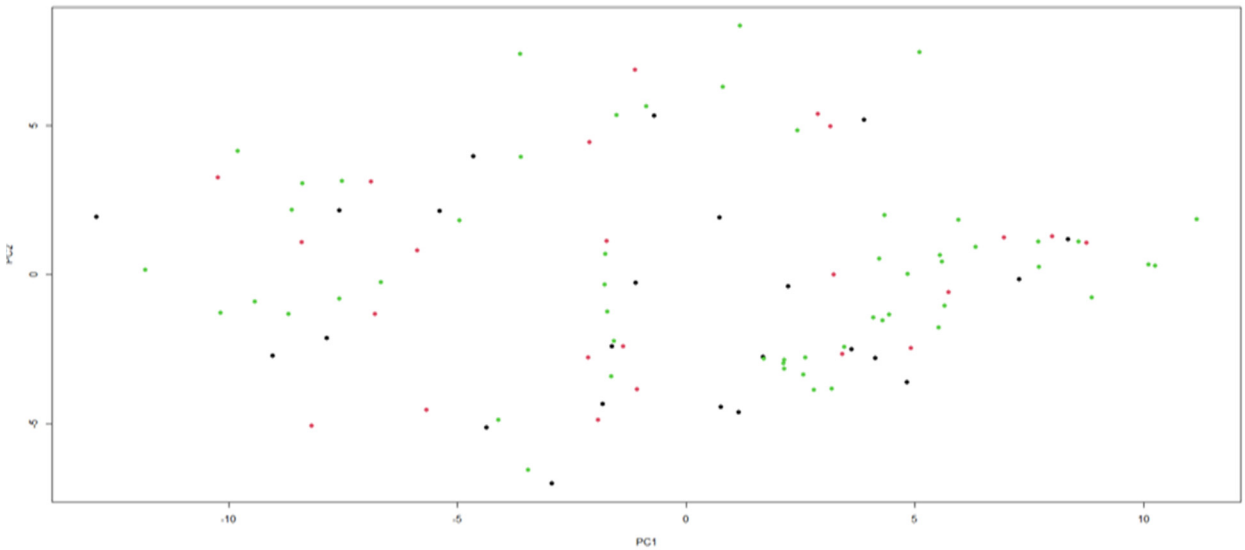
Figure 2. (**a**) PCA Biplot for productivity parameters of *Pleurotus ostreatus*, (**b**) PCA Biplot for productivity parameters of *Pleurotus djamor*.

Figure 4 shows the PCA Biplot of the plane 1–2, graphic (a) indicates the accumulated inertia amounts to 52.4%, while graphic (b) presents the accumulated inertia amounts to 62.6%. The three clusters have been calculated using the Biplot coordinates, the overview of clusters is based on seven variables. Graphic (a) shows important differences between clusters, Cluster 1 (color blue) indicates the presence of fruit bodies of 16 strains of *Pleurotus ostreatus* cultivated on the two mixtures of agricultural wastes with a higher relation to the crude fiber contents and antibacterial activities, while Cluster 2 (color green) indicates the presence of mushrooms of 28 strains of *Pleurotus ostreatus* cultivated on the two mixtures of food wastes with a higher relation to the antioxidant activities, and Cluster 3 (color red) indicates the presence of fruit bodies of 56 strains of *Pleurotus ostreatus* culti-

vated on the two mixtures of agricultural wastes with a higher relation to the protein, ash, fat and carbohydrate contents. On the other hand, in graphic (b), there are also differences between the clusters, Cluster 1 (color red) indicates the presence of 19 strains of mushrooms of *Pleurotus djamor* growth on the two mixtures of agricultural wastes with a higher relation to protein contents and antibacterial activities, whereas Cluster 2 (color green) indicates the presence of fruit bodies of 34 strains of *Pleurotus djamor* cultivated on the two mixtures of substrates with a higher relation to ash and fat contents and also antioxidant activities, and Cluster 3 (color blue) indicates the presence of mushrooms of 47 strains of *Pleurotus djamor* growth on the two mixtures of substrates with a higher relation to carbohydrate and crude fiber contents.



(**a**)



(**b**)

Figure 3. (**a**) K-means using 3 clusters for biological properties of *Pleurotus ostreatus* cultivated on two mixtures of agricultural wastes, (**b**) K-means using 3 clusters for biological properties of *Pleurotus djamor* cultivated on two mixtures of agricultural wastes.

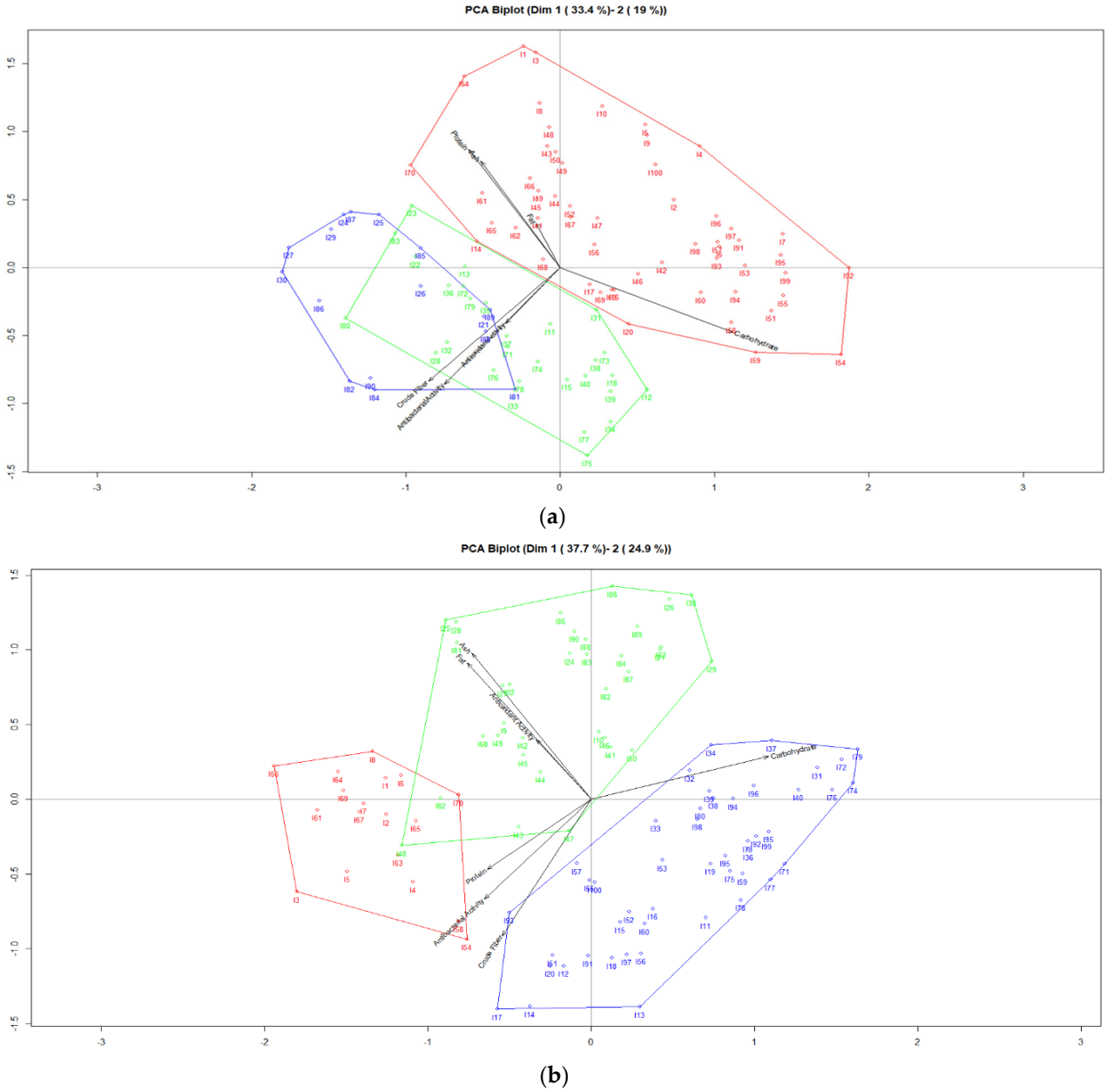


Figure 4. (**a**) PCA Biplot for biological properties of *Pleurotus ostreatus*, (**b**) PCA Biplot for biological properties of *Pleurotus djamor*.

The moisture and fat contents of the mushrooms are influenced by the composition of the agricultural wastes used in the cultivation of edible fungi [39,40]. The nutritional composition of the mushrooms is influenced by the strains of the edible fungi and also by the agricultural wastes used in the cultivation, so we indicate based on the results that the food wastes from the province of Guayas can be used to produce fruit bodies with the highest biological properties.

4. Conclusions

The K-means clustering algorithm was used to obtain proper grouping data using three clusters and providing visualization about the relationships between strains of edible fungi *Pleurotus ostreatus* and *Pleurotus djamor* cultivated on the most representative agricultural wastes from the province of Guayas, with the commercial parameters measured in experimental procedures.

PCA Biplots presented that the use of mixture 1 in the cultivation of the strains of edible fungi *Pleurotus ostreatus* and *Pleurotus djamor* has a higher relation to the productivity parameters: biological efficiencies, crop yields and productivity rates.

The use of the K-means clustering algorithm on the commercial parameters of edible fungi *Pleurotus ostreatus* and *Pleurotus djamor*, cultivated on two mixtures of agricultural wastes, allowed the indication of how to obtain the highest values in productivity parameters or biological properties due to the strain grown on a specific substrate.

Author Contributions: Conceptualization, F.G.-V. and J.D.V.-C.; Formal analysis, J.D.V.-C.; Investigation, F.G.-V.; Methodology, F.G.-V. and J.D.V.-C.; Supervision, P.G.-V. and P.V.-G.; Writing—original draft, F.G.-V., J.D.V.-C., P.G.-V., and P.V.-G.; Writing—review and editing, P.G.-V. and P.V.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by Universidad Estatal de Milagro (UNEMI) Scholarship.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are grateful to Facultad de Ciencias e Ingeniería de la Universidad Estatal de Milagro (UNEMI) and Ecuahidrolizados Industry.

Conflicts of Interest: The authors state no conflict of interest.

References

1. Delgado, A. Guayaquil. *Cities* **2013**, *31*, 515–532. [CrossRef]
2. Mora, E. *Auge y Crisis de una Economía Agroexportadora: El Período Cacaotero*; Corporación Editora Nacional-Editorial Grijalbo Ecuatoriana: Quito, Ecuador, 1988.
3. Pineo, R. Guayaquil and coastal Ecuador during the cacao era. In *The Ecuador Reader*; De La Torre, C., Striffler, S., Eds.; Duke University Press: Durham, NC, USA; London, UK, 2008; pp. 136–147.
4. Striffer, S. The united fruit company's legacy in Ecuador. In *The Ecuador Reader*; De La Torre, C., Striffler, S., Eds.; Duke University Press: Durham, NC, USA; London, UK, 2008; pp. 239–249.
5. Swanson, K. Revanchist urbanism heads south: The regulation of indigenous beggars and street vendors in Ecuador. *Antipode* **2007**, *39*, 708–728. [CrossRef]
6. Crisan, E.V.; Sands, A. *Nutritional Value*; Academic Press: New York, NY, USA, 1978; pp. 137–168.
7. Jong, S.C.; Peng, J.T. Identity and cultivation of a new commercial mushroom in Taiwan. *Mycologia* **1975**, *67*, 1235–1240. [CrossRef] [PubMed]
8. Farr, D.F. Mushroom industry: Diversification with additional species in the United States. *Mycologia* **1983**, *75*, 351–360. [CrossRef]
9. Kaul, T.N. *Cuftivated Edible Mushrooms*; Regional Research Laboratory: Jammu, India, 1983.
10. Kaul, T.N.; Kapur, Y.B.M. (Eds.) *Indian Mushroom Science, 11. Procceding Intern. Conference on Science and Cultivation Technol. Edible Fungi*; Regional Research Laboratory: Jammu, India, 1987.
11. Chang, S.T.; Miles, P.G. *Mushrooms: Cultivation, Nutritional Value, Medicinal Effect, and Environmental Impact*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2008.
12. Singh, R.S.; Bhari, R.; Kaur, H.P. Mushroom lectins: Current status and future perspectives. *Crit. Rev. Biotechnol.* **2010**, *30*, 99–126. [CrossRef] [PubMed]
13. Valenzuela-Cobos, J.D.; Páramo, E.D.; Arce, R.V.; Sánchez-Hernández, A.; Aguilar, M.E.G.; Lara, H.L.; Valencia del Toro, G. Production of hybrid strains among *Pleutorus* and *Lentinula* and evaluation of their mycelial growth kinetics on malt extract agar and wheat grain using the Gompertz and Hill models. *Emir. J. Food Agric.* **2017**, *29*, 927–935.
14. Sánchez-Hernández, A.; Valenzuela Cobos, J.D.; Herrera Martínez, J.; Arce, R.V.; Gómez y Gómez, Y.M.; Segura, P.B.Z.; Aguilar, M.E.G.; Lara, H.L.; Valencia del Toro, G. Characterization of *Pleurotus djamor* neohaplonts recovered by production of protoplasts and chemical dedikaryotization. *3 Biotech* **2019**, *9*, 24. [CrossRef] [PubMed]
15. Manzi, P.; Aguzzi, A.; Pizzoferrato, L. Nutritional value of mushrooms widely consumed in Italy. *Food Chem.* **2001**, *73*, 321–325. [CrossRef]
16. Reis, F.S.; Barros, L.; Martins, A.; Ferreira, I. Chemical composition and nutritional value of the most widely appreciated cultivated mushrooms: An inter-species comparative study. *Food Chem. Toxicol.* **2012**, *50*, 191–197. [CrossRef]
17. Mori, K.; Fukai, S.; Zennyoji, A. Hybridization of shiitake (*Lentinus edodes*) between cultivated strains of Japan and wild strains grown in Taiwan and New Guinea. *Mushroom. Sci.* **1974**, *9*, 391–403.
18. Fultz, S.A. Fruiting at high temperature and its genetic control in the basidiomycete *Flammulina velutipes*. *Appl. Environ. Microbiol.* **1988**, *54*, 2460–2463. [CrossRef]

19. Kashangura, C.; Hallsworth, J.E.; Mswaka, A.Y. Phenotypic diversity amongst strains of *Pleurotus sajor-caju*: Implications for cultivation in arid environments. *Mycol. Res.* **2006**, *110*, 312–317. [CrossRef]
20. Bakir, T.; Karadeniz, M.; Unal, S. Investigation of antioxidant activities of *Pleurotus ostreatus* stored at different temperatures. *Food Sci. Nutr.* **2018**, *6*, 1040–1044. [CrossRef]
21. Guler, C.; Thyne, G.D.; McCray, J.E.; Turner, A.K. Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. *Hydrogeol. J.* **2002**, *10*, 455–474. [CrossRef]
22. Hot, E.; Popović-Bugarin, V. Soil data clustering by using K-means and fuzzy K-means algorithm. *Telfor. J.* **2016**, *8*, 51–56. [CrossRef]
23. Bezdek, J.C.; Ehrilich, R.; Fill, W. FCM: The Fuzzy C-means Clustering Algorithm. *Comput. Geosci.* **1984**, *10*, 191–203. [CrossRef]
24. Valenzuela-Cobos, J.D.; Rodríguez-Grimón, R.O.; Jara-Bastidas, M.L.; Grijalva-Endara, A.; Zied, D.C.; Garín-Aguilar, M.E.; Valencia del Toro, G. Modelling of mycelial growth of parental, hybrid and reconstituted strains of *Pleurotus* and *Lentinula*. *Rev. Mex. Ing. Quim.* **2020**, *19*, 165–174. [CrossRef]
25. Thongsook, T.; Kongbangkerd, T. Influence of calcium and silicon supplementation into *Pleurotus ostreatus* substrates on quality of fresh and canned mushrooms. *Food Sci. Technol. Int.* **2011**, *17*, 351–365. [CrossRef]
26. Salmones, D.; Gaitán-Hernández, R.; Pérez, R.; Guzmán, G. Estudios sobre el género Pleurotus. VIII. Interacción entre crecimiento micelial y productividad. *Rev. Iberoam. Micol.* **1997**, *14*, 173–176.
27. Cardoso, R.V.C.; Carocho, M.; Fernandes, Â.; Zied, D.C.; Cobos, J.D.V.; González-Paramás, A.M.; Ferreira, I.C.F.R.; Barros, L. Influence of Calcium Silicate on the Chemical Properties of *Pleurotus ostreatus* var. florida (Jacq.) P. Kumm. *J. Fungi* **2020**, *6*, 299. [CrossRef] [PubMed]
28. AOAC. *Official Methods of Analysis of AOAC International*, 20th ed.; AOAC: Rockville, MD, USA, 2016.
29. Mocan, A.; Fernandes, A.; Barros, L.; Crişan, G.; Smiljković, M.; Soković, M.; Ferreira, I.C.F. Chemical composition and bioactive properties of the wild mushroom *Polyporus squamosus* (Huds.) Fr: A study with samples from Romania. *Food Sci. Nutr.* **2018**, *9*, 160–170. [CrossRef]
30. Kostic, M.; Smiljkovic, M.; Petrovic, J.; Glamocilija, J.; Barros, L.; Ferreira, I.C.F.R.; Ciric, A.; Sokovic, M. Chemical, nutritive composition and wide-broad bioactive properties of honey mushroom *Armillaria mellea* (Vahl: Fr.) Kummer. *Food Funct.* **2017**, *8*, 3239–3249. [CrossRef]
31. Tsukatani, T.; Suenaga, H.; Shiga, M.; Noguchi, K.; Ishiyama, M.; Ezoe, T.; Matsumoto, K. Comparison of the WST-8 colori-metric method and the CLSI broth microdilution method for susceptibility testing against drug-resistant bacteria. *J. Micro-biol. Methods* **2012**, *90*, 160–166.
32. Stolz, T.; Huertas, M.E.; Mendoza, A. Assessment of air quality monitoring networks using an ensemble clustering method in the three major metropolitan areas of Mexico. *Atmos. Pollut. Res.* **2020**, *11*, 1271–1280. [CrossRef]
33. Wang, Q.; Wang, C.; Feng, Z.; Ye, J. Review of K-means clustering algorithm. *Electron. Des. Eng.* **2012**, *20*, 21–24.
34. Govender, P.; Sivakumar, V. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmos. Pollut. Res.* **2020**, *11*, 40–56. [CrossRef]
35. Pasqualoto, K.F.; Teófilo, R.F.; Guterres, M.; Pereira, F.S.; Ferreira, M. A study of physicochemical and biopharmaceutical properties of Amoxicillin tablets using full factorial design and PCA biplot. *Anal. Chim. Acta* **2007**, *595*, 216–220. [CrossRef]
36. Cardoso, R.V.C.; Carocho, M.; Fernandes, Â.; Pinela, J.; Stojkovic, D.; Sokovic, M.; Zied, D.C.; Cobos, J.D.V.; González-Paramás, A.M.; Ferreira, I.C.F.R.; et al. Antioxidant and Antimicrobial Influence on Oyster Mushrooms (*Pleurotus ostreatus*) from Substrate Supplementation of Calcium Silicate. *Sustainability* **2021**, *13*, 5019. [CrossRef]
37. Abrar, A.S.; Kadam, J.A.; Mane, V.P.; Patil, S.S.; Baig, M.M.V. Biological efficiency and nutritional contents of *Pleurotus florida* (Mont.) singer cultivated on different agro-wastes. *Nat. Sci.* **2009**, *7*, 1545–1740.
38. Da Silva, M.C.S.; Naozuka, J.; da Luz, J.M.R.; de Assunção, L.S.; Oliveira, P.V.; Vanetti, M.C.D.; Bazzolli, D.M.S.; Kasuya, M.C.M. Enrichment of *Pleurotus ostreatus* mushrooms with selenium in coffee husks. *Food. Chem.* **2012**, *131*, 558–563. [CrossRef]
39. Liu, H.; Chen, N.; Feng, C.; Tong, S.; Li, R. Impact of electrostimulation on denitrifying bacterial growth and analysis of bacterial growth kinetics using a modified Gompertz model in a bio-electrochemical denitrification reactor. *Bioresour. Technol.* **2017**, *232*, 344–353. [CrossRef] [PubMed]
40. Valencia del Toro, G.; Ramírez-Ortiz, M.E.; Flores-Ramírez, G.; Costa-Manzano, M.R.; Robles-Martínez, F.; Garín Aguilar, M.E.; Leal-Lara, H. Effect of *Yucca schidigera* bagasseas substrate for Oyster mushroom on cultivation parameters and fruit body quality. *Rev. Mex. Ing. Quim.* **2018**, *17*, 835–846. [CrossRef]

CAPÍTULO IV

SEGUNDO CASO DE CONTRIBUCIÓN A LA MINERÍA DE DATOS

CAPÍTULO IV

4. SEGUNDO CASO DE CONSTRIBUCIÓN A LA MINERÍA DE DATOS

4.1. Metodología

Los datos de características miceliales y culturales de las cepas de *Pleurotus* spp. fueron obtenidos de forma experimental por el autor de esta tesis, con la colaboración el Ing. Cristian Vargas (Gerente General de Ecuahidrolizados S.A.).

4.1.1. Material biológico

Para este estudio se utilizaron 50 cepas híbridas de *Pleurotus ostreatus* (PO) y 50 cepas híbridas de *Pleurotus djamor* (PD). Las cepas son híbridos de apareamiento de neohaplontes compatibles de *Pleurotus djamor* o monokariones de *Pleurotus ostreatus*, los neohoplantes se obtuvieron mediante dedicariotización química, las cepas de *Pleurotus* se mantienen en placas MEA y se depositaron en la colección de hongos del Laboratorio de Investigación y Desarrollo de Ecuahidrolizados.

4.1.2. Preparación de mezclas de medios de cultivo

Las cepas se cultivaron utilizando dos mezclas de medios de cultivo:

M1 = 18 g de extracto de malta, 15 g de agar bacteriológico y 20 g de harina de arroz en 1 L de agua destilada.

M2 = 18 g de extracto de malta, 15 g de agar bacteriológico y 20 g de harina de soya en 1 L de agua destilada.

Las placas con el medio solidificado se incubaron a 28 ° C durante 24 h para comprobar la esterilidad.

4.1.3. Determinación del área micelial

El diámetro de la colonia se midió diariamente hasta que el micelio colonizó las placas de Petri con M1 y las placas de Petri con M2, ver Eq. (6) (Valenzuela-Cobos et al., 2020):

$$A = \frac{\pi d^2}{4} \quad (6)$$

Ecuación 6. Área micelial.

4.1.4. Modelo matemático

Para calcular la velocidad de crecimiento del micelio (μ_{max}) y la fase de latencia (λ) en placas de Petri con M1 y placas de Petri con M2, el área del micelio se ajustó al modelo de Baranyi, ver Eq. (7) (Baty & Delignette-Muller, 2004):

$$y(t_{max}) = \frac{y_{max} + \ln\left((-1 + e^{\mu max\, \lambda} + e^{\mu max t})\right)}{(-1+e^{\mu max t})+e^{(\mu max\, \lambda + y_{max} - y_0)})} \quad (7)$$

Ecuación 7. Modelo Baranyi.

4.1.5. Producción de biomasa

Se cortaron dos discos de micelio (5,5 mm) de cepas de *Pleurotus ostreatus* (PO) y *Pleurotus djamor* (PD) del borde de placas de Petri con sólido líquido M1 y luego se inocularon en 100 mL de la solución de cultivo líquido (L1 = 1 L de agua destilada con maltosa (40 g L^{-1}), extracto de levadura (3 g L^{-1}) y harina de arroz 2 g L^{-1}). De lo contrario, dos discos de micelio de *Pleurotus* spp. se cortaron del borde de las placas con el sólido líquido M2 y luego se inocularon en 100 mL de la solución de cultivo líquido (L2 = 1 L de agua destilada con maltosa (40 g L^{-1}), extracto de levadura (3 g L^{-1}) y harina de soya (2 g L^{-1}).

Todos los estudios de producción se llevaron a cabo a 28 °C y 150 rpm en una incubadora con agitación durante 7 días. Las biomasas celulares se separaron usando una centrífuga de 20000 rpm a 4 °C, luego se lavó del tamiz con agua destilada, se filtró a través de papel de filtro Whatman # 1 y se secó hasta peso constante a 80 °C (Lakzian et al., 2008).

4.1.6. Producción de exopolisacáridos

El caldo de cultivo y el agua usados para lavar la biomasa de los tamices se filtraron a través de papel de filtro Whatman # 1 y se evaporaron a 50 ml a 80ºC usando una placa calefactora. Este volumen reducido se añadió a 150 mL de etanol (98%), con el fin de precipitar los exopolisacáridos (EPS). Los exopolisacáridos precipitados se filtraron y se secaron hasta peso constante a 40ºC (Rasulov et al., 2013; Wagner et al., 2004).

4.1.7. Técnicas Multivariantes

Las herramientas de análisis de agrupamientos utilizadas fueron PCA-Biplot y el algoritmo de K-medoids. Adicionalmente se realizó análisis utilizando la técnica de reglas de asociación.

4.1.8. K-medoids

El algoritmo K-medoids es un método de clasificación no supervisado (Razavi Zadegan et al., 2013). La secuencia del algoritmo K-medoids es:

Seleccione una función de comparación entre objetos. Por ejemplo, si se trata de variables cualitativas, se suele utilizar la distancia euclidiana.

$$\|X_i - X_j\| = \sqrt{\sum_{k=1}^{p} (X_{ik} - X_{jk})^2}$$

2. Calcule la Matriz Global de semejanza o diferencia, es decir, la matriz de distancias.

3. Seleccione los K patrones más lejanos como atractores iniciales.

4. Calcule y almacene la similitud o diferencia entre cada patrón y cada uno de los K objetos atractores.

5. Divida el espacio en grupos, asignando cada patrón al grupo de atractores más cercano.

6. Calcule, para cada grupo definido, su medoid.

7. Considere los medoids recién calculados como nuevos patrones de atracción.

8. Regrese al paso (4)

9. Terminar cuando el conjunto de medoids sea idéntico al de la iteración anterior.

El algoritmo K-medoids tiene un mecanismo para agrupar (mediante particiones) objetos en cualquier espacio de representación. Al calcular medoids en lugar de

centroides, el algoritmo k-medoids converge más rápido a la única solución global posible en ese espacio de representación y con ese conjunto de objetos.

La matriz de datos se transforma a formato .txt y luego se procede a cargar la matriz en el software estadístico R, con la ayuda de Rstudio, utilizando la siguiente instrucción:

```
>DATA<-KMEDIODSCINPD2
>DATA
>rownames(DATA)<-DATA$Strains
>COMPOACTM1BIPLOT<-data.frame(KMEDIODSCINPD2)
>COMPOACTM1BIPLOT
>rownames(COMPOACTM1BIPLOT)<-COMPOACTM1BIPLOT$Strains
>COMPOACTM1BIPLOT
>COMPOACTM1BIPLOT<-COMPOACTM1BIPLOT[,2:5]
>COMPOACTM1BIPLOT
>nutri<-COMPOACTM1BIPLOT
>dim(nutri)
>class(nutri)
>row.names(nutri)
>library(gplots)
>clas<-c(rep("M1",50) , rep("M2",50))
>pca1<-prcomp(nutri)
>pca1$x
>plot(pca1$x[,1:2], pch=19, col=as.factor(clas) )
>legend("bottomleft", pch=19, legend=unique(as.factor(clas)), col=unique(as.factor(clas)) , cex=1.7)
>kmeans1<-kmeans(nutri, 4, algorithm="Forgy")
>kmeans1$cluster
>library(gplots)
>balloonplot(table(kmeans1$cluster, clas))
>plot(pca1$x[,1:2], pch=19, col=as.factor(kmeans1$cluster) )
>library(ClusterR)
>set.seed(1)
>kplus<-KMeans_rcpp(nutri, 4, initializer="kmeans++")
>balloonplot(table(kplus$cluster, clas))
>plot(pca1$x[,1:2], pch=19, col=as.factor(kplus$cluster) )
```

4.1.9. PCA-Biplot

Biplot es una aproximación de una matriz realizada sin hacer suposiciones sobre distribuciones probabilísticas subyacentes que proporciona la estructura geométrica de los datos gráficamente, mostrando la variabilidad del conjunto de individuos y variables. El prefijo bi se refiere a la representación de filas y columnas simultáneas de la matriz.

Teóricamente, en una Biplot una matriz rectangular Y de orden (nxp) y rango r, por otra de rango q (q <r), mediante su descomposición en valores singulares (DVS), es decir, Y≅UΣV '

donde U y V son matrices de vectores singulares ortonormales tales que U´U = V´V = I (donde I es la matriz identidad) y ∑ es una matriz diagonal que contiene los α_k valores singulares más grandes.

Para garantizar la representación es necesaria una factorización como: Y ≅ (UΣS) (Σ1-SV´) = AB ', siendo A y B las matrices que contienen las coordenadas de los (n + p) vectores o marcadores filas ai y columnas bj para usar sobre el gráfico (i = 1,…, n; j = 1,…, p) (Cárdenas et al., 2007).

La matriz de datos se transforma a formato .txt , el software utilizado para realizar estos análisis fue MULTIBIPLOT, desarrollado por (Vicente-Villardón, 2010a), disponible en la página web: http://biplot.usal.es/multbiplot y el MultbiplotR en R (Vicente-Villardón, 2010b), disponible en el sitio web: https://CRAN.R-project.org/package=MultBiplotR. Este programa fue escrito en lenguaje R.

```
>library(MultBiplotR)
>data<-KMEDIODSCINPD2
>data
>X= data[,2:5]
>X
>bipUNEMI=PCA.Biplot(X, Scaling = 5)
>bipUNEMI
>summary(bipUNEMI)
>Inercias=data.frame(paste("Eje",1:length(bipUNEMI$EigenValues)),bipUNEMI$EigenValue
s, >bipUNEMI$Inertia, bipUNEMI$CumInertia)
>colnames(Inercias)=c("Eje", "Valor Propio", "Inercia", "Inercia acumulada")
>library(knitr)
>kable(Inercias)
>kable(bipUNEMI$ColContributions)
>plot(bipUNEMI, mode="ah", margin=0.05, ShowBox=TRUE)
>bipUNEMI=AddCluster2Biplot(bipUNEMI, NGroups=4, ClusterType="hi",
method="ward.D", Original=TRUE)
>plot(bipUNEMI, PlotClus=TRUE,ShowAxis=TRUE)
```

4.1.10. Reglas de asociación

La minería de reglas de asociación se considera la tarea principal en la minería de datos. Una regla de asociación expresa una relación interesante entre diferentes atributos

(Abdel-Basset et al., 2018).

Una regla de asociación implica la forma X ⇒ Y, donde X e Y son conjuntos de elementos; X es el cuerpo e Y es la cabeza. Una regla se puede evaluar mediante dos medidas, denominadas confianza y apoyo. El soporte para la regla de asociación X ⇒ Y es el porcentaje de transacciones que contienen tanto el conjunto de elementos X como Y entre todas las transacciones. La confianza para la regla de asociación X ⇒ Y es el porcentaje de transacciones que contienen un conjunto de elementos Y entre las transacciones que contienen un conjunto de elementos X. El soporte representa la utilidad de las reglas descubiertas y la confianza representa la certeza de las reglas (Choi et al., 2005).

La matriz de datos se transforma a formato .txt y luego se procede a cargar la matriz en el software estadístico R, con la ayuda de Rstudio, utilizando la siguiente instrucción:

```
>installed.packages("data.table")
>library(data.table)
>data<-KMEDIODSCINPD2
>trx<- as.data.table(data)
>head(trx)
>trx<- melt.data.table(data= trx, id.vars = "Strains", na.rm = T)[value==1, .(Strains,variable)]
>trx
>trx<- split(trx$variable, trx$Strains)
>trx
>library(arules)
>trx<- as(trx, "transactions")
>trx
>reglas1<- apriori(trx,parameter = list(support=0.02, confidence=0.90))
>inspect(reglas1)
>reglas1=sort(reglas1, by="confidence", decreasing = T)
>inspect(reglas1)
>library(arulesViz)
>plot(reglas1, method="graph")
```

4.2. Resultados y discusiones

El propósito de esta investigación fue utilizar técnicas de minería de datos y métodos de agrupación para determinar las especies biológicas de hongos comestibles como: *Pleurotus ostreatus* y *Pleurotus djamor* con las más altas características miceliales y culturales. La numeración de las cepas se realizó siguiendo la siguiente distribución:

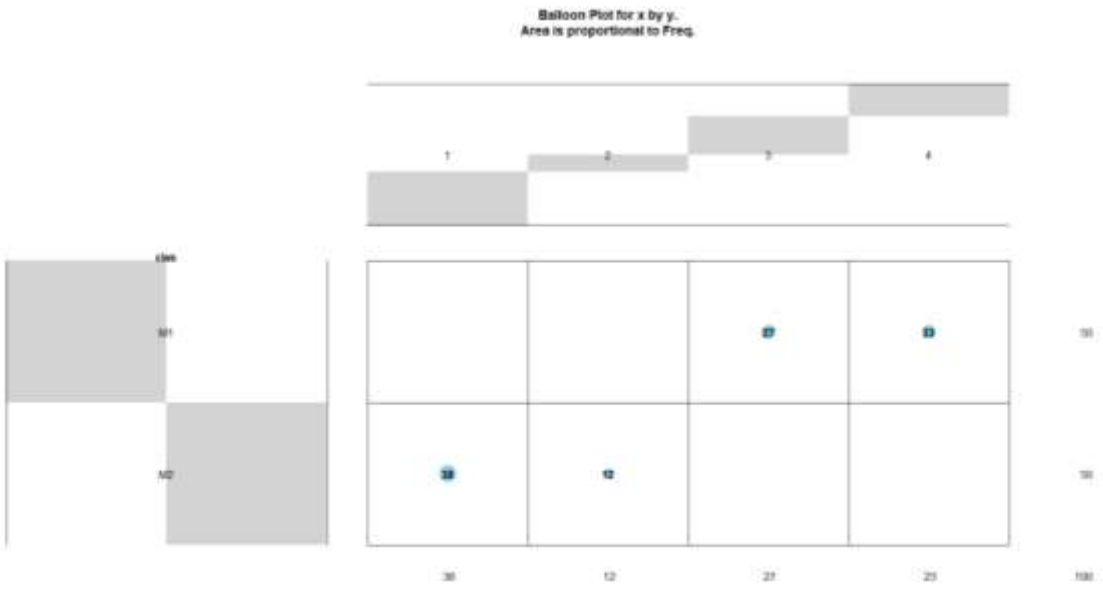
1-50: Cepas de *Pleurotus ostreatus* o *Pleurotus djamor* cultivadas en (M1 = agar extracto de malta con harina de arroz y L1 = maltosa, extracto de levadura y harina de arroz).

51-100: Cepas de *Pleurotus ostreatus* o *Pleurotus djamor* cultivadas en (M2 = agar extracto de malta con harina de soja y L2 = maltosa, extracto de levadura y harina de soya).

4.2.1. Algoritmo de K-medoids para las características miceliales y culturales de *Pleurotus* spp.

La Figura 7 muestra la distribución de los clusters para 100 objetos con 4 variables cada uno. El gráfico (a) indica la distribución de los cuatro clusters para las características miceliales y culturales de cepas de *Pleurotus ostreatus* cultivadas en cultivo sólido (M1 y M2) y también en cultivo líquido (L1 y L2), mientras que en el gráfico (b) se presenta la distribución de los cuatro grupos para las características miceliales y culturales de cepas de *Pleurotus djamor* cultivadas en medio de cultivo (M1 y M2) y también en medio líquido (L1 y L2).

(**a**)



(**b**)

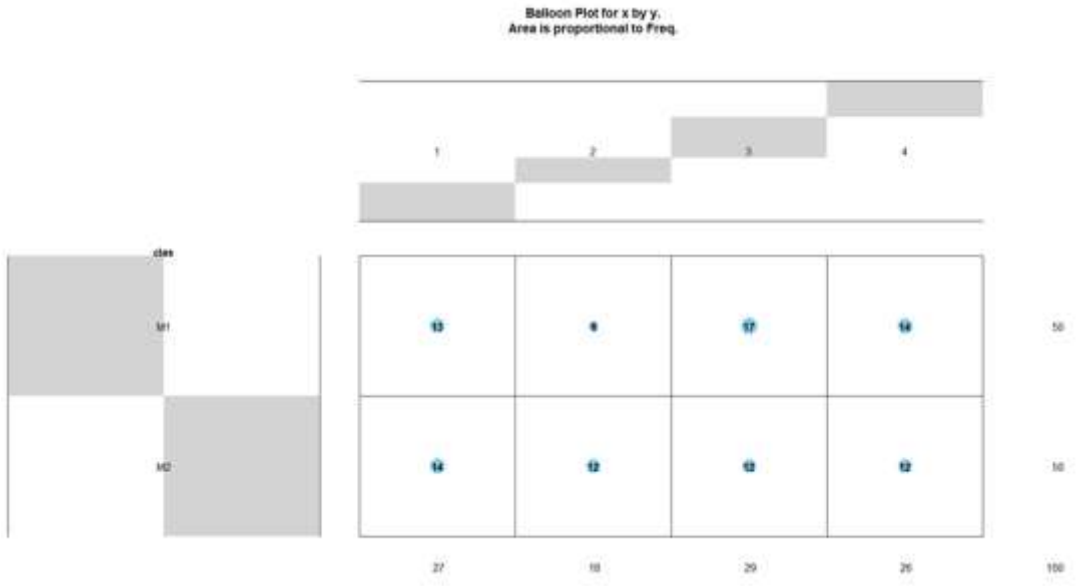


Figura 6. (a) Determinación de los 4 clusters para características miceliales y culturales de cepas de Pleurotus ostreatus cultivadas en cultivo sólido M1 y M2, y cultivo líquido L1 y L2, (b) Determinación de los 4 clusters para características miceliales y culturales de cepas de Pleurotus djamor cultivadas en cultivo sólido M1 y M2 y cultivo líquido L1 y L2.

La Figura 8 presenta los resultados de la distribución normal de 100 puntos de datos alrededor de cuatro grupos en cada gráfico. El tamaño de cada cluster tiene relación con el número de puntos de datos, Gráfico (a): el tamaño del Cluster 1 (color rojo) es 38, el tamaño del Cluster 2 (color verde) es 12, el tamaño del Cluster 3 (color azul) es 27, y el tamaño del Cluster 4 (color violeta) es 23. Las cepas de *Pleurotus ostreatus* crecen en cultivo sólido (M2) y también en cultivo líquido (L2) pertenecientes al Cluster 1 y Cluster 2, mientras que las cepas de *Pleurotus ostreatus* cultivadas en medio de cultivo (M1) y también en medio líquido (L1) pertenecientes al Cluster 3 y Cluster 4, este resultado indica que las cepas de *Pleurotus ostreatus* cultivadas en cultivo sólido (M1) y también en cultivo líquido (L1) perteneciente al Cluster 1 presentó las características miceliales y culturales más altas. Por otro lado, Gráfico (b): el tamaño del Cluster 1 (color rojo) es 27, el tamaño del Cluster 2 (color verde) es 18, el tamaño del Cluster 3 (color azul) es 29, y el tamaño del Cluster 4 (color púrpura) es 26. Cepas de *Pleurotus djamor* crecimiento en cultivo sólido (M1) y también en cultivo líquido (L1) pertenecientes al Cluster 1, Cluster 2 y Cluster 3, mientras que las cepas de *Pleurotus djamor* cultivadas en medio de cultivo (M2) y también en medio líquido (L2) pertenecientes al Cluster 3 y Cluster 4, este resultado indica que las cepas de *Pleurotus djamor* crecen en cultivo sólido (M1) y

también en cultivo líquido (L1) perteneciente al Cluster 1, Cluster 2 y Cluster 3 presentaron las características miceliales y culturales más altas. Los puntos de datos se distribuyen normalmente, los grupos varían en tamaño con puntos de datos máximos y puntos de datos mínimos (Guevara-Viejó et al., 2021).

(**a**)

(**b**)

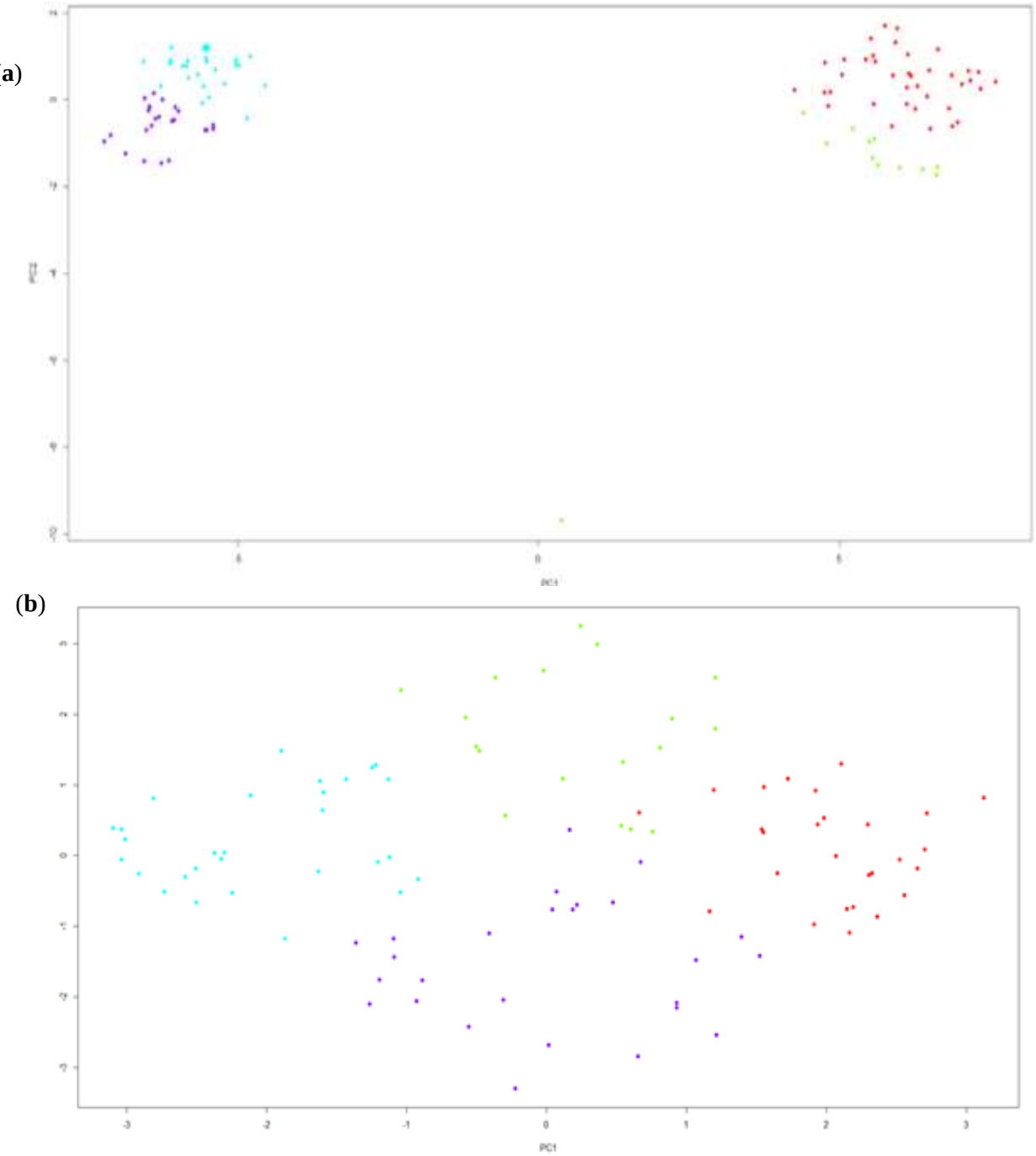


Figura 7. (a) K-medoids usando 4 clusters para características miceliales y culturales de cepas de Pleurotus ostreatus cultivadas en medios de cultivo sólidos M1 y M2, y cultivo líquido L1 y L2, (b) K-medoids usando 4 clusters para características miceliales y culturales de cepas de Pleurotus djamor cultivadas en cultivo sólido M1 y M2 y cultivo líquido L1 y L2.

La Figura 9 presenta el gráfico factorial del plano 1-2 (PCA Biplot), el Gráfico (a) muestra que la inercia acumulada asciende al 97,4%, mientras que el Gráfico (b) presenta la inercia acumulada asciende al 58,1% Además, se han calculado los clusters utilizando las coordenadas de Biplot, la descripción general de los clusters se basa en cuatro variables. Podemos ver en el Gráfico (a) diferencias importantes entre clusters, el Cluster 1 (color rojo) muestra la presencia de 50 cepas de *Pleurotus ostreatus* en crecimiento sobre crecimiento en cultivo sólido (M1) y también en cultivo líquido (L1) con mayor relación a los siguientes parámetros: velocidad máxima, biomasa y contenido de exopolisacáridos, mientras que el Cluster 2 (color verde), el Cluster 3 (color azul) y el Cluster 4 (color violeta) indican la presencia de 50 cepas de *Pleurotus ostreatus* cultivadas en cultivo sólido (M2) y también en cultivo líquido (L2) con mayor relación a la fase de latencia. De lo contrario, en el Gráfico (b) también existen diferencias entre los clus-ters, el Cluster 1 (color rojo) indica la presencia de 42 cepas de *Pleurotus djamor* crecimiento en los dos medios de cultivo y también en el cultivo líquido (L1 y L2) con mayor relación a la velocidad máxima y contenido de exopolisacáridos, mientras que el Cluster 2 (color verde) muestra la presencia de 22 cepas de *Pleurotus djamor* cultivadas en los medios de cultivo (M1 y M2) y también en el cultivo líquido (L1 y L2) con mayor relación con la fase de retardo y el contenido de biomasa, y el Cluster 3 (color púrpura) y el Cluster 4 (color azul) indican la presencia de 36 cepas de *Pleurotus djamor* crecimiento en los dos medios de cultivo y en los dos cultivos líquidos con mayor relación al contenido de exopolisacáridos.

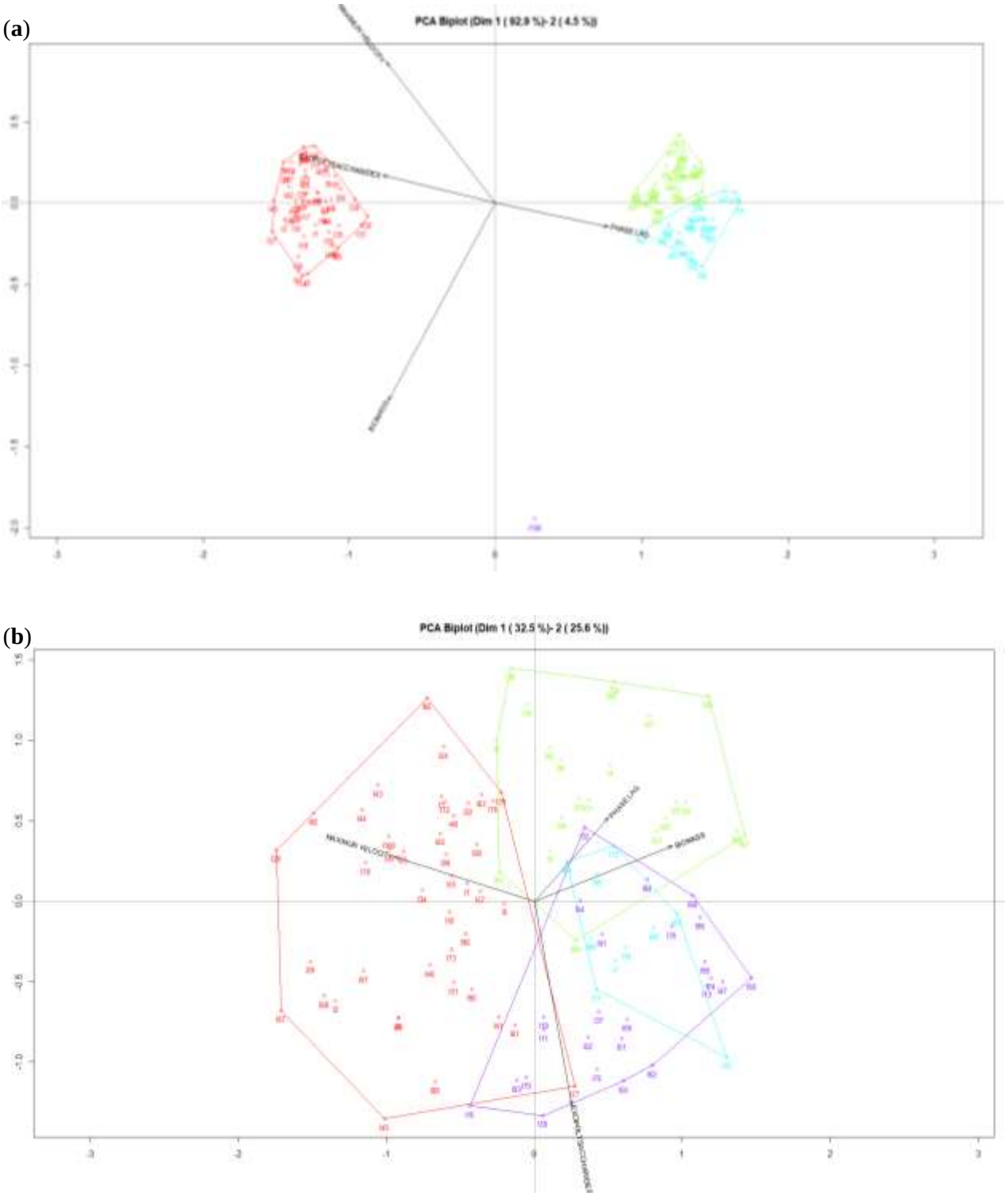


Figura 8.. (a) PCA Biplot para características miceliales y culturales de cepas de Pleurotus ostreatus cultivadas en cultivos sólidos M1 y M2, y cultivo líquido L1 y L2, (b) PCA Biplot para características miceliales y culturales de cepas de Pleurotus djamor cultivadas en cultivo sólido M1 y M2 y cultivo líquido L1 y L2.

4.2.2. Algoritmo de reglas de asociación para características miceliales y culturales de *Pleurotus* spp.

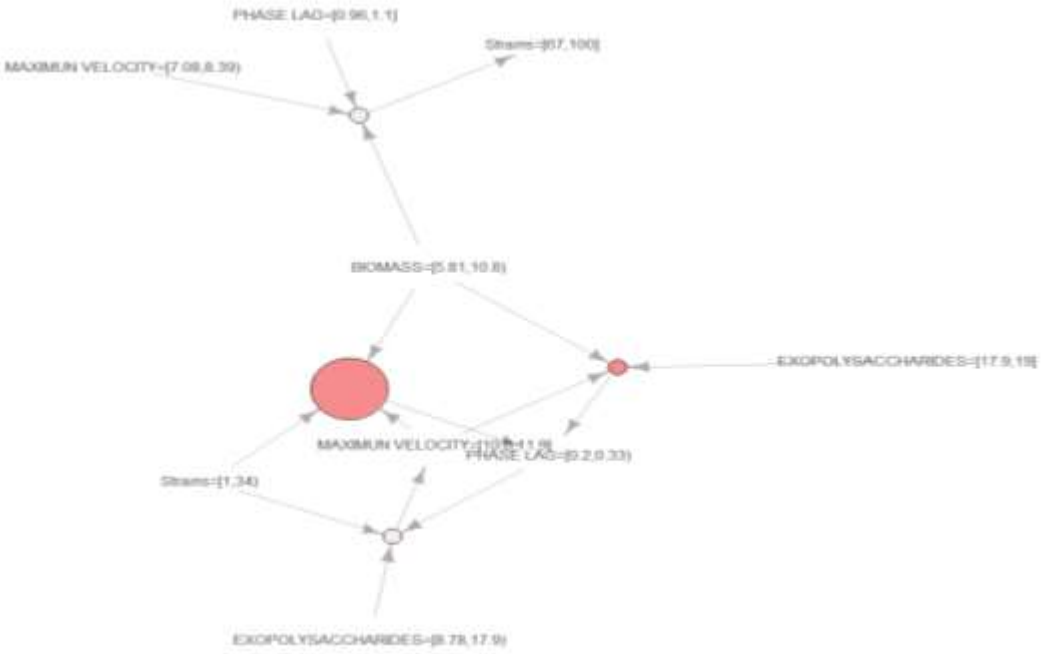
La Figura 10 presenta el uso de reglas de asociación al conjunto de datos de *Pleurotus ostreatus* y *Pleurotus djamor* cultivados en cultivo sólido (M1 y M2) y también en cultivo líquido (L1 y L2). El algoritmo de reglas de asociación es una solución exitosa para extraer reglas alternativas, ya que proporciona una imagen completa de las asociaciones en un gran conjunto de datos.

El gráfico (a) muestra un grupo de cepas de *Pleurotus ostreatus* (1 a 33) cultivadas en cultivo sólido (M1) y el cultivo líquido (L1) con las siguientes características miceliales y culturales: velocidad máxima entre 10,8 y 11,9 h^{-1}, Lag fase desde 0.2 a 0.33 h, el contenido de biomasa varió entre 5.81 y 10.8% y exopolisacáridos entre 8.78 y 17.9%, también presenta un grupo de cepas de *Pleurotus ostreatus* (67 a 100) crecimiento en medio sólido (M2) y el líquido medio (L2) con las siguientes características miceliales y culturales: velocidad máxima entre 7.08 y 8.39 h^{-1}, fase de rezago de 0.96 a 1.1 h, y contenido de biomasa entre 5.81 y 10.8%.

El gráfico (b) presenta un grupo de cepas de *Pleurotus djamor* (1 a 33) que crecen en cultivo sólido (M1) y el cultivo líquido (L1) con las siguientes características miceliales y culturales: velocidad máxima desde 13,3 a 15 h^{-1}, fase de rezago entre 0.33 y 0.42 h, contenido de biomasa entre 16.7 y 18.4%, y exopolisacáridos entre 21.2 y 22.5%, también muestra un grupo de cepas de *Pleurotus djamor* (34 a 66) con las siguientes características miceliales y culturales: velocidad máxima de 10,1 a 11,5 h^{-1}, el contenido de biomasa osciló entre el 16,7 y el 18,4% y los exopolisacáridos entre el 22,5 y el 24%.

El algoritmo de reglas de asociación proporciona una visualización que indica el grupo de cepas de *Pleurotus ostreatus* y *Pleurotus djamor* con características miceliales y culturales específicas: velocidad máxima, fase de latencia, contenido de biomasa y producción de exopolisacáridos.

(**a**)



(**b**)

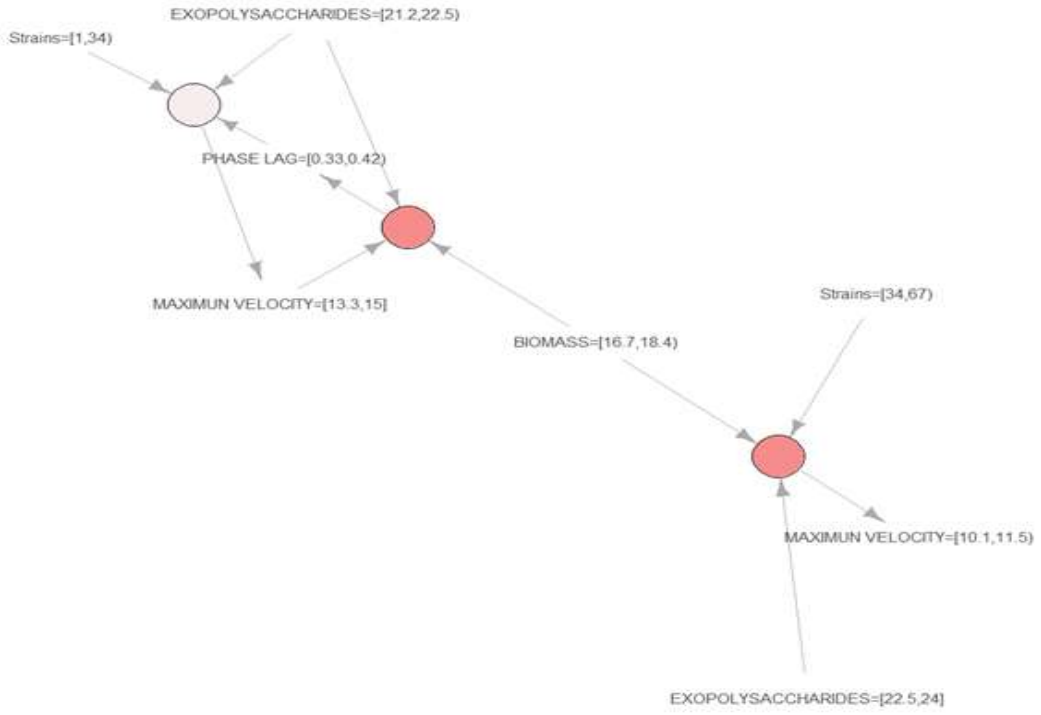


Figura 9. (a) Algoritmo de reglas de asociación para características miceliales y culturales de cepas de Pleurotus ostreatus cultivadas en cultivos sólidos M1 y M2, y cultivos líquidos L1 y L2, (b) Algoritmo de reglas de asociación para características miceliales y culturales de cepas de Pleurotus djamor cultivadas en cultivo sólido M1 y M2, y cultivo líquido L1 y L2.

Características de la Revista en que se publicó el artículo

Nombre de Revista: Journal of Fungi

Nivel de Cuartil: JCR – Q1

Factor de Impacto: 5.816

Journal of *Fungi*

MDPI

Article

Data-Mining Techniques: A New Approach to Identifying the Links among Hybrid Strains of *Pleurotus* with Culture Media

Fabricio Guevara-Viejó [1], Juan Diego Valenzuela-Cobos [1,2], Purificación Vicente-Galindo [3,4] and Purificación Galindo-Villardón [3,5,*]

1 Facultad de Ciencias e Ingeniería, Universidad Estatal de Milagro (UNEMI), Milagro 091050, Ecuador; jguevarav@unemi.edu.ec (F.G.-V.); juan_diegova@hotmail.com (J.D.V.-C.)
2 School of Medicine, Universidad Espíritu Santo, Guayaquil 092301, Ecuador
3 Department of Statistics, University of Salamanca, 37008 Salamanca, Spain; purivg@usal.es
4 Institute for Biomedical Research of Salamanca (IBSAL), 37006 Salamanca, Spain
5 Centro de Investigación Institucional, Universidad Bernardo O'Higgins, Av. Viel 1497, Santiago 8320000, Chile
* Correspondence: pgalindo@usal.es; Tel.: +34-646665034

Abstract: In this study, a data set of mycelial and cultural characteristics of hybrid strains of *Pleurotus ostreatus* and *Pleurotus djamor* were analyzed using three data-mining techniques: the K-medoids clustering algorithm, PCA biplot and the association rules algorithm. The characteristics evaluated were as follows: maximum velocity; lag phase; biomass; and exopolysaccharides content in the cultivation of 50 hybrid strains of *Pleurotus ostreatus* and 50 hybrid strains of *Pleurotus djamor.* Different mixtures of culture media were used to supplement Ecuadorian agricultural products. Data of the parameters obtained in the experimental methods were grouped into four clusters, obtaining a presentation of the hybrid strains of *Pleurotus* with a higher relation to each characteristic measured. Data-mining tools showed the hybrid strains cultivated on solid-culture media (M1 = malt extract agar and rice flour) and liquid-culture media (L1 = maltose, yeast extract and rice flour) presented the highest mycelial and cultural characteristics. These results are good indicators to improve the industrial production of edible fungi by using rice flour in the cultivation, contributing to the mushroom market and circular economy.

Keywords: circular economy; data-mining techniques; mushroom market

check for updates

Citation: Guevara-Viejó, F.; Valenzuela-Cobos, J.D.; Vicente-Galindo, P.; Galindo-Villardón, P. Data-Mining Techniques: A New Approach to Identifying the Links among Hybrid Strains of *Pleurotus* with Culture Media. *J. Fungi* **2021**, *7*, 882. https://doi.org/10.3390/jof7100882

Academic Editors: Monika Gąsecka and Zuzanna Magdziak

Received: 27 September 2021
Accepted: 13 October 2021
Published: 19 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Data mining is the process of extracting understandable and useful information from big data, with its main objective being to find hidden or implicit information which is not possible to obtain by methods of conventional statistics. The beginning of the mining process is formed generally by records from operational databases or data warehouses [1]. Data mining is an extraction process of information that involves the search for behavior patterns that are hidden at first glance among large amounts of information [2]. There are various algorithms and techniques that describe the interesting relationship between different attributes, such as the K-medoids clustering algorithm, PCA biplot and the association rules algorithm.

In the agricultural sector, where agribusinesses must make countless decisions every day, considering various intricate complexities and factors which influence them, it is necessary that an accurate estimation of the yield of the crops is involved in agricultural planning. Data-mining techniques are a necessary approach to achieve practical and effective solutions to this problem; therefore, agriculture is an obvious target for big data [3]. Agriculture in Ecuador is one of the main sectors that maintains economic dynamics and supplies raw materials to the food industry, promoting the country's food security and sovereignty [4]. The main effects of agriculture are to reduce hunger and malnutrition, improve living conditions, increase income and generate employment for

vulnerable groups living in poverty [5]. The main Ecuadorian export products that have gained global popularity over time are bananas, cocoa and flowers [6]. However, there are some agricultural products that have presented economic losses for farmers, such as rice and soybeans [7,8], because these products are offered at cheap prices in countries near to Ecuador. Therefore, the official prices of these products are not respected by the industry [9], and in some cases, the rice and soybeans are used as animal feed.

Pleurotus genus is one of the most commercialized groups of mushrooms in the world, due to the ease of its cultivation, great economic potential, and flavor [10,11]. Moreover, it only requires tropical and subtropical climates to produce fruit bodies [12]. However, the production of *Pleurotus* in small-scale companies presents some problems, such as product contamination and a difficulty in obtaining good-quality spawns [13]. The development of hybrid strains allows for the improvement of commercial attributes by reducing the incubation time, using different agricultural residues for the cultivation of the fungi and making it more adaptable to climatic conditions [14,15]. A number of studies have focused on the biotechnological development of fungi strains and inoculum, such as the use of solid-culture media with the supplementation of agricultural products to improve the growth speed, and the influence of nutrients from the inoculum in liquid media to increase the production of biomass and exopolysaccharides [16–18]. The nutritional and environmental requirements of the edible fungi strains show direct relationships with the productivity parameters after cultivation [19]. The production of edible fungi is a profitable business due to the use of low-cost agricultural products and food waste [20].

The main goal of this study was to use data-mining techniques such as the K-medoids clustering algorithm, PCA biplot and the association rules algorithm to identify the hybrid strains of *Pleurotus ostreatus* and *Pleurotus djamor,* using culture media supplemented with the agricultural products of rice and soybeans that obtained the highest values in mycelial and cultural characteristics.

2. Materials and Methods

2.1. Mushroom Strains

For this study, 50 hybrid strains of Pleurotus ostreatus (PO) and 50 hybrid strains of Pleurotus djamor (PD) were used. The strains were hybrids obtained through the pairing of compatible neohaplonts of Pleurotus djamor or monokaryons of Pleurotus ostreatus. The neohaplonts were obtained by chemical dedikaryotization, and the Pleurotus strains were maintained on MEA dishes and deposited at the fungal collection of the Research and Development Laboratory of Ecuahidrolizados.

2.2. Chemical Dedikaryotization

The mycelium of *Pleurotus* spp. on MEA dishes was divided into four parts and put into a blender (Model N.4237, Mark: Marnie). Then, it was homogenized with 50 mL of sterile water for 60 s, and 25 µL homogenate was inoculated in 100 mL flasks with 50 mL dedikaryotization solution (20% of anhydrous glucose and 20% of peptone) and incubated at 28 °C. When the mycelium growth was noticeable in the flasks with the dedikaryotization solution, the flasks were homogenized with 50 mL sterile distilled water for 60 s, and 25 µL of this homogenate was inoculated on MEA plates and incubated at 28 °C until colonies were formed. Growing colonies were observed under the microscope and identified as neohaplonts, characterized by the absence of clamp connections [21,22].

2.3. Identification of Neohaplonts' Compatibility Types and Production of Reconstituted Strains

To identify the two types of neohaplonts (mycelium with an absence of clamp connections) in the parental strains of *Pleurotus ostreatus* or *Pleurotus djamor*, a monokaryotic component of *Pleurotus ostreatus* was randomly selected and paired in the MEA dishes with all remaining neohaplonts in the *Pleurotus ostreatus*. The same product was realized with the monokaryons in the *Pleurotus djamor*. The dikaryotic mycelium was characterized by the presence of clamp connections and verified under the microscope 10(x) [23]. Authors

have reported that this method for the production of hybrid strains presented a high degree of polymorphism among the hybrid strains and the parental strains. The high degree of polymorphism indicated the genetic diversity that existed between the strains [24].

2.4. Preparation of Mixtures of Culture Media

Hybrid strains of *Pleurotus* were cultivated using two mixtures of culture media:

M1 = 18 g of malt extract, 15 g of bacteriological agar and 20 g of rice flour in 1 L of distilled water.

M2 = 18 g of malt extract, 15 g of bacteriological agar and 20 g of soybean flour in 1 L of distilled water.

The dishes with the solidified media were incubated at 28 °C for 24 h to check for sterility.

2.5. Determination of Mycelial Area

The diameter of the colony was measured daily until the mycelium colonized the Petri dishes with M1 and the Petri dishes with M2, as seen in Equation (1) [25]:

$$A = \frac{\pi d^2}{4} \quad (1)$$

Equation (1). Mycelial Area.

2.6. Mathemical Model

To calculate the mycelial growth speed (μ_{max}) and the lag time (λ) on the Petri dishes with M1 and the Petri dishes with M2, the mycelial area was fitted to the Baranyi Model, as seen in Equation (2) [26]:

$$y(t_{max}) = \frac{y_{max} + \ln\left((-1 + e^{\mu_{max}\,\lambda} + e^{\mu_{max} t}\right)}{\left(-1 + e^{\mu_{max} t}\right) + e^{\left(\mu_{max}\,\lambda + y_{max} - y_0\right)}} \quad (2)$$

Equation (2). Baranyi Model.

2.7. Biomass Production

Two discs of mycelium (5.5 mm) from the hybrid strains of *Pleurotus ostreatus* (PO) and *Pleurotus djamor* (PD) were cut from the edge of the Petri dishes with solid liquid M1, and were then inoculated in 100 mL of the solution of liquid culture (L1 = 1 L of distilled water with maltose (40 g L^{-1}), yeast extract (3 g L^{-1}) and rice flour (2 g L^{-1})). Furthermore, two discs of mycelium from the hybrid strains of *Pleurotus* were cut from the edge of the plates with solid liquid M2, and were then inoculated in 100 mL of the solution of liquid culture (L2 = 1 L of distilled water with maltose (40 g L^{-1}), yeast extract (3 g L^{-1}) and soybean flour (2 g L^{-1})).

All production studies were carried out at 28 °C and 150 rpm in a shaking incubator for 7 days. Cellular biomasses were separated by using a 20,000 rpm centrifuge at 4 °C, and were then washed from the sieve with distilled water, filtered through Whatman #1 filter paper, and dried to a constant weight at 80 °C [27].

2.8. Exopolysaccharides Production

The culture broth and the water used to wash the biomass off the sieves were filtered through Whatman #1 filter paper and evaporated to 50 mL at 80 °C using a heating plate. This reduced volume was added to 150 mL of ethanol (98%), in order to precipitate the exopolysaccharides (EPS). The precipitated exopolysaccharides was filtered out and dried to constant weight at 40 °C [28,29].

2.9. Statistical Analysis

The mycelial characteristics of maximum velocity and lag phase, as well as the cultural characteristics of biomass and exopolysaccharides content, were measured in triplicates (for

each hybrid strain growing on the different culture media). The data-mining techniques (K-medoids, PCA biplot and association rules) were realized using R software version 4.1.1.

2.9.1. K-Medoids Clustering

The K-medoids algorithm is a method of unsupervised classification [30]. The sequence of the K-medoids algorithm is as follows:

1. Select a comparison function between objects. For example, if we are dealing with qualitative variables, we usually use the Euclidean distance;

$$\|X_i - X_j\| = \sqrt{\sum_{k=1}^{p} \left(X_{ik} - X_{jk}\right)^2}$$

2. Calculate the Global Matrix of similarity or difference, that is, the distance matrix;
3. Select the K-farthest patterns as initial attractors;
4. Calculate and store the similarity or difference between each pattern and each of the K-attractor objects;
5. Partition the space into groups, assigning each pattern to the closest attractor group;
6. Calculate, for each defined group, its medoid;
7. Consider the newly calculated medoids as new attractor patterns;
8. Return to step (4);
9. Terminate when the set of medoids is identical to that of the previous iteration.

The K-medoids algorithm has a mechanism for grouping (by partitioning) objects in any representation space. By calculating medoids instead of centroids, the K-medoids algorithm converges faster to the only possible global solution in that representation space, and with that set of objects.

2.9.2. PCA Biplot

A biplot approximates a matrix performed without making assumptions about the underlying probabilistic distributions that provide the geometric structure of the data graphically, showing the variability of the set of individuals and variables. The prefix *bi* refers to the representation of simultaneous rows and columns of the matrix.

Theoretically, in a biplot, a rectangular matrix Y of order (*nxp*) and rank r, by another of rank q (q < r), has its decomposition into singular values (DVS) given by:

$$Y \cong U\Sigma V'$$

where U and V are matrices of orthonormal singular vectors such that $U'U = V'V = I$ (where I is the identity matrix) and $\sum$ is a diagonal matrix containing the α_k greatest singular values.

To guarantee the representation is necessary, a factorization such as: $Y \cong (U\Sigma^S)$ $(\Sigma^{1-S}$ V′) = AB′, with A and B being the matrices that contain the coordinates of the $(n + p)$ vectors or markers rows a_i and columns b_j to use over the graphic (i = 1, . . . , *n*; j = 1, . . . , *p*) [31].

2.9.3. Association Rules

The mining of association rules is considered to be the main task in data mining. An association rule expresses an interesting relationship between different attributes [32].

An association rule implies the form X ⇒ Y, where X and Y are itemsets; X is the body and Y is the head. A rule can be evaluated by two measures, called confidence and support. The support for the association rule X ⇒ Y is the percentage of transactions that contain both itemset X and Y among all transactions. The confidence for the association rule X ⇒ Y is the percentage of transactions that contain an itemset Y among the transactions that contain an itemset X. Support represents the usefulness of the discovered rules, and confidence represents the certainty of the rules [33].

3. Results and Discussion

The focus of this research was to identify the links among hybrid strains of *Pleurotus* with the culture media that obtained the highest mycelial and cultural characteristics.

The mycelial characteristics measured were maximum velocity and lag phase, whereas the cultural characteristics determined were biomass and exopolysaccharides content.

The numeration of the strains followed this distribution:

Hybrid strains 1–50 of *Pleurotus ostreatus* or *Pleurotus djamor* cultivated on (M1 = malt extract agar with rice flour and L1 = maltose, yeast extract and rice flour);

Hybrid strains 51–100 of *Pleurotus ostreatus* or *Pleurotus djamor* cultivated on (M2 = malt extract agar with soybean flour and L2 = maltose, yeast extract and soybean flour).

3.1. Clustering K-Medoids Algorithm for Mycelial and Cultural Characteristics of the Hybrid Strains of Pleurotus

Table 1 indicates the distribution of the four clusters for the mycelial and cultural characteristics of the hybrid strains of *Pleurotus ostreatus* cultivated on solid culture (M1 and M2) and on liquid culture (L1 and L2). The size of cluster 1 is 33, the size of cluster 2 is 17, the size of cluster 3 is 16, and the size of cluster 4 is 34. The hybrid strains of *Pleurotus ostreatus* growing on solid culture (M2) and on liquid culture (L2) belonged to cluster 1 and cluster 2, whereas the hybrid strains of *Pleurotus ostreatus* cultivated on culture media (M1) and on liquid media (L1) belonged to cluster 3 and cluster 4. We found that the hybrid strains of *Pleurotus ostreatus* cultivated on solid culture (M1) and on liquid culture (L1) belonging to cluster 1 (yellow) presented the highest mycelial and cultural characteristics, and the hybrid strains of *Pleurotus ostreatus* grown on solid culture (M2) and on liquid culture (L2) belonging to cluster 4 (yellow) also presented the highest mycelial and cultural characteristics. The K-medoids clustering algorithm indicated that the number of strains cultivated on the different culture media presented the highest mycelial and cultural characteristics.

Table 1. Number of hybrid strains of *Pleurotus ostreatus* growing on different culture media belonging to each cluster using the clustering K-medoids algorithm.

Culture Media	Cluster 1	Cluster 2	Cluster 3	Cluster 4
M1 + L1	33	17		
M2 + L2			16	34

The yellow represents the number of *Pleurotus ostreatus* strains growing on the solid culture and the liquid media that showed the highest mycelial and cultural characteristics.

The culture medium M1 contained 18 g of malt extract, 15 g of bacteriological agar and 20 g of rice flour in 1 L of distilled water. The culture medium L1 contained 40 g of maltose, 3 g of yeast extract and 2 g of rice flour in 1 L of distilled water.

The culture medium M2 contained 18 g of malt extract, 15 g of bacteriological agar and 20 g of soybean flour in 1 L of distilled water. The culture medium L2 contained 40 g of maltose, 3 g of yeast extract and 2 g of soybean flour in 1 L of distilled water.

Table 2 presents the distribution of the four clusters for the mycelial and cultural characteristics of the hybrid strains of *Pleurotus djamor* grown on culture media (M1 and M2) and on liquid media (L1 and L2). The size of cluster 1 is 27, the size of the cluster 2 is 18, the size of the cluster 3 is 30, and the size of the cluster 4 is 25. The hybrid strains of *Pleurotus djamor* were grown on solid culture (M1) and on liquid culture (L1) belonging to cluster 1, cluster 2 and cluster 3, whereas the hybrid strains of *Pleurotus djamor* cultivated on culture medium (M2) and on liquid medium (L2) belonged to cluster 3 and cluster 4; this result indicated that the hybrid strains of *Pleurotus djamor* grown on solid culture (M1) and on liquid culture (L1) belonging to cluster 1, cluster 2 and cluster 3 (yellow) presented the highest mycelial and cultural characteristics, whereas the hybrid strains of *Pleurotus djamor* cultivated on solid culture (M2) and on liquid culture (L2) belonging to

cluster 3 (yellow) presented the highest mycelial and cultural characteristics. Data points are normally distributed, and clusters may vary in size with maximum data points and minimum data points.

Table 2. Number of hybrid strains of *Pleurotus djamor* cultivated on different culture media belonging to each cluster using the clustering K-medoids algorithm.

Culture Media	Cluster 1	Cluster 2	Cluster 3	Cluster 4
M1 + L1	13	6	14	17
M2 + L2	14	12	16	8

The variability in the characteristics of the hybrid strains (the high rates of invasion in substrates, and high productivities) was due to the separation of the nuclei during the formation of the monokaryons and their subsequent union during the formation of the dikaryon (hybrid strains) [34]. The K-medoids clustering algorithm allowed us to determine the presence of four clusters indicating the relations among the hybrid strains of *Pleurotus ostreatus* and the hybrid strains of *Pleurotus djamor* with the mycelial characteristics of maximum velocity and phase lag, and the cultural characteristics of biomass content and exopolysaccharides.

The yellow indicates the number of *Pleurotus djamor* strains cultivated on solid culture and liquid media that showed the highest mycelial and cultural characteristics.

The culture medium M1 contained 18 g of malt extract, 15 g of bacteriological agar and 20 g of rice flour in 1 L of distilled water. The culture medium L1 contained 40 g of maltose, 3 g of yeast extract and 2 g of rice flour in 1 L of distilled water.

The culture medium M2 contained 18 g of malt extract, 15 g of bacteriological agar and 20 g of soybean flour in 1 L of distilled water. The culture medium L2 contained 40 g of maltose, 3 g of yeast extract and 2 g of soybean flour in 1 L of distilled water.

3.2. PCA Biplot Algorithm for Mycelial and Cultural Characteristics of the Hybrid Strains of Pleurotus

Figure 1 presents the factorial graph of the plane 1-2 (PCA biplot). Figure 1a shows the accumulated inertia amounts to 97.4%, whereas Figure 1b presents the accumulated inertia amounts of 58.1% In addition, the clusters were calculated using the biplot coordinates, and the overview of the clusters was based on four variables. We can see in Figure 1a the important differences between the clusters; cluster 1 (red) shows the presence of 50 hybrid strains of *Pleurotus ostreatus* growing on solid culture (M1) and on liquid culture (L1) with a higher relation to the following parameters: maximum velocity, biomass and exopolysaccharides content. On the other hand, cluster 2 (brown), cluster 3 (green) and cluster 4 (blue) indicate the presence of 50 hybrid strains of *Pleurotus ostreatus* cultivated on solid culture (M2) and on liquid culture (L2) with a higher relation to the lag phase. Furthermore, Figure 1b demonstrates that there are differences between the clusters; cluster 1 (red) indicates the presence of 42 hybrid strains of *Pleurotus djamor* growing on the two culture media and in the liquid culture (L1 and L2) with a higher relation to the maximum velocity and exopolysaccharides content, whereas cluster 2 (brown) shows the presence of 22 hybrid strains of *Pleurotus djamor* cultivated on the culture media (M1 and M2) and on the liquid culture (L1 and L2) with a higher relation to the lag phase and biomass content. Cluster 3 (blue) and cluster 4 (green) indicate the presence of 36 hybrid strains of *Pleurotus djamor* growing on the two culture media and on the two liquid cultures with a higher relation to exopolysaccharides content.

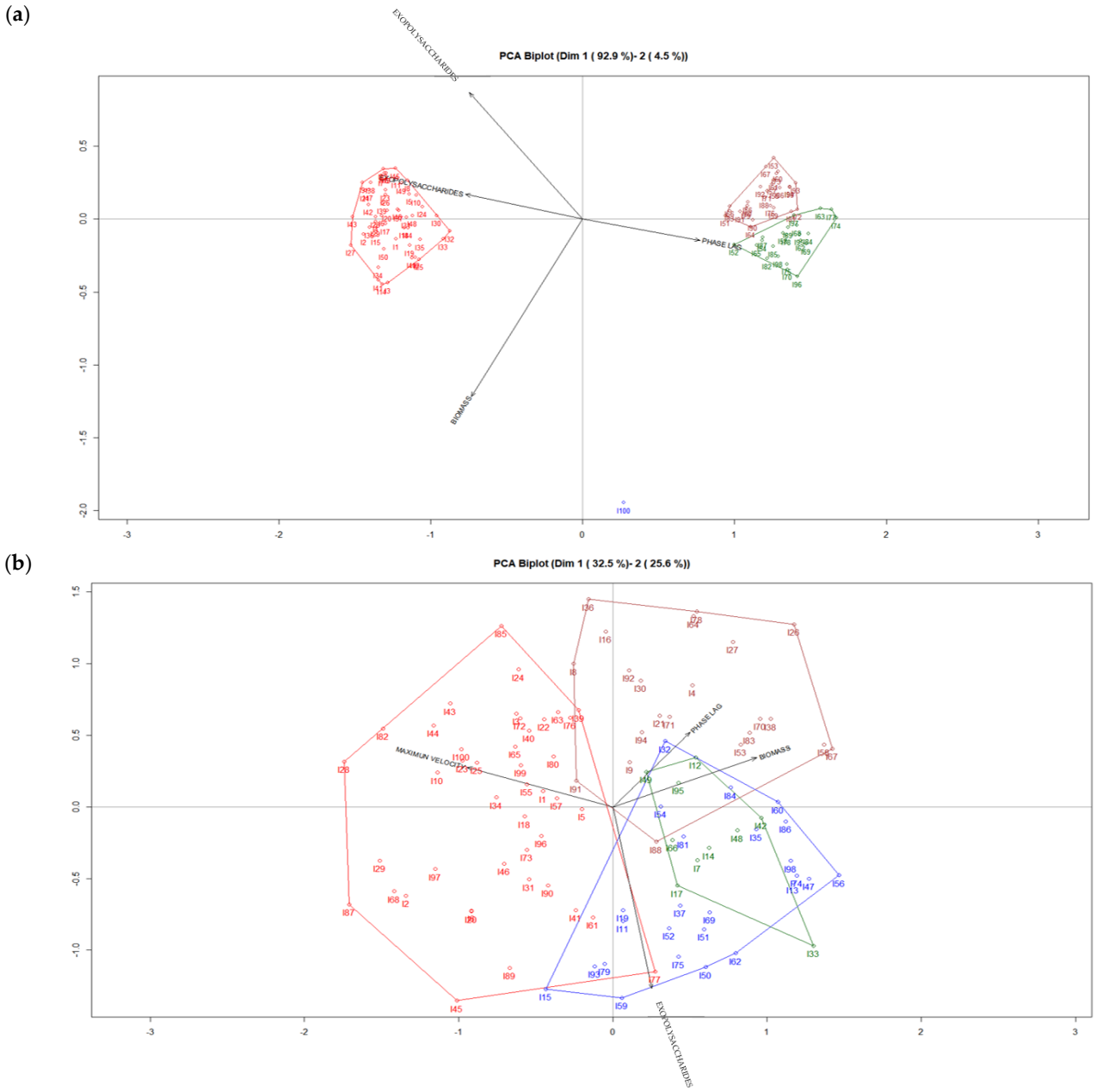


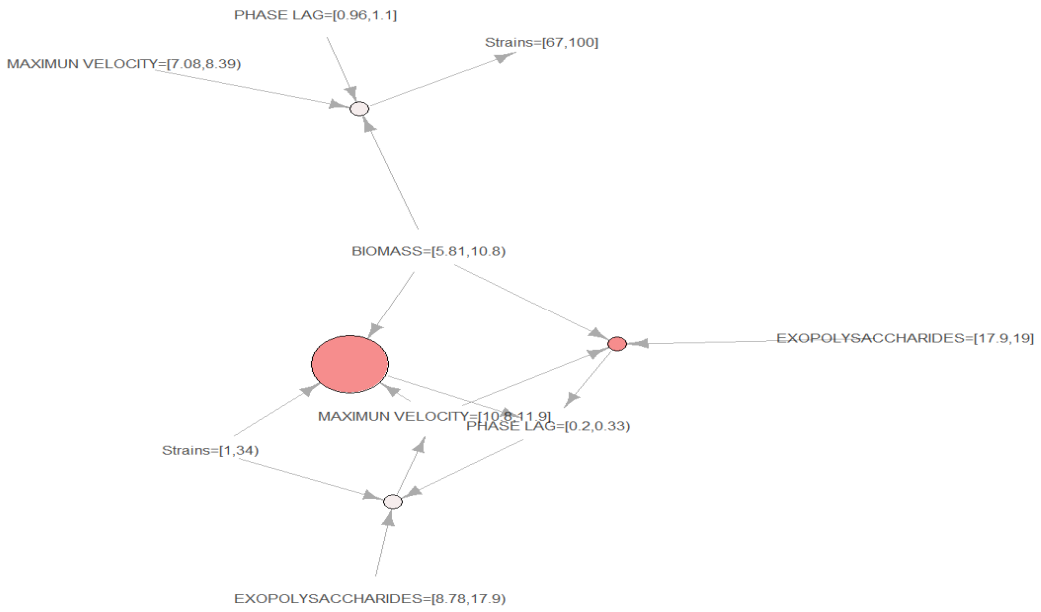
Figure 1. (**a**) PCA biplot for the mycelial and cultural characteristics of the hybrid strains of *Pleurotus ostreatus* cultivated on solid culture M1 and M2, and liquid culture L1 and L2, (**b**) PCA biplot for the mycelial and cultural characteristics of the hybrid strains of *Pleurotus djamor* cultivated on solid culture M1 and M2 and liquid culture L1 and L2.

Maltose is the most suitable carbon source for both mycelial biomass and exopolysaccharides content, whereas yeast extract is the favorable nitrogen source for both mycelial biomass and EPS production [35]. The use of submerged culture in the cultivation of edible fungi represents an alternative method of the rapid and efficient production of biomass and the production of exopolysaccharides (EPS) [36]. The main interest in the production of EPS by fungi is due to its biological and pharmacological activities, such as its immunostimulation, antitumor and hypoglycemic qualities [37]. The use of culture media supplemented with rice flour allowed the strains to colonize the substrate in a short amount of time, in comparison with the culture media supplemented with soybean flour. The solid culture M1 (malt extract agar with rice flour) and liquid culture L1 (maltose, yeast extract and rice flour) can be used to obtain the highest mycelial and cultural characteristics in the growing of hybrid strains of *Pleurotus*.

3.3. Association Rules Algorithm for Mycelial and Cultural Characteristics of Hybrid Strains of Pleurotus

Figure 2 presents the use of association rules to a data set of hybrid strains of *Pleurotus ostreatus* and *Pleurotus djamor* cultivated on solid culture (M1 and M2) and on liquid culture (L1 and L2). The association rules algorithm is a successful solution for extracting alternate rules, because it provides a complete picture of associations in a large data set.

(**a**)



(**b**)

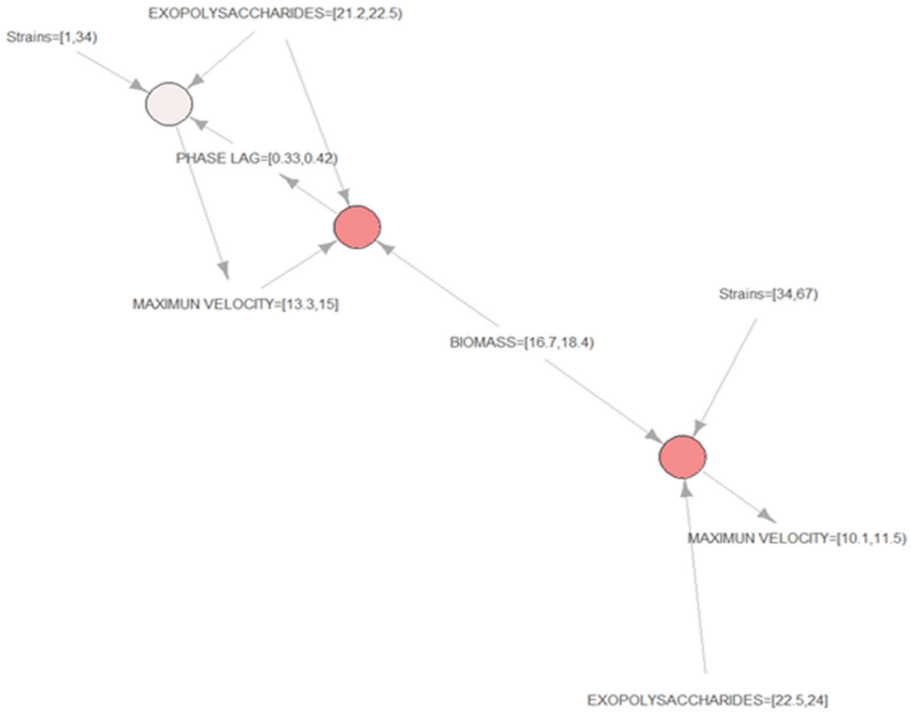


Figure 2. (**a**) Association rules algorithm for mycelial and cultural characteristics of hybrid strains of *Pleurotus ostreatus* cultivated on solid culture M1 and M2, and liquid culture L1 and L2, (**b**) Association rules algorithm for mycelial and cultural characteristics of hybrid strains of *Pleurotus djamor* cultivated on solid culture M1 and M2, and liquid culture L1 and L2.

Figure 2a shows a group of hybrid *Pleurotus ostreatus* strains (1 to 33) cultivated on solid culture (M1) and liquid culture (L1) with the following mycelial and cultural characteristics: maximum velocity between 10.8 and 11.9 h^{-1}, lag phase from 0.2 to 0.33 h, biomass content ranging from 5.81% to 10.8% and exopolysaccharides between 8.78% and 17.9%. Moreover, it also presents a group of hybrid strains of *Pleurotus ostreatus* (67 to 100) growing on solid medium (M2) and liquid medium (L2) with the following mycelial and cultural characteristics: maximum velocity between 7.08 and 8.39 h^{-1}, lag phase from 0.96 to 1.1 h, and biomass content ranging from 5.81% to 10.8%.

Figure 2b presents a group of hybrid strains of *Pleurotus djamor* (1 to 33) growing on solid culture (M1) and liquid culture (L1) with the following mycelial and cultural characteristics: maximum velocity from 13.3 to 15 h^{-1}, lag phase between 0.33 and 0.42 h, biomass content ranging from 16.7% to 18.4%, and exopolysaccharides between 21.2% and 22.5%. It also shows a group of hybrid strains of *Pleurotus djamor* (34 to 66) with the following mycelial and cultural characteristics: maximum velocity from 10.1 to 11.5 h^{-1}, biomass content ranging from 16.7% to 18.4%, and exopolysaccharides between 22.5% and 24%.

The mycelial and cultural characteristics of each strain can be used as a selection criterion for edible-mushroom-cultivation programs. The strains with the highest maximum velocity and lowest lag phases have the potential to obtain the highest commercial parameters, because they colonize the substrate in the least amount of time, while still allowing for the obtainment of the fruiting characteristics [38]. The optimal submerged culture conditions for maximum mycelial growth and exopolysaccharides production depend strongly on the type of substrates and fungal species [39].

The association rules algorithm allowed us to identify a link among a group of *Pleurotus ostreatus* and *Pleurotus djamor* hybrid strains. The solid and liquid culture media allowed us to obtain the highest mycelial characteristics of maximum velocity and phase lag, and the highest cultural characteristics of biomass content and exopolysaccharides.

4. Conclusions

The K-medoids clustering algorithm allowed us to determine the presence of four clusters that indicated the relation between the hybrid strains of *Pleurotus ostreatus* and *Pleurotus djamor,* with determined mycelial and cultural characteristics.

PCA biplot presented the specific hybrid strains of *Pleurotus ostreatus* and the specific hybrid strains of *Pleurotus djamor* cultivated on the different culture media (solid and liquid) with the characteristics measured.

The association rules algorithm identified the link between the hybrid strains of *Pleurotus ostreatus* and the hybrid strains of *Pleurotus djamor* with the solid culture (M1) and the liquid culture (L1) that obtained the highest mycelial and cultural characteristics.

Author Contributions: Conceptualization, F.G.-V. and J.D.V.-C.; formal analysis, J.D.V.-C.; investigation, F.G.-V.; methodology, F.G.-V. and J.D.V.-C.; supervision, P.G.-V. and P.V.-G.; writing—original draft, F.G.-V., J.D.V.-C., P.G.-V. and P.V.-G.; writing—review and editing, P.G.-V. and P.V.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by Universidad Estatal de Milagro (UNEMI) Scholarship.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are grateful to Facultad de Ciencias e Ingeniería de la Universidad Estatal de Milagro (UNEMI) and Ecuahidrolizados Industry.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rodríguez, J.P.F.; Pérez, A.G. Aplicación de la minería de datos en la bioinformática. *ACIMED* **2002**, *10*, 69–76.
2. Suárez, Y.R.; Amador, A.D. Herramientas de minería de datos. *Rev. Cuba. Cienc. Inform.* **2011**, *3*, 3–4.
3. Majumdar, J.; Naraseeyappa, S.; Ankalaki, S. Analysis of agriculture data using data mining techniques: Application of big data. *J. Big Data* **2017**, *4*, 20. [CrossRef]
4. Pacheco, J.; Ochoa-Moreno, W.-S.; Ordoñez, J.; Izquierdo-Montoya, L. Agricultural Diversification and Economic Growth in Ecuador. *Sustainability* **2018**, *10*, 2257. [CrossRef]
5. FAO. La contribución del crecimiento agrícola a la reducción de la pobreza, el hambre y la malnutrición. In *El Estado de la Inseguridad Alimenaria en el Mundo*; Food and Agricultural Organization of the United Nations: Washington, DC, USA, 2012; pp. 30–39.
6. Camino, S.; Diaz, V.; Pezantez, D. Posicionamiento y eficiencia del banano, cacao y flores del Ecuador en el mercado mundial. *Rev. Cien. UNEMI* **2016**, *9*, 48–53. [CrossRef]
7. Zambrano, C.E.; Arias, M.S.A.; Rodríguez, W.V.C. Factores que inciden en la productividad del cultivo de arroz en la provincia Los Ríos. *Univ. Sociedad.* **2019**, *11*, 270–277.
8. Painii-Montero, V.F.; Santillán-Muñoz, O.; Barcos-Arias, M.; Portalanza, D.; Durigon, A.; Garcés-Fiallos, F.R. Towards indicators of sustainable development for soybeans productive units: A multicriteria perspective for the Ecuadorian coast. *Ecol. Indic.* **2020**, *119*, 106800. [CrossRef]
9. Díaz, N.; Intriago, R.; Tomalá, V.; López, A.; Paredes, A.; Reyes, S. El cultivo de soya y su importancia para el Ecuador. *INNOVA Res. J.* **2016**, *1*, 77–85. [CrossRef]
10. Rosado, F.R.; Germano, S.; Carbonero, E.R.; Costa, S.M.; Iacomini, M.; Kemmelmeier, C. Biomass and exopolysaccharide production in submerged cultures of *Pleurotus ostreatoroseus* Sing. and *Pleurotus ostreatus* "florida" (Jack.: Fr.) Kummer. *J. Basic Microbiol.* **2003**, *43*, 230–237. [CrossRef]
11. Cardoso, R.V.C.; Carocho, M.; Fernandes, Â.; Zied, D.C.; Cobos, J.D.V.; González-Paramás, A.M.; Ferreira, I.C.F.R.; Barros, L. Influence of Calcium Silicate on the Chemical Properties of *Pleurotus ostreatus* var. florida (Jacq.) P. Kumm. *J. Fungi* **2020**, *6*, 299. [CrossRef]
12. Valenzuela-Cobos, J.D.; Páramo, E.D.; Arce, R.V.; Hernández, A.S.; Aguilar, M.E.G.; Lara, H.L.; del Toro, G.V. Production of hybrid strains among *Pleutorus* and *Lentinula* and evaluation of their mycelial growth kinetics on malt extract agar and wheat grain using the Gompertz and Hill models. *Emir. J. Food Agric.* **2017**, *29*, 927–935.
13. De León-Monzón, J.H.; Sánchez, J.E.; Nahed-Toral, J. El cultivo de *Pleurotus ostreatus* en los altos de Chiapas. *Rev. Mex. Micol.* **2004**, *18*, 31–38.
14. Eichlerová, I.; Homolka, H. Preparation and crossing of basidiospore-derived monokaryons—A useful tool for obtaining laccase and other ligninolytic enzyme higher-producing dikaryotic strains of *Pleurotus ostreatus*. *Antonie Leeuwenhoek* **1999**, *75*, 321–327. [CrossRef] [PubMed]
15. Chakraborty, U.; Sikdar, S.R. Production and characterization of somatic hybrids raised through protoplast fusion between edible mushroom strains *Volvariella volvacea* and *Pleurotus florida*. *World J. Microbiol. Biotechnol.* **2008**, *24*, 1481–1492. [CrossRef]
16. Chegwin, C.; Nieto, I.J. Influencia del medio de cultivo en la producción de metabolitos secundarios del hongo comestible *Pleurotus ostreatus* cultivados por fermentación en estado líquido empleando harinas de cereal como fuente de carbono. *Rev. Mex. Micol.* **2013**, *37*, 1–9.
17. Díaz-Talamantes, C.; Burrola-Aguilar, C.; Aguilar-Miguel, X.; Mata, G. In vitro mycelial growth of wild edible mushrooms from the central Mexican highlands. *Rev. Chapingo Ser. Cienc. For. Ambiente* **2017**, *23*, 3. [CrossRef]
18. Economou, C.N.; Diamantopoulou, P.A.; Philippoussis, A.N. Valorization of spent oyster mushroom substrate and laccase recovery through successive solid state cultivation of *Pleurotus*, *Ganoderma*, and *Lentinula* strains. *Appl. Microbiol. Biotechnol.* **2017**, *101*, 5213–5222. [CrossRef]
19. Arana-Gabriel, Y.; Burrola-Aguilar, C.; Garibay-Orijel, R.; Franco-Maass, S. Obtención de cepas y producción de inóculo de cinco especies de hongos silvestres comestibles de alta montaña en el centro de México. *Rev. Chapingo Ser. Cienc. For. Ambiente* **2014**, *20*, 213–226.
20. Suárez, C.; Nieto, J. Cultivo biotecnológico de macrohongos comestibles: Una alternativa en la obtención de nutracéuticos. *Rev. Iberoam. Micol.* **2016**, *30*, 1–8. [CrossRef]
21. Leal-Lara, H.; Eger-Hummel, G. A monokaryotization method and its use for genetic studies in wood-rotting basidiomycetes. *Theor. Appl. Genet.* **1982**, *61*, 1–4. [CrossRef]
22. Del Toro, G.V.; Leal-Lara, H. Estudios de compatibilidad entre cepas de *Pleurotus* spp. con cuerpos fructíferos de diversos colores. *Rev. Mex. Mic.* **1999**, *15*, 65–71.
23. Hernández, A.A.S.; Cobos, J.D.V.; Martínez, J.H.; Arce, R.V.; Gomez, Y.d.G.y.; Segura, P.B.Z.; Aguilar, M.E.G.; Lara, H.L.; del Toro, G.V. Characterization of *Pleurotus djamor* neohaplonts recovered by production of protoplasts and chemical dedikaryotization. *3 Biotech* **2019**, *9*, 24. [CrossRef]
24. Aguilar, D.L.; Zárate, S.P.B.; Villanueva, A.R.; Hernández, J.L.Y.; Aguilar, M.E.G.; Mendoza, P.C.G.; del Toro, G.V. Utilización de marcadores ITS e ISSR para la caracterización molecular de cepas híbridas de *Pleurotus djamor*. *Rev. Iberoam. Micol.* **2018**, *35*, 49–55. [CrossRef]

25. Valenzuela-Cobos, J.D.; Rodríguez-Grimón, R.O.; Jara-Bastidas, M.L.; Grijalva-Endara, A.; Zied, D.C.; Garín-Aguilar, M.E.; del Toro, G.V. Modeling of micelial growth of parental, hybrid and reconstituted strains of *Pleurotus* and *Lentinula*. *Rev. Mex. Ing. Quim.* **2020**, *19*, 165–174. [CrossRef]
26. Baty, F.; Delignette-Muller, M.L. Estimating the bacterial lag time: Which model, which precision? *Int. J. Food Microbiol.* **2004**, *91*, 261–277. [CrossRef] [PubMed]
27. Lakzian, A.; Berenji, A.R.; Karimi, E.; Razavi, S. Adsorption capability of lead, nickel and zinc by exopolysaccharide and dried cell of Ensifer meliloti. *Asian J. Chem.* **2008**, *20*, 6075–6080.
28. Rasulov, B.A.; Yili, A.; Aisa, H.A. Biosorption of metal ions by exopolysaccharide produced by *Azotobacter chroococcum* XU1. *J. Environ. Prot.* **2013**, *4*, 989–993. [CrossRef]
29. Wagner, D.; Mitchell, A.; Sassaki, G.L.; de Almeida Amazonas, M.A.L. Links between morphology and physiology of *Ganoderma lucidum* in submerged culture for the production of exopolysaccharide. *J. Biotechnol.* **2004**, *114*, 153–164. [CrossRef]
30. Razavi Zadegan, S.M.; Mirzaie, M.; Sadoughi, F. Ranked k-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets. *Knowl.-Based Syst.* **2013**, *39*, 133–143. [CrossRef]
31. Cárdenas, O.; Galindo, M.P.; Vicente-Villardón, J.L. Los métodos Biplot: Evolución y aplicaciones. *Rev. Venez. Anál. Coyunt.* **2007**, *13*, 279–303.
32. Abdel-Basset, M.; Mohamed, M.; Smarandache, F.; Chang, V. Neutrosophic association rule mining algorithm for big data analysis. *Symmetry* **2018**, *10*, 106. [CrossRef]
33. Choi, D.H.; Ahn, B.S.; Kim, S.H. Prioritization of association rules in data mining: Multiple criteria decision approach. *Expert Syst. Appl.* **2005**, *29*, 867–878. [CrossRef]
34. Clark, T.A.; Anderson, J.B. Dikaryons of the basidiomycete fungus *Schizophyllum commune*: Evolution in long term culture. *Genetics* **2004**, *167*, 1663–1675. [CrossRef] [PubMed]
35. Jia, L.; Hu, S.Z.; Xu, M. Optimization of submerged culture conditions for the production of mycelial biomass and exopolysaccharide by *Pleurotus nebrodensis*. *Ann. Microbiol.* **2007**, *57*, 389–393.
36. Confortin, F.G.; Marchetto, R.; Bettin, F.; Camassola, M.; Salvado, M.; Dillon, A.J. Production of *Pleurotus sajor-caju* strain PS-2001 biomass in submerged culture. *J. Ind. Microbiol. Biotechnol.* **2008**, *35*, 1149–1155. [CrossRef] [PubMed]
37. Lee, C.; Bae, J.T.; Pyo, H.B.; Choe, T.B.; Kim, S.W.; Hwang, H.J.; Yun, J.W. Submerged culture conditions for the production of mycelial biomass and exopolysaccharides by the edible basidiomycete *Grifola frondosa*. *Enzyme Microb. Technol.* **2004**, *35*, 369–376. [CrossRef]
38. Guadarrama-Mendoza, P.C.; del Toro, G.V.; Ramírez-Carrillo, R.; Robles-Martinez, F.; Yáñez-Fernández, J.; Garín-Aguilar, M.E.; Bravo-Villa, G. Morphology and mycelial growth rate of *Pleurotus* spp. strains from the Mexican Mixtec region. *Braz. J. Microbiol.* **2014**, *45*, 861–872. [CrossRef]
39. Ogidi, C.O.; Ubaru, A.M.; Ladi-Lawal, T.; Thonda, O.A.; Aladejana, O.M.; Malomo, O. Bioactivity assessment of exopolysaccharides produced by *Pleurotus pulmonarius* in submerged culture with different agro-waste residues. *Heliyon* **2020**, *6*, 1–7. [CrossRef]

CAPÍTULO V

TERCER CASO DE CONTRIBUCIÓN A LA MINERÍA DE DATOS

CAPÍTULO V

5. TERCER CASO DE CONSTRIBUCIÓN A LA MINERÍA DE DATOS

5.1. Metodología

Los datos de crecimiento, composición nutricional de camarones juveniles *Litopenaeus vannamei* fueron obtenidos de forma experimental por el autor de esta tesis, con la colaboración el Ing. Cristian Vargas (Gerente General de Ecuahidrolizados S.A.).4

5.1.1. Diseño experimental

Se distribuyeron aleatoriamente diez camarones *Litopenaeus vannamei* juveniles (0,70 g) por tanque (total 180 tanques de 20 L de capacidad). El experimento se realizó en la Camaronera "La Chorrera" (Xie et al., 2017; Arambul Munoz et al., 2019).

5.1.2. Preparación de las mezclas

Las dos mezclas utilizadas en el experimento se realizaron utilizando la siguiente composición:

Mezcla 1 (M1): Pellets mezclados con el ligante (hidrolizado de subproductos de sardina al 20%). La formulación fue de 200 mL de aglutinante de sardina en 2 L de agua para 25 kg de pellets.

Mezcla 2 (M2): Pellets mezclados con el ligante (hidrolizado de subproductos de sardina al 30%). La formulación fue de 200 mL de aglutinante de sardina en 2 L de agua para 25 kg de pellets.

5.1.3. Alimentación y dietas experimentales

Se distribuyeron aleatoriamente diez camarones juveniles por tanque (180 tanques), y los camarones juveniles del Nilo se alimentaron cuatro veces al día. En los comederos se pusieron 200 g de mezcla (M1) o 200 g de mezcla (M2). Los camarones fueron alimentados cuatro veces al día en los siguientes horarios: 09:00 am, 11:30 am, 2:00 pm y 4:30 pm, y las heces se retiraron todos los días. Las pruebas se realizaron durante siete semanas consecutivas (Mmanda et al., 2020). Los experimentos se realizaron por triplicado.

5.1.4. Parámetros de crecimiento de camarones juveniles

Se contaron, pesaron y midieron tres camarones de cada tanque (después de siete semanas de la prueba de alimentación) para determinar: rendimiento de crecimiento, incluida la ganancia de peso (WG %), tasa de crecimiento específica (SGR %), eficiencia alimenticia (FE %), proteína índice de eficiencia (PER) y porcentaje de supervivencia (S %) usando las siguientes ecuaciones (8-12) (Mohanty et al., 1999; Bae et al., 2020).

$$WG\ (\%) = \frac{peso\ final\ (g) - peso\ inicial\ (g)}{peso\ inicial\ (g)} \times 10 \quad (8)$$

Ecuación 8. Ganancia de peso de camarones juveniles.

$$SGR\ (\%) = \frac{ln\ peso\ final\ (g) - ln\ peso\ inicial\ (g)}{días} \quad (9)$$

Ecuación 9. Tasa de crecimiento específica.

$$FE\ (\%) = \frac{peso\ final\ (g) - peso\ inicial\ (g)}{ración\ alimenticia\ (g)} \times 100 \quad (10)$$

Ecuación 10. Eficiencia alimenticia.

$$PER = \frac{peso\ ganado\ húmedo\ (g)}{ingesta\ de\ proteínas\ (g)} \quad (11)$$

Ecuación 11. Índice de eficiencia de proteína.

$$S\ (\%) = \frac{número\ inicial\ de\ especies}{número\ final\ de\ especies} \times 100 \quad (12)$$

Ecuación 12. Porcentaje de supervivencia de camarones juveniles.

5.1.5. Composición nutricional de los camarones juveniles

Se realizó un análisis proximal de las muestras de acuerdo con la Asociación de Químicos Analíticos Oficiales (AOAC). La humedad se determinó secando muestras de camarones a 100 °C hasta peso constante. El nitrógeno (N) se determinó mediante el método de Kjeldahl y el contenido de proteína se calculó mediante el factor 6,25. Se utilizó el método soxhlet para determinar el lípido crudo. El contenido de cenizas se midió calentando las muestras a 600 °C durante 24 h (Valencia del Toro et al., 2018; Valenzuela-Cobos et al., 2019; AOAC, 2002). Todos los análisis de prueba se realizaron por triplicado por dieta experimental.

5.1.6. Análisis estadístico

Las técnicas de minería de datos (Algoritmo de agrupamiento de K-means y PCA Biplot) se realizaron utilizando el software R versión 4.1.1.

5.1.7. Algoritmo de agrupamiento de K-means

El algoritmo K-means consiste en agrupar un conjunto de datos (M bloques o vectores de muestra extraídos del conjunto de entrenamiento) en grupos o clusters (K celdas de cuantización, tales que K < M + 1), de manera que los vectores de un mismo grupo presentan alta similitud entre sí y tienen poca similitud con vectores de otros grupos. Esta técnica indica que cada vector de entrenamiento (bloque de muestras del conjunto de datos original) pertenece a una y solo una celda de cuantificación.

Sea X = {~ xj}, j = 1, 2, ..., M + 1 un conjunto de entrenamiento compuesto por M vectores N-dimensionales, con M ≫ K. El algoritmo K-means divide el espacio vectorial R^N asignado a cada vector de entrenamiento a un solo grupo a través de la búsqueda del vecino más cercano (VMP). Precisamente, $\overrightarrow{x_j}$ pertenecerá al grupo (celda de cuantificación) V ($\overrightarrow{w_i}$ if d ($\overrightarrow{x_j}$, $\overrightarrow{w_i}$) <d ($\overrightarrow{x_j}$, $\overrightarrow{w_a}$), ∀a ≠ i, donde d ($\overrightarrow{x_j}$, $\overrightarrow{w_i}$) denota la distancia euclidiana cuadrática entre $\overrightarrow{x_j}$ y $\overrightarrow{w_i}$. En este caso, $\overrightarrow{w_i}$ se dice que es el VMP de $\overrightarrow{x_j}$. La búsqueda de VMP se puede asociar con una función de pertenencia, definida por.

$$\mu_i(\overrightarrow{x_j}) = \begin{cases} 1, si\ \overrightarrow{w_i} = VMP(\overrightarrow{x_j}) \\ 0, otros\ casos \end{cases}$$

Así, la distorsión obtenida al representar todos los vectores del conjunto de entrenamiento por los respectivos VMP viene dada por

$$J_1 = \sum_{i=1}^{K}\sum_{j=1}^{M} \mu_i(\overrightarrow{x_j}) d(\overrightarrow{x_j}, \overrightarrow{w_i})$$

Para minimizar J_1, los vectores $\overrightarrow{w_i}$ se actualizan de la siguiente manera:

$$\overrightarrow{w_i} = \frac{\sum_{j=1}^{M} \mu_i(\overrightarrow{x_j}) d(\overrightarrow{x_j})\, \overrightarrow{x_j}}{\sum_{j=1}^{M} \mu_i(\overrightarrow{x_j})}\ , i = 1,2, \dots, K$$

Después de inicializar el conjunto de vectores $\overrightarrow{w_i}$, i = 1, 2, ..., K, el algoritmo K-means se puede resumir de la siguiente manera:

1. Partición: el conjunto de entrenamiento se divide en K grupos de acuerdo con la regla VMP.

2. Los nuevos vectores de código son los centroides de los clusters, calculados según la Ecuación $\overrightarrow{w_i}$.

3. Prueba de convergencia: criterio de parada del algoritmo.

Los pasos de partición y actualización se llevan a cabo hasta que se satisfacen los criterios de parada. Precisamente, el algoritmo se detiene al final de la t-ésima iteración si

$$\frac{J_1(t-1) - J_1(t)}{J_1(t)} \leq \in,$$

donde $\in$ es un parámetro del algoritmo, llamado umbral de distorsión, y $J_1(t)$ denota una distorsión obtenida en la partición de la t-ésima iteración (Madeiro et al., 2012).

Adicionalmente, se aplicó un PCA Biplot (Pasqualoto et al., 2017) para explorar y visualizar los diferentes parámetros y las respuestas más relevantes.

5.2. Resultados y discusiones

El enfoque de este trabajo fue determinar la viabilidad del uso de aglutinante acuícola con sardina en dietas de camarones juveniles *L. vannamei* y evaluar su influencia en parámetros comerciales: ganancia de peso, tasa de crecimiento específica, eficiencia alimenticia, eficiencia proteica ratio, porcentaje de supervivencia, contenido de humedad, proteína bruta, lípidos brutos y contenido de cenizas.

La numeración de los camarones *L. vannamei* juveniles por tanque se realizó siguiendo la siguiente distribución:

1–90: Promedio de tres camarones *L. vannamei* juveniles por tanque alimentados con la mezcla M1 o M2.

5.2.1. Minería de datos para parámetros de crecimiento de camarones juveniles

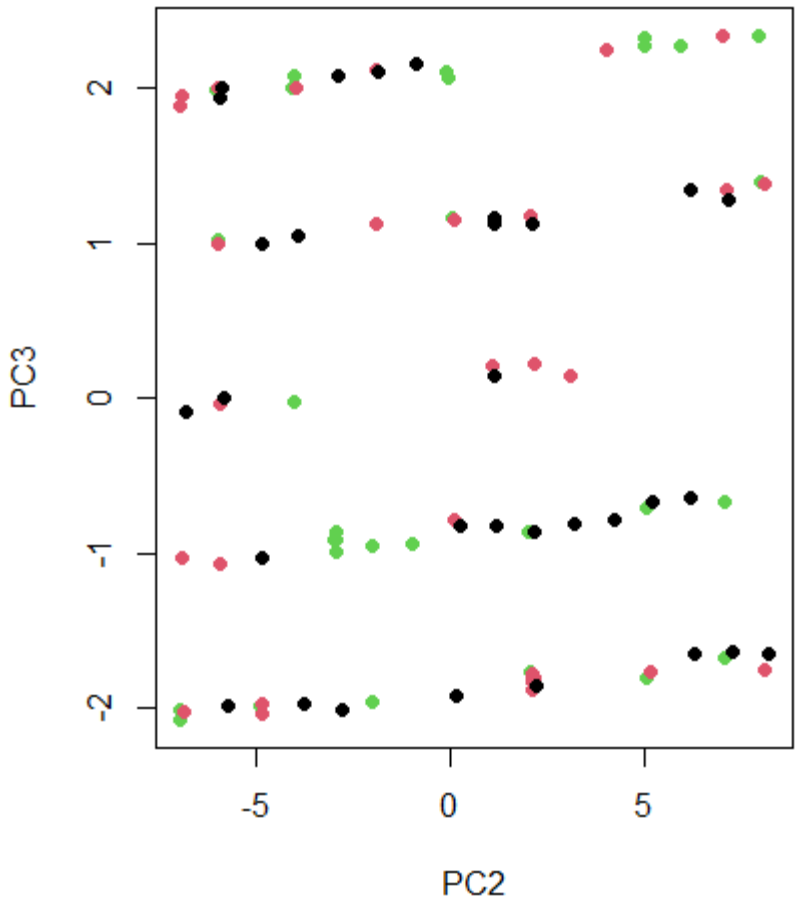
La Figura 11 presenta el uso del algoritmo de agrupamiento del método K-means para 90 objetos de 5 variables cada uno, utilizando el software RStudio. El gráfico (a) presenta el uso de tres grupos para el crecimiento de camarones juveniles *L. vannamei* alimentados con la mezcla M1 después de siete semanas, mientras que en el gráfico (b) muestra el uso de tres grupos para el crecimiento de camarones juveniles *L. vannamei* alimentados con la mezcla M2 después de siete semanas. Los resultados muestran la distribución normal de 90 puntos de datos alrededor de tres grupos en cada gráfico. El color de los diferentes clusters mostró las muestras específicas que presentaron la característica más alta medida. Los clusters permiten la separación de un conjunto de objetos en subconjuntos que no se superponen; los objetos en el grupo son similares y diferentes a los objetos en el otro grupo (Razavi et al., 2013).

El tamaño de cada grupo tiene relación con el número de puntos de datos, el gráfico (a): el tamaño del Grupo 1 (color rojo) es 28, el tamaño del Grupo 2 (color negro) es 32 y el tamaño del Grupo 3 (color verde) es 30. Los camarones juveniles *L. vannamei* alimentados con la mezcla M1 perteneciente al Grupo 2 mostraron los valores más altos de rendimiento de crecimiento. De lo contrario, el gráfico (b): el tamaño del Grupo 1 (color negro) es 30, el tamaño del Grupo 2 (color verde) es 29 y el tamaño del Grupo 3 (color rojo) es 3 camarones juveniles *L. vannamei* alimentados con la mezcla M2 perteneciente al Cluster 3 presentó los mayores valores de rendimientos de crecimiento. Dado que los puntos de datos se distribuyen normalmente, los grupos varían en tamaño

con puntos de datos máximos y puntos de datos mínimos.

El aglutinante de acuicultura se puede utilizar para mezclarlo con otros ingredientes como: antibióticos, vitaminas y ácidos orgánicos para controlar las infecciones bacterianas y mejorar el rendimiento reproductivo de los camarones y la tasa de eclosión de los huevos (Valenzuela-Cobos & Vargas-Farías, 2020). No existe un efecto beneficioso de aumentar la frecuencia de alimentación o el tamaño de la ración sobre el crecimiento o la supervivencia de los camarones (*L. vannamei*) (Velasco et al., 1999).

(**a**)



(**b**)

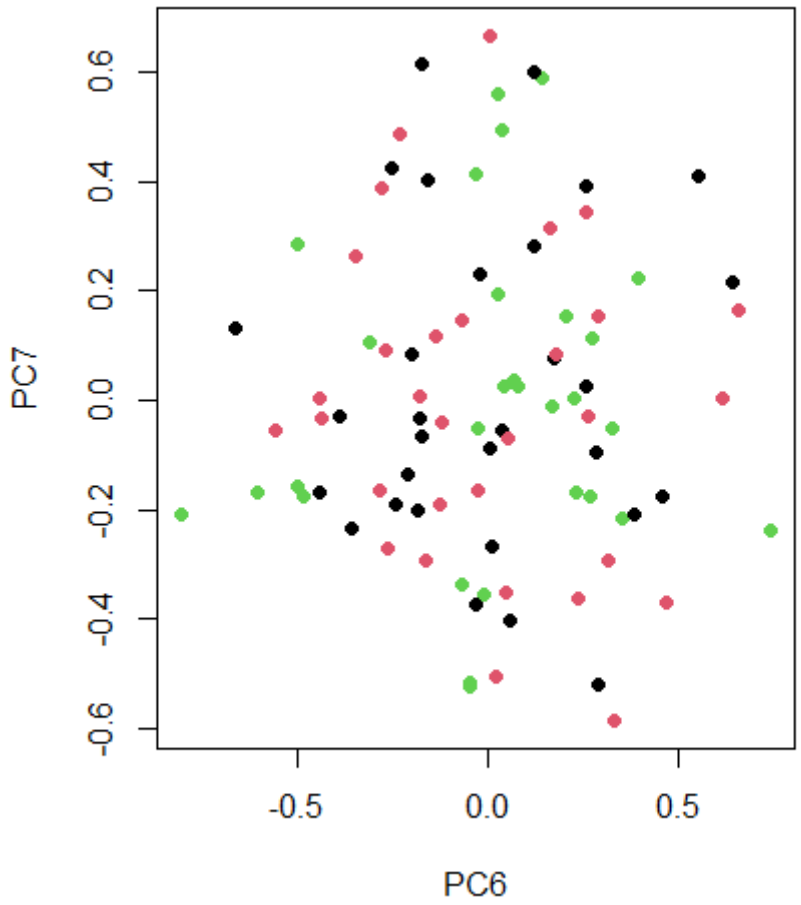


Figura 10. (a) Algoritmo de agrupamiento de K means para el rendimiento de crecimiento de camarones juveniles L. vannamei alimentados con la mezcla 1; (b) Algoritmo de agrupamiento de K means para el crecimiento de camarones juveniles L. vannamei alimentados con la mezcla 2.

La Figura 12 muestra el gráfico factorial del plano 1–2 (PCA-Biplot); el gráfico (a) presenta la inercia acumulada asciende al 49,1%, mientras que el gráfico (b) presenta la inercia acumulada asciende al 47%.

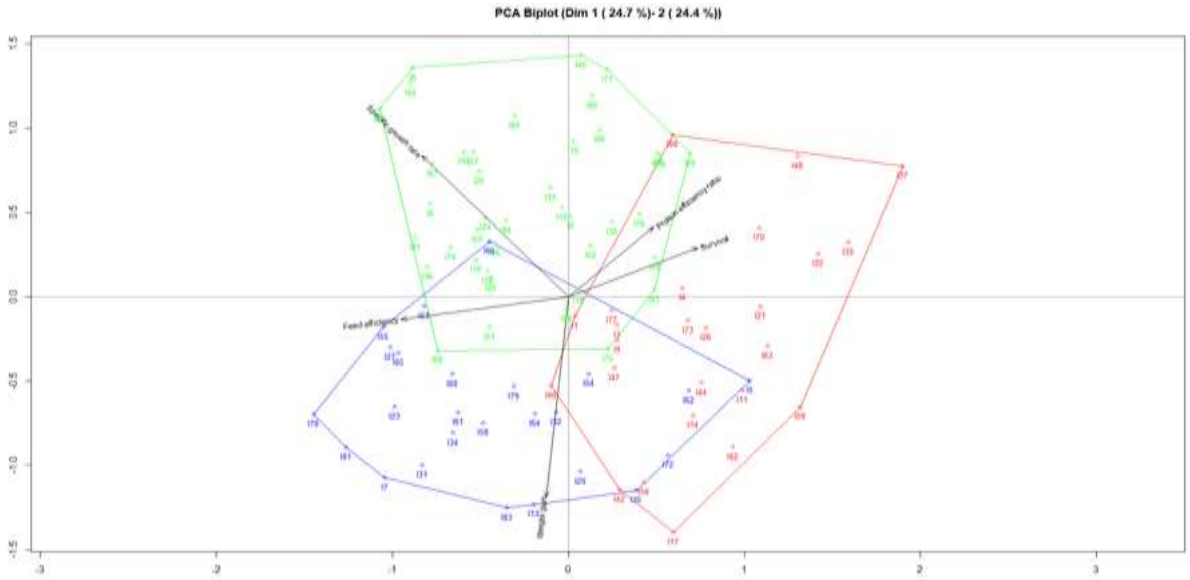
Además, los clusters se han calculado utilizando las coordenadas Biplot; la descripción general de los clusters se basa en cinco variables. Observamos en el gráfico (a) diferencias importantes entre clusters. El grupo 1 (color azul) indica la presencia de 25 juveniles de camarón *L. vannamei* alimentados con la mezcla M1 con mayor relación con la eficiencia alimenticia y la ganancia de peso, mientras que el grupo 2 (color verde) indica la presencia de 40 camarones juveniles *L. vannamei* alimentados con la mezcla M1 con mayor relación con la tasa de crecimiento específico y la relación de eficiencia proteica, y el Grupo 3 (color rojo) indica la presencia de 25 camarones juveniles *L.*

vannamei alimentados con la mezcla M1 con mayor relación con la supervivencia específica.

Por lo demás, en el gráfico (b) también hay diferencias entre los clusters. El grupo 1 (color rojo) indica la presencia de 19 camarones juveniles *L. vannamei* alimentados con la mezcla M2 con mayor relación con la eficiencia alimenticia, mientras que el grupo 2 (color azul) indica la presencia de 30 camarones juveniles *L. vannamei* alimentados con la mezcla M2 con mayor relación a todos los parámetros, y el Grupo 3 (color verde) indica la presencia de 41 camarones juveniles *L. vannamei* alimentados con la mezcla M2 con mayor relación a la tasa de crecimiento específico, ganancia de peso y índice de eficiencia proteica.

Las dietas que contenían más harina de pescado produjeron el mejor crecimiento, supervivencia y eficiencia proteica, lo que se puede atribuir a los parámetros deseables para los camarones de alimentación mencionados anteriormente, que son alta digestibilidad y atractivo, así como un perfil equilibrado de aminoácidos (Huang et al., 2017). Los camarones alimentados durante el día crecieron tan bien como los alimentados durante la noche y tuvieron una mejor eficiencia alimenticia y supervivencia que los alimentados durante la noche (Tacon et al., 2002). Los valores de la relación de eficiencia de proteína están relacionados con el nivel de proteína, y esto se atribuye al uso del exceso de proteína como fuente de energía en lugar de la formación de masa (Shahkar et al., 2014). El uso de aglutinantes con sardina en las dietas puede mejorar los rendimientos de crecimiento debido al aumento de la alimentación y al menor desperdicio de alimentos.

(**a**)



(**b**)

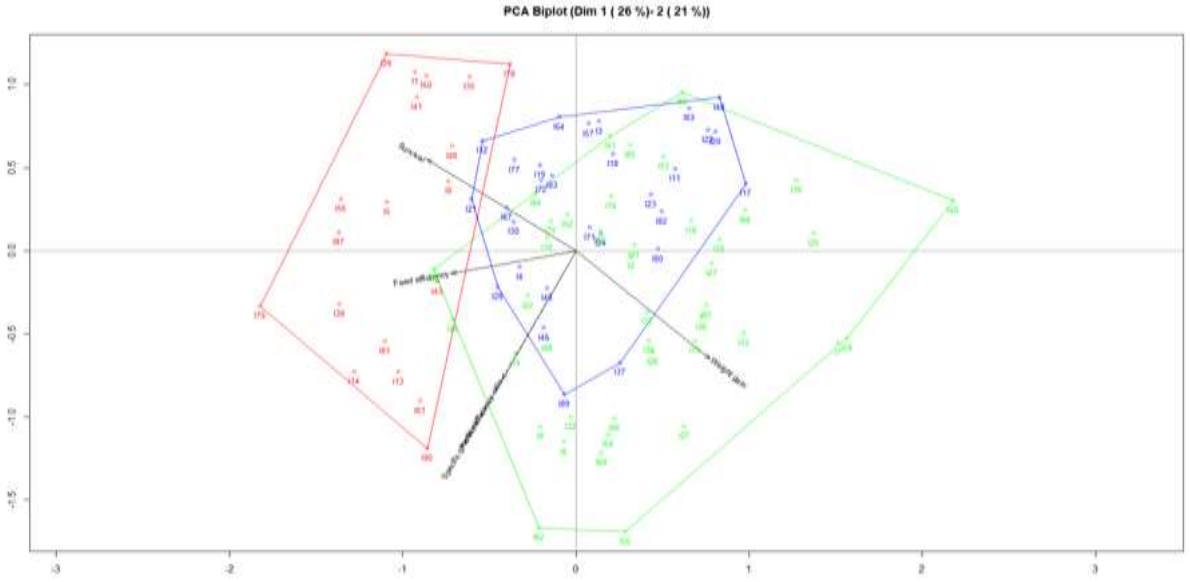


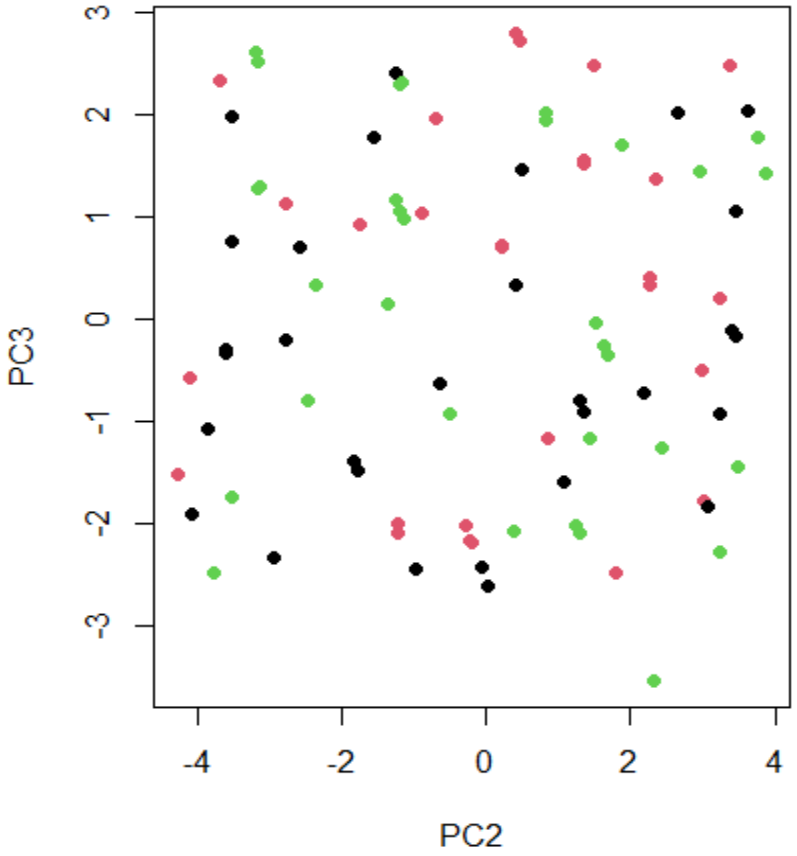
Figura 11. (a) PCA Biplot para rendimientos de crecimiento de camarones juveniles L. vannamei alimentados con la mezcla 1; (b) PCA Biplot para rendimientos de crecimiento de camarones juveniles L. vannamei alimentados con la mezcla 2.

5.2.2. Minería de datos para la composición nutricional de camarones juveniles

La Figura 13 presenta la aplicación del algoritmo de agrupamiento del método K-means a 90 objetos de 5 variables, cada uno utilizando el software RStudio. El gráfico (a) presentó el uso de tres grupos para la composición nutricional de camarones juveniles *L. vannamei* alimentados con la mezcla 1, mientras que en el gráfico (b) el uso de tres grupos para la composición nutricional de camarones juveniles de *L. vannamei* alimentados con la mezcla 2 también se mostró. Los resultados muestran la distribución normal de 90 puntos de datos alrededor de tres grupos en cada gráfico.

El tamaño de cada cluster está en relación con la cantidad de puntos de datos, gráfico (a): el tamaño del Cluster 1 (color rojo) es 31, el tamaño del Cluster 2 (color negro) es 30 y el tamaño del Cluster 3 (color verde) es de 29. Los camarones juveniles *L. vannamei* alimentados con la mezcla M1 pertenecientes al Cluster 1 mostraron los valores más altos de parámetros nutricionales. De lo contrario, en el gráfico (b): el tamaño del Cluster 1 (color negro) es 30, el tamaño del Cluster 2 (color verde) es 28 y el tamaño del Cluster 3 (color rojo) es 32. Los camarones juveniles *L. vannamei* alimentados con la mezcla M2 pertenecientes al Cluster 3 presentaron los valores más altos de parámetros nutricionales. Dado que los puntos de datos se distribuyen normalmente, los grupos varían en tamaño con puntos de datos máximos y puntos de datos mínimos.

(**a**)



(**b**)

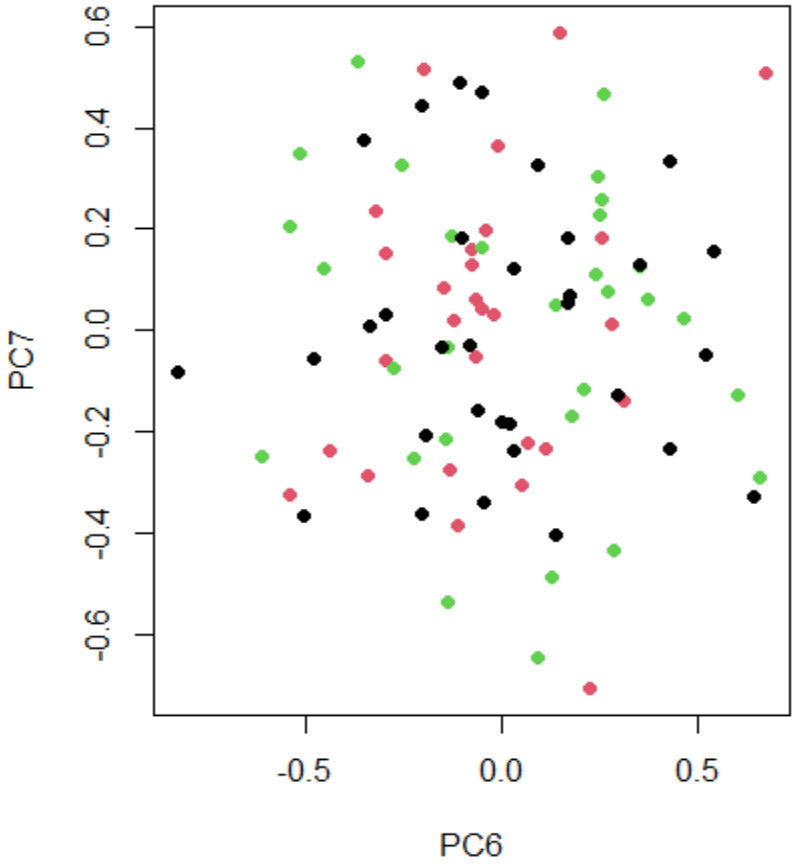


Figura 12. (a) Algoritmo de agrupamiento de K-means para la composición nutricional de camarones juveniles L. vannamei alimentados con la mezcla 1; (b) Algoritmo de agrupamiento de K-means para la composición nutricional de camarones juveniles L. vannamei alimentados con la mezcla 2.

La Figura 14 presenta el gráfico factorial del plano 1–2 (PCA-Biplot); el gráfico (a) presenta la inercia acumulada asciende al 56%, mientras que el gráfico (b) presenta la inercia acumulada asciende al 54,1% Además, los clusters se han calculado utilizando las coordenadas Biplot; la descripción general de los clusters se basa en cuatro variables.

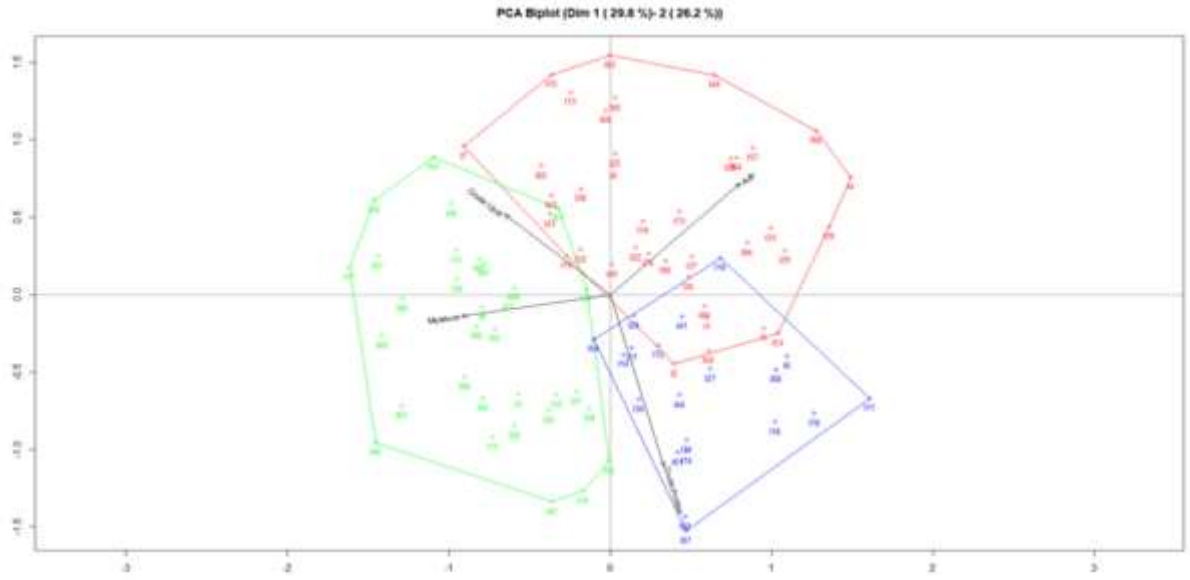
Observamos, en el gráfico (a) diferencias importantes entre conglomerados. El grupo 1 (color verde) indica la presencia de 35 camarones juveniles *L. vannamei* alimentados con la mezcla M1 con mayor relación de humedad y lípido crudo, mientras que el grupo 2 (color rojo) indica la presencia de 36 camarones juveniles *L. vannamei* alimentados con mezcla M1 con mayor relación a cenizas, y el Grupo 3 (color azul) indica la presencia de 19 camarones juveniles *L. vannamei* alimentados con la mezcla M1 con mayor relación a proteína cruda. Por otro lado, en el gráfico (b) también hay diferencias entre los conglomerados. El grupo 1 (color verde) indica la presencia de 23 camarones juveniles

L. vannamei alimentados con la mezcla M2 con mayor relación a proteína cruda y lípidos crudos, mientras que el grupo 2 (color azul) indica la presencia de 30 camarones juveniles *L. vannamei* alimentados con la mezcla M2 con mayor relación a cenizas, y el Grupo 3 (color rojo) indica la presencia de 37 camarones juveniles *L. vannamei* alimentados con la mezcla M2 con mayor relación a cenizas y humedad.

La composición de nutrientes de los camarones se ve afectada por la especie de camarón y la región de reproducción (Liu et al., 2021). Una buena fuente de proteínas para fines de nutrición animal es aquella con un contenido equilibrado de aminoácidos (Gil-Nuñez et al., 2020). El alto contenido de proteína cruda en los mariscos consiste en 70 a 80 % de fibronectina miogénica y 20 a 30 % de proteína sarcoplásmica. El contenido de ceniza refleja el contenido de compuestos inorgánicos en muestras biológicas, hasta cierto punto (Halim et al., 2016). En relación con los lípidos crudos, el hepatopáncreas es el almacén de lípidos, incluidos los triglicéridos y los fosfolípidos (Gulzar et al., 2020). Los camarones y los subproductos del camarón son los tipos de mariscos más consumidos debido a su valor nutricional (Nirmal et al., 2020). Los valores más altos de composición nutricional en los cuerpos de los camarones obtenidos con estas dietas pueden ayudar a mejorar la acuicultura en los camaroneros de pequeña escala.

El PCA Biplot depende del preprocesamiento de datos y la selección de variables y también utiliza la descomposición de valores singulares (SVD) de la matriz de datos (Ringner, 2008), mientras que el algoritmo K-means asigna cada objeto al grupo que tiene el centroide más cercano (Kodinariya et al., 2013). Los resultados indican que la minería de datos puede describir una buena visualización de las condiciones de alimentación con el objetivo de obtener parámetros comerciales específicos de camarones juveniles *Litopenaeus vannamei*, como el rendimiento de crecimiento o la composición nutricional.

(**a**)



(**b**)

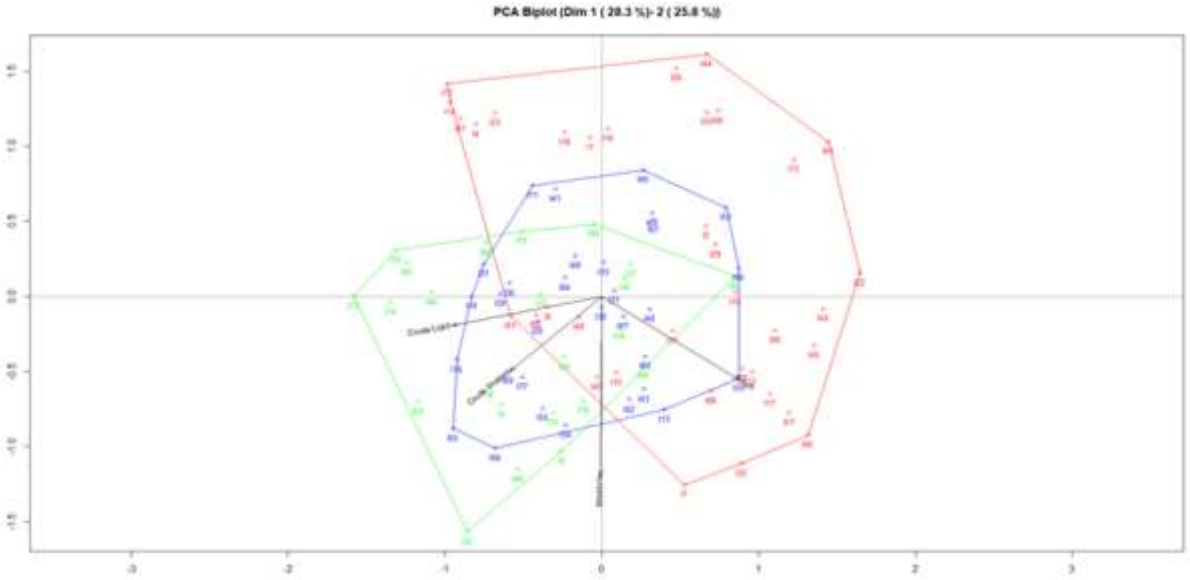


Figura 13. (a) PCA Biplot para la composición nutricional de camarones juveniles L. vannamei alimentados con la mezcla 1; (b) PCA Biplot para la composición nutricional de camarones juveniles L. vannamei alimentados con la mezcla 2.

Características de la Revista en que se publicó el artículo

Nombre de Revista: Sustainability

Nivel de Cuartil: JCR – Q1

Factor de Impacto: 3.251

sustainability

MDPI

Article

Data Mining Techniques: New Method to Identify the Effects of Aquaculture Binder with Sardine on Diets of Juvenile *Litopenaeus vannamei*

Fabricio Guevara-Viejó [1], Juan Diego Valenzuela-Cobos [1,2], Ana Grijalva-Endara [3], Purificación Vicente-Galindo [1,3,4,*] and Purificación Galindo-Villardón [1,3,5]

1 Centro de Gestión de Estudios Estadísticos, Universidad Estatal de Milagro (UNEMI), Milagro 091050, Ecuador; jguevarav@unemi.edu.ec (F.G.-V.); juan_diegova@hotmail.com (J.D.V.-C.); pgalindo@usal.es (P.G.-V.)
2 I+D+i Department, Ecuahidrolizados Industry, Guayaquil 090154, Ecuador
3 Department of Statistics, University of Salamanca, 37008 Salamanca, Spain; anagrijalvae@gmail.com
4 Institute for Biomedical Research of Salamanca (IBSAL), 37008 Salamanca, Spain
5 Centro de Investigación Institucional, Universidad Bernardo O'Higgins, Av. Viel 1497, Santiago 8370809, Chile
* Correspondence: purivg@usal.es; Tel.: +34-664-038-513

Abstract: In this research, a dataset of growth performances and nutritional composition of juvenile *Litopenaeus vannamei* after being fed two diets that include aquaculture binder with sardine for 7 weeks was analyzed using data mining techniques: the K-Means Clustering Algorithm and PCA Biplot, to have a visualization of each parameter (vector) measured. The parameters evaluated were: weight gain, specific growth rate, feed efficiency, protein efficiency ratio, survival percent, moisture content, crude protein, crude lipid, and ash content. Data mining tools showed the juvenile *Litopenaeus vannamei* fed with mixture 2 (pellets mixed with the binder of sardine subproducts) presented the highest growth performances and nutritional composition, 23 juvenile *L. vannamei* shrimps showed higher relation with crude protein and crude lipid, 30 *L. vannamei* shrimps presented higher relation with ash, and 37 juvenile *L. vannamei* shrimps showed higher relation with ash and moisture. The results obtained in experimental procedures indicate that the use of a binder of sardine subproducts in shrimp diets improves the commercial parameters, improving the aquaculture field.

Keywords: aquaculture; binder of sardine subproducts; *Litopenaeus vannamei*

check for updates

Citation: Guevara-Viejó, F.; Valenzuela-Cobos, J.D.; Grijalva-Endara, A.; Vicente-Galindo, P.; Galindo-Villardón, P. Data Mining Techniques: New Method to Identify the Effects of Aquaculture Binder with Sardine on Diets of Juvenile *Litopenaeus vannamei*. *Sustainability* **2022**, *14*, 4203. https://doi.org/10.3390/su14074203

Academic Editor: George P. Kraemer

Received: 19 February 2022
Accepted: 25 March 2022
Published: 1 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Ecuador is considered one of the most important countries for shrimp production. There is a growing interest in expanding aquaculture using alternative species and technologies [1,2]. Aquaculture is the second largest component of the Ecuadorian economy, after fossil fuels [3]. This expansion is almost entirely attributable to shrimp aquaculture and has led to land use or land cover transitions in Ecuadorian estuaries, with historic mangroves and other estuarine land to be used as shrimp ponds [4]. The highest production of shrimp in Ecuador is due two factors: Ecuador shrimp production has traditionally been semi-intensive, using feed and water exchange but no aeration, and there is much uninhabited land suitable for large ponds and farms in Ecuador [5]. By the end of 2009, the country had 175 thousand hectares of active shrimp farms representing 2578 aquaculture companies, with an export production of 450 million pounds representing 34% of the total manufactured products [6]. The increase in shrimp farming activity has also represented a great source of work in the country [7]. In order to diversify aquaculture production in Ecuador, several projects have been carried out for the production of shrimp, *Litopenaeus stylirostris* [8], *Sciaenops ocelatus* [9], and *Seriola rivoliana* [10]. Ecuador produced 510,000 metric tons of white shrimp (*Litopenaeus vannamei*) in 2018 [11].

In 2002, as a consequence of the white spot syndrome, alternative shrimp production methods were developed, such as culture in covered ponds that allowed reduced water exchange and a more constant temperature level and the so-called "onshore" system, which consisted of cultivating shrimp at very low salinities using water from wells and rivers in agricultural areas of the provinces of Manabí and Guayas. Supplementary feeding is associated with this activity, which influences the cost of shrimp production [12]. The strategy and the optimization of the feeding are aspects of importance in aquaculture that imply the formulation of different diets (pellets). The nutrient content present in pellets will influence shrimp growth, survival, and excreted waste products [13]. In the formulation of pellets, it is necessary to maintain the valuable dietary nutrients using binders [14]; binders affect pellet stability in three ways: by reducing voids, resulting in a more compact and durable pellet acting as adhesives, sticking particles together, and exerting a chemical action on the ingredients and altering the nature of the feed, obtaining a more durable pellet [15]. Binders are used to reduce the leaching of medication applied to balanced foods and drugs such as antibiotics, vitamins, organic acids. This type of product as a mixture of gluten in the diet can be used to obtain the highest values of apparent protein digestibility (ADP) and apparent dry matter digestibility (ADMD) [16]. The use of binders as attractants in shrimp feed is not common. In order to illustrate adequate visualization, it is necessary to use data mining tools, such as K-means clustering algorithm and PCA Biplot, to study the benefits of binders in the shrimps' diets.

Data mining refers to the process of extracting knowledge from databases by discovering anomalous situations, trends, patterns, and sequences in the data. Data mining is a stage within the complete knowledge discovery process that tries to obtain patterns or models from the collected data. The algorithms of data mining techniques usually have three components: (1) The model contains parameters to be set from the input data, (2) the preference criterion compares alternative models, and (3) the search algorithm is similar to other artificial intelligence programs [17]. There are various data mining techniques that describe the interesting relationship between different attributes, such as the K-Means Clustering Algorithm and PCA Biplot [18,19].

The main goal of this study was to use data mining techniques such as the K-Means Clustering Algorithm and PCA Biplot, identifying the growth performance and the nutritional composition of juvenile *Litopenaeus vannamei* shrimps after feeding diets that include an aquaculture binder with sardine.

2. Materials and Methods

2.1. Experimental Design

Ten juvenile *Litopenaeus vannamei* shrimps (0.70 g) were randomly distributed per tank (total 180 tanks of 20 L capacity). The experiment was realized at the Shrimp Farm "La Chorrera" [20,21].

2.2. Preparation of the Mixtures

The two mixtures used in the experiment were made using the following composition:

Mixture 1 (M1): Pellets mixed with the binder (hydrolyzed of sardine subproducts 20%). The formulation was 200 mL of sardine binder on 2 L of water for 25 kg of pellets.

Mixture 2 (M2): Pellets mixed with the binder (hydrolyzed of sardine subproducts 30%). The formulation was 200 mL of sardine binder on 2 L of water for 25 kg of pellets.

2.3. Experimental Diets and Feeding

Ten juvenile shrimps were randomly distributed per tank (180 tanks), and the juvenile shrimps were fed four times a day. In the feeders were put 200 g of mixture (M1) or 200 g of mixture (M2). The shrimps were fed four times a day at the following times: 09:00 a.m., 11:30 a.m., 2:00 p.m., and 4:30 p.m., and the faeces were removed every day. Tests were performed for seven consecutive weeks [22]. The experiments were realized in triplicate.

2.4. Growth Performances of Juvenile Shrimps

Three shrimps from each tank were counted, weighed, and measured (after seven weeks of the feeding trial) to determine: growth performance including weight gain (*WG* %), specific growth rate (*SGR* %), feed efficiency (*FE* %), protein efficiency ratio (*PER*), and survival percent (*S* %) using the following equations [23,24].

$$WG\ (\%) = \frac{final\ weight\ (g) - initial\ weight\ (g)}{initial\ weigth\ (g)} \times 100 \quad (1)$$

Equation (1). Weight gain of juvenile shrimps.

$$SGR\ (\%) = \frac{ln\ final\ weight\ (g) - ln\ initial\ weight\ (g)}{days} \quad (2)$$

Equation (2). Specific growth rate of juvenile shrimps.

$$FE\ (\%) = \frac{final\ weight\ (g) - initial\ weight\ (g)}{feed\ ration\ (g)} \times 100 \quad (3)$$

Equation (3). Feed efficiency of juvenile shrimps.

$$PER = \frac{wet\ weight\ gain\ (g)}{protein\ intake\ (g)} \quad (4)$$

Equation (4). Protein efficiency ratio of juvenile shrimps.

$$S\ (\%) = \frac{final\ number\ of\ fish}{initial\ number\ of\ fish} \times 100 \quad (5)$$

Equation (5). Survival percent of juvenile shrimps.

2.5. Nutritional Composition of Juvenile Shrimps

A proximate analysis of samples was performed according to the Association of Official Analytical Chemists (AOAC). The moisture was determined by drying shrimp samples at 100 °C to constant weight. Nitrogen (N) was determined using the Kjeldahl method, and the protein content was calculated using the 6.25 factor. The soxhlet method was used to determined crude lipid. The ash content was measured by heating the samples at 600 °C for 24 h [25–27]. All test analyses were realized in triplicate per experimental diet.

2.6. Statistical Analysis

The data-mining techniques (K-Means Clustering Algorithm and PCA Biplot) were realized using R software version 4.1.1. (R Core Team, Vienna, Austria) .

K-Means Clustering Algorithm

The K-Means algorithm consists of grouping a set of data (M blocks or sample vectors extracted from the training set) in groups or clusters (K quantization cells, such that $K < M + 1$), so that the vectors of the same group present high similarity to each other and have little similarity with vectors from other groups. This technique indicates that each training vector (block of samples from the original dataset) belongs to one and only one quantization cell.

Let $X = \{\sim x_j\}, j = 1, 2, \ldots, M + 1$ be a training set composed of M vectors N-dimensional, with $M \gg K$. The K-Means algorithm divides the vector space R^N assigned to each training vector to a single cluster via Nearest Neighbor Search (*VMP*). Precisely, $\vec{x_j}$ will belong to the group (cell of quantization) V ($\vec{w_i}$ if d ($\vec{x_j}$, $\vec{w_i}$) < d ($\vec{x_j}$, $\vec{w_a}$), ∀a ≠ i, where d ($\vec{x_j}$, $\vec{w_i}$) denotes

the quadratic Euclidean distance between $\vec{x_j}$ and $\vec{w_i}$. In this case, $\vec{w_i}$ is said to be the *VMP* of $\vec{x_j}$. The *VMP* search can be associated with a membership function, defined by

$$\mu_i\left(\vec{x_j}\right) = \begin{cases} 1, \; if \; \vec{w_i} = VMP\left(\vec{x_j}\right) \\ 0, \; otherwise \end{cases}$$

Thus, the distortion obtained by representing all the vectors of the training set by the respective *VMP*s is given by

$$J_1 = \sum_{i=1}^{K} \sum_{j=1}^{M} \mu_i\left(\vec{x_j}\right) d\left(\vec{x_j}, \vec{w_i}\right)$$

To minimize J_1, the vectors $\vec{w_i}$ are updated as follows:

$$\vec{w_i} = \frac{\sum_{j=1}^{M} \mu_i\left(\vec{x_j}\right) d\left(\vec{x_j}\right) \vec{x_j}}{\sum_{j=1}^{M} \mu_i\left(\vec{x_j}\right)}, \; i = 1, 2, \ldots, K$$

After initializing the $\vec{w_i}$ vector set, i = 1, 2, . . . , K, the K-Means algorithm can be summarized as follows:

1. Partitioning—the training set is partitioned into K clusters according to the *VMP* rule.
2. The new code vectors are the centroids of the clusters, calculated according to Equation $\vec{w_i}$.
3. Convergence test—algorithm stop criterion.

The partitioning and updating steps are carried out until the stop criteria is satisfied. Precisely, the algorithm stops at the end of the tth iteration if

$$\frac{J_1(t-1) - J_1(t)}{J_1(t)} \leq \in,$$

where $\in$ is an algorithm parameter, called distortion threshold, and $J_1(t)$ denotes a distortion obtained in the partitioning of the tth iteration [28].

Additionally, a PCA Biplot [29] was applied to explore and visualize the different parameters and the most relevant responses.

3. Results and Discussion

The focus of this work was to determine the viability of the use of aquaculture binder with sardine on diets of juvenile *L. vannamei* shrimps and to assess its influence on commercial parameters: weight gain, specific growth rate, feed efficiency, protein efficiency ratio, survival percent, moisture content, crude protein, crude lipid, and ash content.

The numeration of the juvenile *L. vannamei* shrimps per tank was made following the next distribution:

1–90: Average of three juvenile *L. vannamei* shrimps per tank fed with the mixture M1 or M2 after seven weeks.

3.1. Data Mining for Growth Performances of Juvenile Shrimps

Figure 1 presents the use of method K-means clustering algorithm to 90 objects having 5 variables each, using the software RStudio. The graphic (a) presents the use of three clusters to growth performances of juvenile *L. vannamei* shrimps fed with the mixture M1 after seven weeks, while in the graphic (b) the use of three clusters to growth performances of juvenile *L. vannamei* shrimps fed with the mixture M2 after seven weeks was also presented. The results show the normal distribution of 90 data points around three clusters in each graphic. The color of the different clusters showed the specific samples that presented the highest characteristic measured. Clusters allow separation of a set of objects

into none-overlapping subsets; the objects in the cluster are similar and dissimilar with the objects in the other cluster [30].

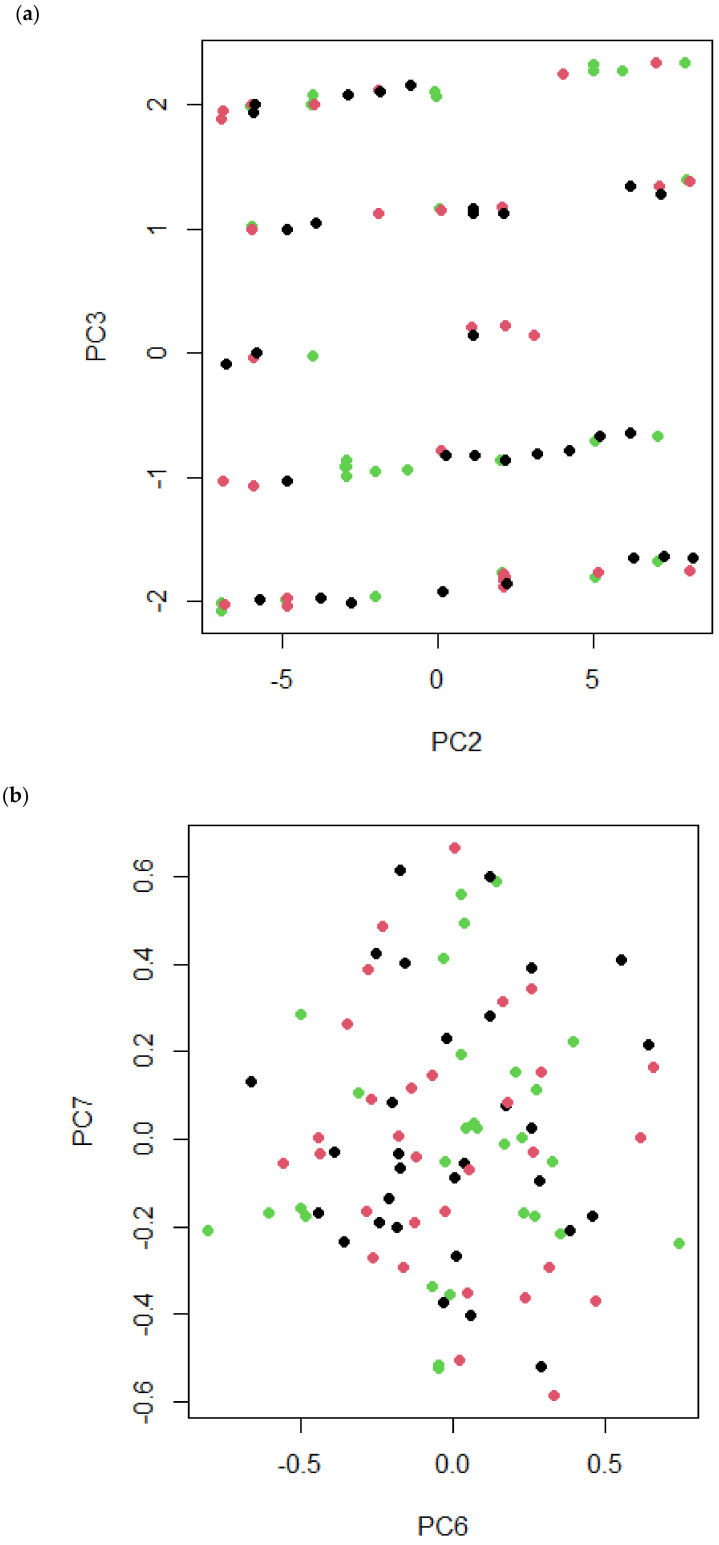


Figure 1. (**a**) K−Means Clustering Algorithm to growth performances of juvenile *L. vannamei* shrimps fed with mixture 1; (**b**) K−Means Clustering Algorithm to growth performances of juvenile *L. vannamei* shrimps fed with the mixture 2.

The size of each cluster has relation with the number of data points, the graphic (a): size of Cluster 1 (color red) is 28, the size of Cluster 2 (color black) is 32, and the size of Cluster 3 (color green) is 30. Juvenile *L. vannamei* shrimps fed with the mixture M1 belonging to Cluster 2 showed the highest values of growth performances. Otherwise, the graphic (b): size of Cluster 1 (color black) is 30, the size of Cluster 2 (color green) is 29, and the size of Cluster 3 (color red) is 31. Juvenile *L. vannamei* shrimps fed with the mixture M2 belonging to Cluster 3 showed the highest values of growth performances. Since the data points are normally distributed, clusters vary in size with maximum data points and minimum data points.

The aquaculture binder can be used to mix with other ingredients such as: antibiotics, vitamins, and organic acids to control bacterial infections and improve shrimp reproductive performance and egg hatching rate [31]. There is no beneficial effect of increasing the feeding frequency or ration size on the growth or survival of shrimp (*L. vannamei*) [32].

Figure 2 shows the factorial graph of the plane 1–2 (PCA-Biplot); graphic (a) presents the accumulated inertia amounts to 49.1%, while graphic (b) presents the accumulated inertia amounts to 47%.

(**a**)

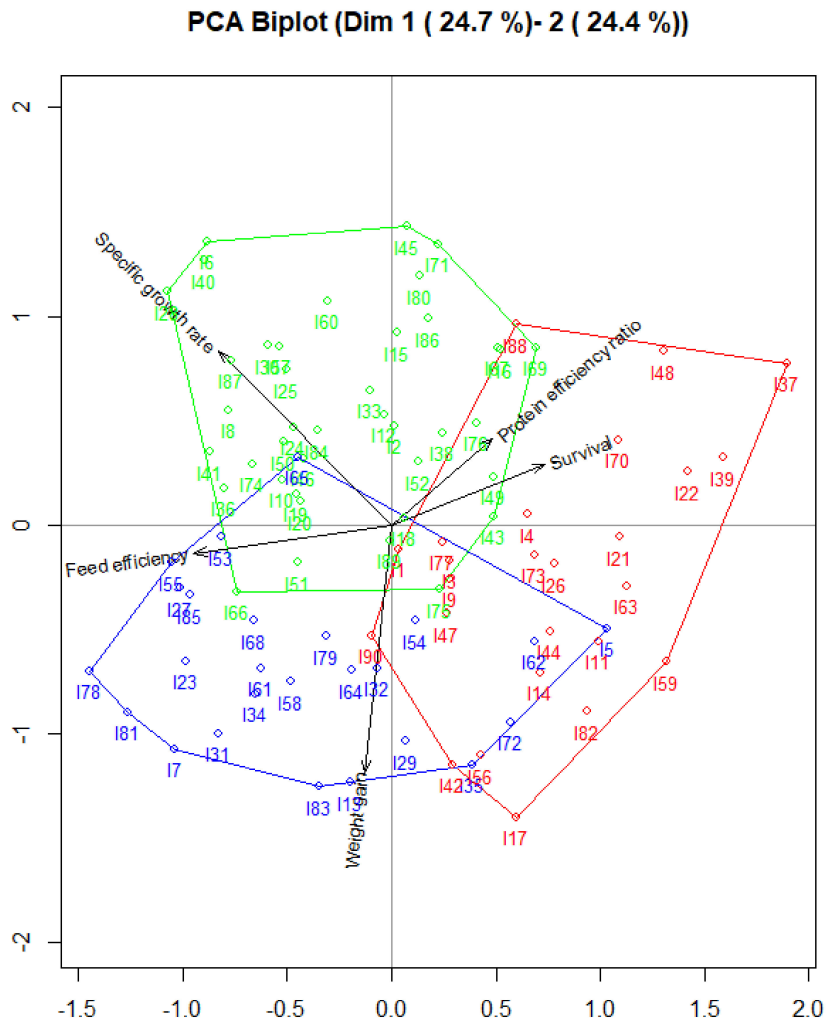


Figure 2. *Cont.*

(**b**)

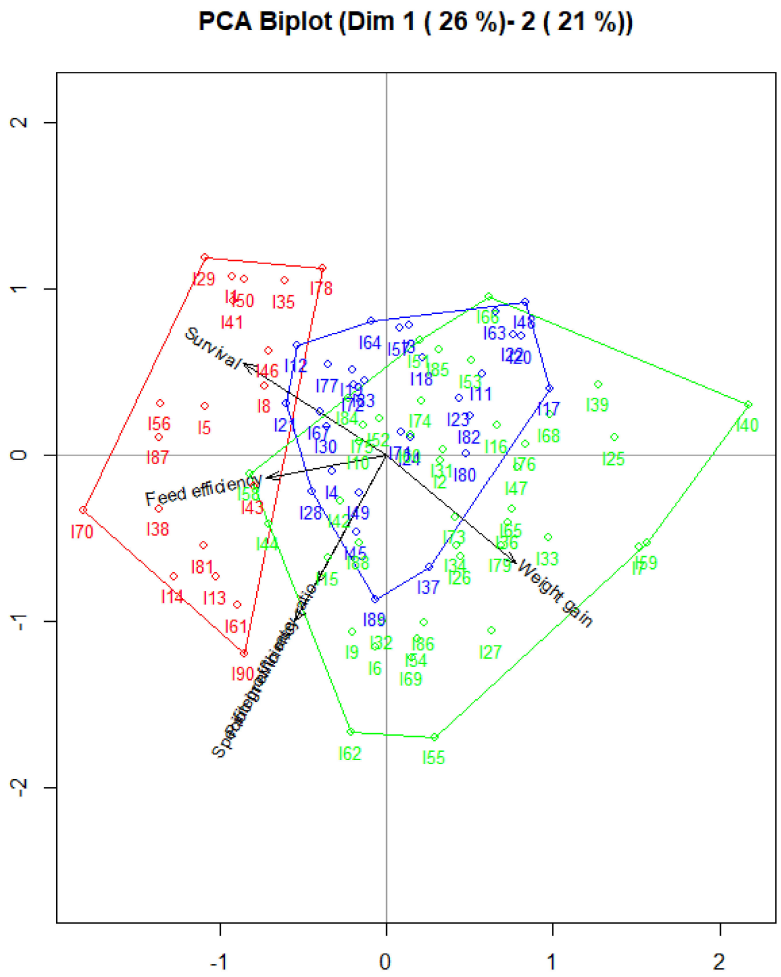


Figure 2. (**a**) PCA Biplot to growth performances of juvenile *L. vannamei* shrimps fed with the mixture 1; (**b**) PCA Biplot to growth performances of juvenile *L. vannamei* shrimps fed with the mixture 2.

In addition, clusters have been calculated using the Biplot coordinates; the overview of clusters is based on five variables. We observe in graphic (a) important differences between clusters. Cluster 1 (color blue) indicates the presence of 25 juvenile *L. vannamei* shrimps fed with the mixture M1 with higher relation to feed efficiency and weight gain, while Cluster 2 (color green) indicates the presence of 40 juvenile *L. vannamei* shrimps fed with the mixture M1 with higher relation to specific growth rate and protein efficiency ratio, and Cluster 3 (color red) indicates the presence of 25 juvenile *L. vannamei* shrimps fed with the mixture M1 with higher relation to specific survival.

Otherwise, in graphic (b) there are also differences between the clusters. Cluster 1 (color red) indicates the presence of 19 juvenile *L. vannamei* shrimps fed with the mixture M2 with higher relation to feed efficiency, whereas Cluster 2 (color blue) indicates the presence of 30 juvenile *L. vannamei* shrimps fed with the mixture M2 with higher relation to all the parameters, and Cluster 3 (color green) indicates the presence of 41 juvenile *L. vannamei* shrimps fed with the mixture M2 with higher relation to specific growth rate, weight gain, and protein efficiency ratio.

Diets that contained more fish meal produced the best growth, survival, and protein efficiency, which can be attributed to the desirable parameters for the feed shrimp above, which are high digestibility and attractiveness as well as a balanced amino acid profile [33]. The shrimp fed during the day grew as well as, and had better feed efficiency and survival than, those fed at night [34]. Protein efficiency ratio values are related with the protein level, and this is attributed to the use of protein excess as an energy source instead of

mass formation [35]. The use of binders with sardine in diets can improve the growth performances due to increasing the feeding and less food waste.

3.2. Data Mining for Nutritional Composition of Juvenile Shrimps

Figure 3 presents the application of method K-means clustering algorithm to 90 objects having 5 variables, each one using the software RStudio. Graphic (a) presented the use of three clusters for nutritional composition of juvenile *L. vannamei* shrimps fed with mixture 1, while in graphic (b) the use of three clusters for nutritional composition of juvenile *L. vannamei* shrimps fed with mixture 2 was also shown. The results show the normal distribution of 90 data points around three clusters in each graphic.

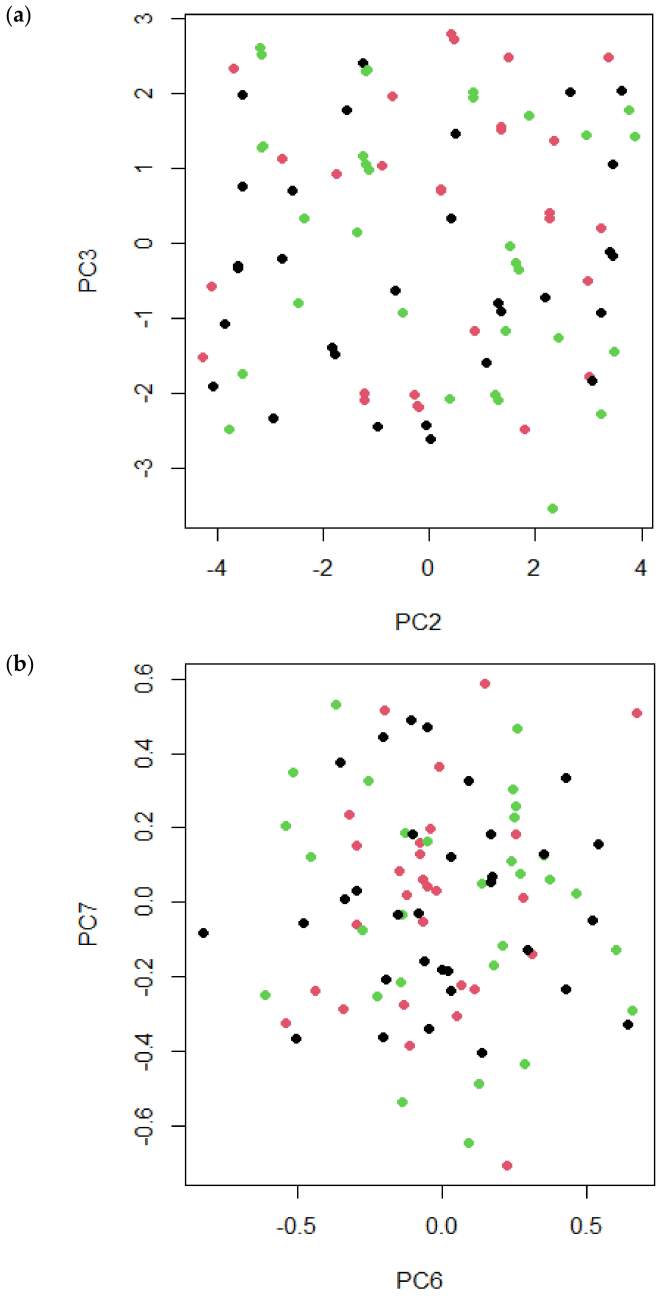


Figure 3. (**a**) K−Means Clustering Algorithm for nutritional composition of juvenile *L. vannamei* shrimps fed with mixture 1; (**b**) K−Means Clustering Algorithm for nutritional composition of juvenile *L. vannamei* shrimps fed with mixture 2.

The size of each cluster is in relation to the number of data points, graphic (a): size of Cluster 1 (color red) is 31, the size of Cluster 2 (color black) is 30, and the size of Cluster 3 (color green) is 29. Juvenile *L. vannamei* shrimps fed with mixture M1 belonging to Cluster 1 showed the highest values of nutritional parameters. Otherwise, in graphic (b): the size of Cluster 1 (color black) is 30, the size of Cluster 2 (color green) is 28, and the size of Cluster 3 (color red) is 32. Juvenile *L. vannamei* shrimps fed with mixture M2 belonging to Cluster 3 showed the highest values of nutritional parameters. Since the data points are normally distributed, clusters vary in size with maximum data points and minimum data points.

Figure 4 presents the factorial graph of the plane 1–2 (PCA-Biplot); graphic (a) presents the accumulated inertia amounts to 56%, while graphic (b) presents the accumulated inertia amounts to 54.1% In addition, clusters have been calculated using the Biplot coordinates; the overview of clusters is based on four variables.

We observe, in graphic (a), important differences between clusters. Cluster 1 (color green) indicates the presence of 35 juvenile *L. vannamei* shrimps fed with mixture M1 with higher relation of moisture and crude lipid, while Cluster 2 (color red) indicates the presence of 36 juvenile *L. vannamei* shrimps fed with mixture M1 with higher relation to ash, and Cluster 3 (color blue) indicates the presence of 19 juvenile *L. vannamei* shrimps fed with mixture M1 with higher relation to crude protein. On the other hand, in graphic (b) there are also differences between the clusters. Cluster 1 (color green) indicates the presence of 23 juvenile *L. vannamei* shrimps fed with mixture M2 with higher relation to crude protein and crude lipid, whereas Cluster 2 (color blue) indicates the presence of 30 juvenile *L. vannamei* shrimps fed with mixture M2 with higher relation to ash, and Cluster 3 (color red) indicates the presence of 37 juvenile *L. vannamei* shrimps fed with mixture M2 with higher relation to ash and moisture.

(**a**)

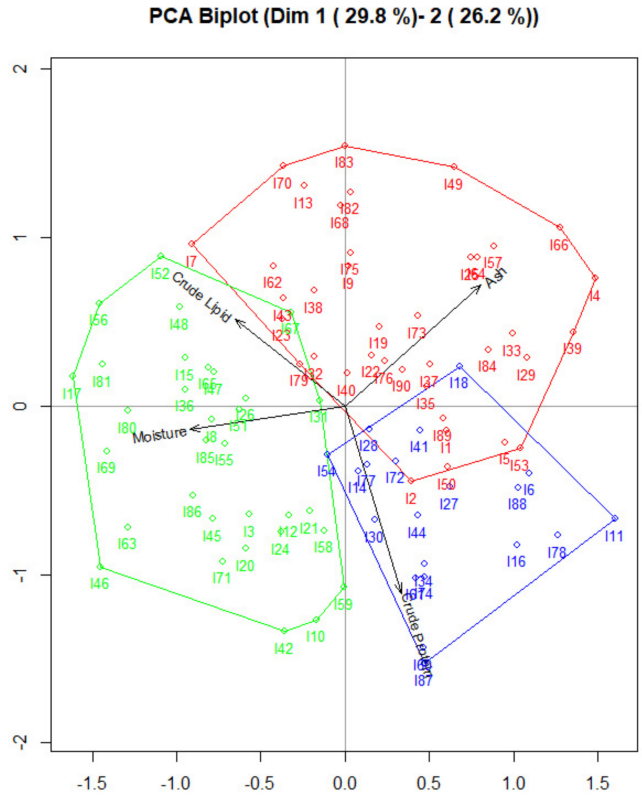


Figure 4. *Cont.*

(**b**)

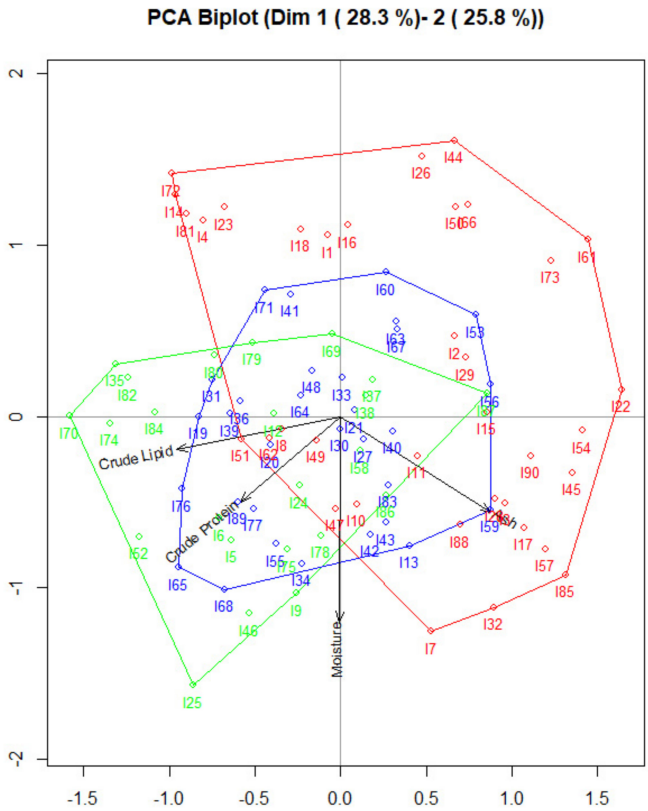


Figure 4. (**a**) PCA Biplot for nutritional composition of juvenile *L. vannamei* shrimps fed with mixture 1; (**b**) PCA Biplot for nutritional composition of juvenile *L. vannamei* shrimps fed with mixture 2.

The nutrient composition of shrimp is affected by shrimp species and breeding region [36]. A good protein source for animal nutrition purposes is that with a balanced content of amino acids [37]. The high crude protein in seafood consists of 70–80% myogenic fibronectin and 20–30% sarcoplasmic protein. The ash content reflects the content of inorganic compounds in biological samples, to a certain extent [38]. In relation to the crude lipid, the hepatopancreas is the storehouse of lipids, including triglycerides and phospholipids [39]. Shrimp and shrimp subproducts are the most consumed types of seafood because of their nutritional value [40]. The highest values of nutritional composition in the shrimps' bodies obtained with these diets can help to improve the aquaculture in small scale shrimp farmers.

The PCA Biplot depends on data preprocessing and variable selection and also uses singular value decomposition (SVD) of the data matrix [41], while the K-means algorithm assigns each object to the group that has the nearest centroid [42]. The results indicate that the data mining can describe a good visualization of the conditions of feeding with the objective to obtain specific commercial parameters of juvenile *Litopenaeus vannamei* shrimps, such as growth performance or nutritional composition.

4. Conclusions

Data mining tools such as PCA Biplot and K-means algorithm presented that juvenile *Litopenaeus vannamei* shrimps fed with mixture 2 presented the highest relation with specific growth rate, weight gain, protein efficiency ratio, crude protein, and crude lipid.

The use of data mining techniques on commercial parameters of juvenile *Litopenaeus vannamei* shrimps allows the conditions of feeding to be determined in order to obtain the highest values in specific parameters such as growth performance or nutritional composition.

The use of a binder with sardine allows a higher consumption of pellets to be obtained; a similar result was presented in other studies that used a mixture of pellet with a tuna binder, and the consumption was higher in comparison with only pellets.

Author Contributions: Conceptualization, F.G.-V. and J.D.V.-C.; formal analysis, J.D.V.-C. and A.G.-E.; investigation, F.G.-V.; methodology, F.G.-V. and J.D.V.-C.; supervision, P.G.-V. and P.V.-G.; writing—original draft, F.G.-V., J.D.V.-C., A.G.-E., P.G.-V. and P.V.-G.; writing—review and editing, P.G.-V. and P.V.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by Universidad Estatal de Milagro (UNEMI) Scholarship.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are grateful to Facultad de Ciencias e Ingeniería de la Universidad Estatal de Milagro (UNEMI) and Ecuahidrolizados Industry.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Naylor, R.L.; Goldberg, R.J.; Mooney, H.; Beveridge, M.; Clay, J.; Folke, C.; Kautsky, N.; Lubchenco, J.; Primavera, J.; Williams, M. Nature's subsidies to shrimp and salmon farming. *Science* **1998**, *282*, 883–884. [CrossRef]
2. Naylor, R.L.; Goldberg, R.J.; Primavera, J.H.; Kautsky, N.; Beveridge, M.C.M.; Clay, J.; Folke, C.; Lubchenco, J.; Mooney, H.; Troell, M. Effect of aquaculture on world fish supplies. *Nature* **2000**, *405*, 1017–1024. [CrossRef] [PubMed]
3. Naylor, R.L. Expanding the Boundaries of Agricultural Development. *Food Secur.* **2011**, *3*, 233–251. [CrossRef]
4. Hamilton, S.E.; Stankwitz, C. Examining the relationship between international aid and mangrove deforestation in coastal Ecuador from 1970 to 2006. *J. Land Use Sci.* **2012**, *7*, 177–202. [CrossRef]
5. Boyd, C.E.; Davis, R.P.; McNevin, A.A. Comparison of resource use for farmed shrimp in Ecuador, India, Indonesia, Thailand, and Vietnam. *Aquac. Fish Fish.* **2021**, *1*, 3–15. [CrossRef]
6. PROECUADOR. *Instituto de Promoción de Exportaciones e Inversiones. El Camarón Congelado es el Tercer Producto de Exportación*; PROECUADOR: Quito, Ecuador, 2013.
7. Carrillo, D. *La Industria de Alimentos y Bebidas en el Ecuador*; Instituto Nacional de Estadística y Censos: Loja, Ecuador, 2009.
8. Rivera, L.M.; Trujillo, L.E.; Pais-Chanfrau, J.M.; Núñez, J.; Pineda, J.; Romero, H.; Tinococo, O.; Cabrera, C.; Dimitrov, V. Functional foods as stimulators of the immune system of *Litopenaeus vannamei* cultivated in Machala, Province of El Oro, Ecuador. *Ital. J. Food Sci.* **2018**, *31*, 227–232.
9. Guartatanga, R.; Schwartz, L.; Wigglesworth, J.M.; Griffith, D.R.W. *Experimental Intensive Rearing of Red Drum (Sciaenops ocellatus) in Ecuador*; CENAIM: San Pedro de Manglaralto, Ecuador, 1993.
10. Blacio, E.; Darquea, J.; Rodríguez, S. Avances en el cultivo de Huayaipe, Seriola rivoliana (Valenciennes 1833), en las instalaciones del CENAIM. *Mundo Acuic.* **2003**, *9*, 23–24.
11. Boyd, C.E.; Davis, R.P.; Wilson, A.G.; Marcillo, F.; Brian, S.; McNevin, A.A. Resource use in whiteleg shrimp *Litopenaeus vannamei* farming in Ecuador. *J. World Aquac. Soc.* **2001**, *52*, 772–788. [CrossRef]
12. Lawrence, A.L.; Lee, P.G. Research in the Americas. In *Crustacean Nutrition. Advances in World Aquaculture*; D'Abramo, L.R., Conklin, D.E., Akiyama, D.M., Eds.; World Aquaculture Society: Baton Rouge, LA, USA, 1997; Volume 6, pp. 566–587.
13. Smith, D.M.; Burford, M.A.; Tabrett, S.J.; Irvin, S.J.; Ward, L. The effect of feeding frequency on water quality and growth of the black tiger shrimp (*Penaeus monodon*). *Aquaculture* **2002**, *207*, 125–136. [CrossRef]
14. Partridge, G.J.; Southgate, P.C. The effect of binder composition on ingestion and assimilation of microbound diets MBD by barramundi Lates calcarifer Bloch larvae. *Aquac. Res.* **1999**, *30*, 879–886. [CrossRef]
15. Palma, J.; Bureau, D.P.; Andrade, J.P. Effects of binder type and binder addition on the growth of juvenile *Palaemonetes varians* and *Palaemon elegans* (*Crustacea: Palaemonidae*). *Aquac. Int.* **2008**, *16*, 427–436. [CrossRef]
16. Argüello-Guevara, W.; Molina-Poveda, C. Effect of binder type and concentrationon prepared feed stability, feed ingestionand digestibility of *Litopenaeus vannamei* broodstock diets. *Aquac. Nutr.* **2013**, *19*, 515–522. [CrossRef]
17. Valcárcel, V. Data Mining y el Descubrimiento del Conocimiento. *Ind. Data* **2004**, *7*, 83–86. [CrossRef]
18. Guevara-Viejó, F.; Valenzuela-Cobos, J.D.; Vicente-Galindo, P.; Galindo-Villardón, P. Application of K-Means Clustering Algorithm to Commercial Parameters of *Pleurotus* spp. Cultivated on Representative Agricultural Wastes from Province of Guayas. *J. Fungi* **2021**, *7*, 537. [CrossRef] [PubMed]
19. Guevara-Viejó, F.; Valenzuela-Cobos, J.D.; Vicente-Galindo, P.; Galindo-Villardón, P. Data-Mining Techniques: A New Approach to Identifying the Links among Hybrid Strains of *Pleurotus* with Culture Media. *J. Fungi* **2021**, *7*, 882. [CrossRef]

20. Xie, S.W.; Li, Y.T.; Zhou, W.W.; Tian, L.X.; Li, Y.M.; Zeng, S.L.; Liu, Y.J. Effect of γ-aminobutyric acid supplementation on growth performance, endocrine hormone and stress tolerance of juvenile Pacific white shrimp, *Litopenaeus vannamei*, fed low fishmeal diet. *Aquac Nutr.* **2017**, *23*, 54–62. [CrossRef]
21. Arambul Munoz, E.; Ponce Palafox, J.; De Los Santos, R.; Aragon Noriega, E.; Rodriguez Dominguez, G.; Castillo Vargasmachuca, S. Influence of Stocking Density on Production and Water Quality of a Photo Heterotrophic Intensive System of White Shrimp (*Penaeus vannamei*) in Circular Lined Grow out Ponds, with Minimal Water Replacement. *Lat. Am. J. Aquat. Res.* **2019**, *47*, 449–455. [CrossRef]
22. Mmanda, F.P.; Lindberg, J.E.; Halden, A.N.; Mtolera, M.S.P.; Kitula, R.; Lundh, T. Digestibility of local feed ingredients in tilapia *Oreochromis niloticus* juveniles, determined on faeces collected by siphoning or stripping. *Fishes* **2020**, *5*, 32. [CrossRef]
23. Mohanty, R.K. Growth performance of *Penaeus monodon* at different stocking densities. *J. Inland Fish. Soc. India* **1999**, *31*, 53–59.
24. Bae, J.; Hamidoghli, A.; Djaballah, M.S.; Maamri, S.; Hamdi, A.; Souffi, I.; Farris, N.W.; Bai, S.C. Effects of three different dietary plant protein sources as fishmeal replacers in juvenile whiteleg shrimp, *Litopenaeus vannamei*. *Fish. Aquat. Sci.* **2020**, *23*, 2. [CrossRef]
25. Valencia del Toro, G.; Ramírez-Ortiz, M.E.; Flores-Ramírez, G.; Costa-Manzano, M.R.; Robles-Martínez, F.; Garín Aguilar, M.E.; Leal-Lara, H. Effect of *Yucca schidigera* bagasse as substrate for Oyster mushroom on cultivation parameters and fruit body quality. *Rev. Mex. Ing. Quim.* **2018**, *17*, 835–846. [CrossRef]
26. Valenzuela-Cobos, J.D.; Vásquez-Véliz, G.; Zied, D.C.; Franco-Hernández, O.M.; SánchezHernández, A.; Garín Aguilar, M.E.; Leal Lara, H.; Valencia del Toro, G. Bioconversion of agricultural wastes using parental, hybrid and reconstituted strains of *Pleurotus* and *Lentinula*. *Rev. Mex. Ing. Quim.* **2019**, *18*, 647–657. [CrossRef]
27. Association of Official Analytical Chemists (AOAC). *International Official Methods of Analysis*; AOAC: Washington, DC, USA, 2002.
28. Madeiro, F.; Galvão, R.R.A.; Ferreira, F.A.B.S.; Cunha, D.C. Uma alternativa de aceleração do algoritmo fuzzy k-means aplicado à quantização vetorial. *TEMA Tend. Mat. Appl. Comput.* **2012**, *13*, 193–206. [CrossRef]
29. Pasqualoto, K.F.; Teófilo, R.F.; Guterres, M.; Pereira, F.S.; Ferreira, M. A study of physicochemical and biopharmaceutical properties of Amoxicillin tablets using full factorial design and PCA biplot. *Anal. Chim. Acta* **2007**, *595*, 216–220. [CrossRef] [PubMed]
30. Razavi Zadegan, S.M.; Mirzaie, M.; Sadoughi, F. Ranked k-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets. *Knowl.-Based Syst.* **2013**, *39*, 133–143. [CrossRef]
31. Valenzuela-Cobos, J.D.; Vargas-Farias, C.J. Study about the use of aquaculture binder with tuna attractant in the feeding of white shrimp (*Litopenaeus vannamei*). *Rev. Mex. Ing. Quim.* **2020**, *19*, 355–361. [CrossRef]
32. Velasco, M.; Lawrence, A.L.; Castille, F.L. Effect of variations in daily feeding frequency and ration size on growth of shrimp, *Litopenaeus vannamei* (Boone), in zero-water exchange culture tanks. *Aquaculture* **1999**, *179*, 141–148. [CrossRef]
33. Huang, F.; Wang, L.; Zhang, C.; Song, K. Replacement of fishmeal with soybean meal and mineral supplements in diets of *Litopenaeus vannamei* reared in low-salinity water. *Aquaculture* **2017**, *473*, 172–180. [CrossRef]
34. Tacon, A.G.J.; Cody, J.J.; Conquest, L.D.; Divakaran, S.; Forster, I.P.; Decamp, O.E. Effect of culture system on the nutrition and growth performance of Pacific white shrimp *Litopenaeus vannamei* (Boone) fed different diets. *Aquac. Nutr.* **2002**, *8*, 121–139. [CrossRef]
35. Shahkar, E.; Yun, H.; Park, G.; Jang, I.K.; Kyoung Kim, S.; Katya, K.; Bai, S.C. Evaluation of optimum dietary protein level for juvenile whiteleg shrimp (*Litopenaeus vannamei*). *J. Crust. Biol.* **2014**, *34*, 552–558. [CrossRef]
36. Liu, Z.; Liu, Q.; Zhang, D.; Wei, S.; Sun, Q.; Xia, Q.; Shi, W.; Ji, H.; Liu, S. Comparison of the Proximate Composition and Nutritional Profile of Byproducts and Edible Parts of Five Species of Shrimp. *Foods* **2021**, *10*, 2603. [CrossRef] [PubMed]
37. Gil-Núñez, J.C.; Martínez-Córdova, L.R.; Servín-Villegas, R.S.; Magallon-Barajas, F.J.; Bórques-López, R.A.; Gonzalez-Galaviz, J.R.; Casillas-Hernández, R. Production of *Penaeus vannamei* in low salinity, using diets formulated with different protein sources and percentages. *Lat. Am. J. Aquat. Res.* **2020**, *48*, 396–405. [CrossRef]
38. Halim, N.R.; Yusof, H.M.; Sarbon, N.M. Functional and bioactive properties of fish protein hydolysates and peptides: A comprehensive review. *Trends Food Sci. Technol.* **2016**, *51*, 24–33. [CrossRef]
39. Gulzar, S.; Raju, N.; Nagarajarao, R.C.; Benjakul, S. Oil and pigments from shrimp processing by-products: Extraction, composition, bioactivities and its application—A review. *Trends Food Sci. Technol.* **2020**, *100*, 307–319. [CrossRef]
40. Nirmal, N.P.; Santivarangkna, C.; Rajput, M.S.; Benjakul, S. Trends in shrimp processing waste utilization: An industrial prospective. *Trends Food Sci. Technol.* **2020**, *103*, 20–35. [CrossRef]
41. Ringner, M. What is principal component analysis? *Nat. Biotechnol.* **2008**, *26*, 303–304. [CrossRef] [PubMed]
42. Kodinariya, T.M.; Makwana, P.R. Review on determining number of cluster in K-means clustering. *Int. J. Adv. Res. Comput. Sci. Manag. Stud.* **2013**, *1*, 90–95.

CONCLUSIONES

CONCLUSIONES

Siguiendo los resultados obtenidos en este trabajo se puede concluir que:

1. El algoritmo de agrupamiento de K-medoids fue utilizado para determinar la presencia de cuatro clusters obteniendo una visualización precisa de los parámetros de especies biológicas de hongos comestibles cultivados en diferentes medios de cultivo suplementados con productos agrícolas ecuatorianos, las cepas cultivadas en cultivo sólido (M1) y el cultivo líquido (L1) mostraron las más altas características miceliales y culturales.

2. PCA Biplot presentó las cepas de *Pleurotus ostreatus* y *Pleurotus djamor* cultivadas en los diferentes medios de cultivo (sólido y líquido) con mayor relación con la característica medida.

3. El algoritmo de reglas de asociación mostró el grupo de cepas de *Pleurotus ostreatus* y *Pleurotus djamor* crecimiento en cultivo sólido (M1) y el cultivo líquido (L1) con las más altas características miceliales y culturales.

4. El uso del algoritmo de K-means en los parámetros comerciales de hongos comestibles *Pleurotus ostreatus* y *Pleurotus djamor* cultivados en dos mezclas de cultivos agrícolas residuos, permitió elegir la cepa cultivada y el sustrato para obtener los mayores parámetros comerciales.

5. PCA Biplot presentó que el uso de la mezcla 1 (S1) en el cultivo de las cepas de los hongos comestibles *Pleurotus ostreatus* y *Pleurotus djamor* tienen una mayor relación con los parámetros de productividad: eficiencias biológicas, rendimiento de los cultivos y tasas de productividad.

6. Las técnicas de minería de datos como PCA Biplot y el algoritmo K-means presentaron que los camarones juveniles *Litopenaeus vannamei* alimentados con la mezcla 2 presentaron la relación más alta con la tasa de crecimiento específico, la ganancia de peso, la relación de eficiencia de proteína, el contenido de proteína cruda y de grasa.

7. El uso de técnicas de minería de datos sobre parámetros comerciales de camarones juveniles *Litopenaeus vannamei* permite determinar las condiciones de alimentación para obtener los valores más altos en parámetros específicos como el crecimiento o la composición nutricional.

RECOMENDACIONES

RECOMENDACIONES

Con la realización de esta tesis, se hacen las siguientes recomendaciones para dar continuidada este tipo de estudios:

1. Con los resultados obtenidos, animamos a que se continúe el interés de la evaluación de técnicas de estadística multivariante, para lograr una correcta elección de especies biológicas que presenten los mayores parámetros comerciales.

2. El Ecuador es un país cuya agricultura y acuicultura tienen una amplia importancia en el sector económico, por lo que es indispensable utilizar técnicas de minería de datos para predecir comportamientos de estos sectores basados en sus productos.

BIBLIOGRAFÍA

BILIOGRAFÍA

Abdel-Basset, M., Mohamed, M., Smarandache, F. & Chang, V. (2018). Neutrosophic association rule mining algorithm for big data analysis. *Symmetry*, *10*, 106.

Abrar, A.S., Kadam, J.A., Mane, V.P., Patil, S.S. & Baig, M.M.V. (2009). Biological efficiency and nutritional contents of *Pleurotus florida* (Mont.) singer cultivated on different agro-wastes. *Natural Sciences*, *7*, 1545–1740

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association between sets of items in massive database. *International proceedings of the ACM-SIGMOD international conference on management of data (pp. 207–216).*

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the international conference on very large data bases* (pp. 407–419).

Arambul Munoz, E., Ponce Palafox, J., De Los Santos, R., Aragon Noriega, E., Rodriguez Dominguez, G. & Castillo Vargasmachuca, S. (2019). Influence of Stocking Density on Production and Water Quality of a Photo Heterotrophic Intensive System of White Shrimp (*Penaeus vannamei*) in Circular Lined Grow out Ponds, with Minimal Water Replacement. *Latin American Journal of Aquatic Research*, *47*, 449–455.

Arana-Gabriel, Y., Burrola-Aguilar, C., Garibay-Orijel, R. & Franco-Maass, S. (2014). Obtención de cepas y producción de inóculo de cinco especies de hongos silvestres comestibles de alta montaña en el centro de México. *Revista Chapingo serie ciencias forestales y del ambiente, 20*, 213–226.

Argüello-Guevara, W. & Molina-Poveda, C. (2013). Effect of binder type and concentrationon prepared feed stability, feed ingestionand digestibility of *Litopenaeus vannamei* broodstock diets. *Aquaculture Nutrition*, 19, 515–522.

Arthur, D. & Vassilvitskii, S. k-means++: The advantages of careful seeding. In Symposium on Discrete Algorithms (SODA), pp. 1027–1035. SIAM, 2007.

AOAC. (2002). International Official Methods of Analysis; Association of OfficialAnalytical Chemists AOAC: Washington DC, USA.

AOAC. (2016). Official Methods of Analysis of AOAC International, 20th edn.

Atkinson, A.C. & Riani, M. (2007). Exploratory tools for clustering multivariate data. *Computational Statistics and Data Analysis*, *52*, 272–285.

Bae, J., Hamidoghli, A., Djaballah, M.S., Maamri, S., Hamdi, A., Souffi, I., Farris, N.W., Bai, S.C. (2020). Effects of three different dietary plant protein sources as fishmeal replacers in juvenile whiteleg shrimp, *Litopenaeus vannamei*. *Fisheries and Aquatic Sciences*, *23*, 2.

Bakir, T., Karadeniz, M. & Unal, S. (2018). Investigation of antioxidant activities of *Pleurotus ostreatus* stored at different temperatures. *Food Science & Nutrition*, *6*, 1040-1044.

Barr, A., & Feigenbaum, E. A. (Eds.). (1981). *The handbook of artificial intelligence*. Los Altos, CA: Morgan Kaufmann.

Baty, F., & Delignette-Muller, M.L. (2004). Estimating the bacterial lag time: which model,which precision?. *International Journal of Food Microbiology*, *91*, 261-277.

Bezdek, J.C., Ehrilich, R. & Fill, W. (1984). The Fuzzy C-means Clustering Algorithm. *Computers and Geosciences*, *10*, 191-203.

Bishop, C.M. (2006) Pattern recognition and machine learning. Springer, New York.

Blacio, E., Darquea, J. & Rodríguez, S. (2003). Avances en el cultivo de Huayaipe, Seriola rivoliana (Valenciennes 1833), en las instalaciones del CENAIM. *Mundo Acuícola*, 9, 23–24.

Bock, H-H. (1996). Probabilistic models in cluster analysis. *Computational Statistics and Data Analysis*, *23*, 5–28

Boyd, C.E., Davis, R.P., Wilson, A.G., Marcillo, F., Brian, S. & McNevin, A.A. (2001). Resource use in whiteleg shrimp *Litopenaeus vannamei* farming in Ecuador. *Journal of the World Aquaculture Society*, 52, 772–788.

Boyd, C.E.; Davis, R.P. & McNevin, A.A. (2021). Comparison of resource use for farmed shrimp in Ecuador, India, Indonesia, Thailand, and Vietnam. *Aquaculture, Fish and Fisheries*, 1, 3–15.

Brudzewski, K., Osowski, S. & Markiewicz, T. (2004) Classification of milk by means of an electronic nose and SVM neural network. *Sensors and*

Actuators B, 291–298.

Bucheli, H., & Thompson, W. (2014). Statistics and Machine Learning at Scale: New Technologies Apply Machine Learning to Big Data. Insights From the SAS Analytics 2014 Conference.

Camps-Valls, G., Gomez-Chova, L., Calpe-Maravilla, J., Soria-Olivas, E., Martin-Guerrero, J.D. & Moreno, J. (2003). Support vector machines for crop classification using hyperspectral data. *Lecture Notes in Computer Science, 2652,* 134–141.

Cárdenas, O., Galindo, M.P. & Vicente-Villardón, J.L (2007). Los métodos Biplot: evolución y aplicaciones. *Revista Venezolana de Análisis de Coyuntura*, *13*, 279-303.

Cardoso, R.V.C., Carocho, M., Fernandes, Â., Zied, D.C., Cobos, J.D.V., González-Paramás, A.M., Ferreira, I.C.F.R. & Barros, L. (2020). Influence of Calcium Silicate on the Chemical Properties of *Pleurotus ostreatus* var. florida (Jacq.) P. Kumm. *Journal of Fungi*, *6*, 299.

Cardoso, R.V.C., Carocho, M., Fernandes, Â., Pinela, J., Stojkovic, D., Sokovic, M., Zied, D.C., Cobos, J.D.V., González-Paramás, A.M., Ferreira, I.C.F.R. & Barros, L. (2021). Antioxidant and Antimicrobial Influence on Oyster Mushrooms (*Pleurotus ostreatus*) from Substrate Supplementation of Calcium Silicate. *Sustainability*, *13*, 5019.

Carrillo, D. (2009). La Industria de Alimentos y Bebidas en el Ecuador; Instituto Nacional de Estadística y Censos: Ecuador.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining worldwide web browsing patterns. *Journal of Knowledge and Information Systems*, *1*, 5–32.

Chang, S.T. & Miles, P.G. (2008). Mushrooms: Cultivation, Nutritional Value, Medicinal Effect, and Environmental Impact. 2nd. Boca Raton, Fla, USA: CRC Press.

Chaturvedi, A., Green, P.E. & Carroll, J.D. (2001). K-modes clustering. *Journal of Classification*, *18*, 35–55.

Chegwin, C. & Nieto, I.J. (2013). Influencia del medio de cultivo en la producción de metabolitos secundarios del hongo comestible *Pleurotus ostreatus* cultivados por fermentación en estado líquido empleando harinas de cereal como fuente de carbono. *Revista Mexicana Micología*,

37, 1–9.

Choi, D.H., Ahn, B.S. & Kim, S.H. (2005). Prioritization of association rules in data mining: Multiple criteria decision approach. *Expert Systems With Applications*, *29*, 867–878.

Crisan, E.V. & Sands, A. (1978). Nutritional value. Academic Press. New York, 137-168.

Das, K.C. & Evans, M.D. (1992). Detecting fertility of hatching eggs using machine vision II: neural network classifiers. *Trans ASAE*, *35*, 2035–2041.

Da Silva, M.C.S., Naozuka, J., da Luz, J.M.R., de Assunção, L.S., Oliveira, P.V., Vanetti, M.C.D., Bazzolli, D.M.S. & Kasuya, M.C.M. (2012). Enrichment of Pleurotus ostreatus mushrooms with selenium in coffee husks. *Food Chemistry*, *131*, 558–563.

Delgado, A. (2013). Guayaquil. *Cities*, *31*, 515-532.

Dempster, A.P., Laird, N.M. & Rubin, R.D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, *39*, 1–38.

Díaz-Talamantes, C., Burrola-Aguilar, C., Aguilar-Miguel, X. & Mata, G. (2017). In vitro mycelial growth of wild edible mushrooms from the central Mexican highlands. *Revista Chapingo serie ciencias forestales y del ambiente*, *23*, 3.

Dhillon, I.S., Guan, Y. & Kulis, B. (2004). Kernel k-means: Spectral clustering and normalized cuts. In: Proc. 10th KDD, pp. 551–556.

Du, C.-J. & Sun, D.-W. (2005) Pizza sauce spread classification using colour vision and support vector machines. *Journal of Food Engineering*, *66*, 137–145.

Dunn, J.C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybern. 3*, 32–57.

Economou, C.N., Diamantopoulou, P.A. & Philippoussis, A.N. (2017). Valorization of spent oyster mushroom substrate and laccase recovery through successive solid state cultivation of *Pleurotus, Ganoderma*, and *Lentinula* strains. *Applied Microbiology and Biotechnology*, *101*, 5213–5222.

Ester, M., Kriegel, H.P., Sander, J. & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In: Paper

presented at International conference on knowledge discovery and data mining.

Fagerlund, S. (2007). Bird species recognition using support vector machines. EURASIP J Adv Signal Processing, Article ID 38637, p 8.

Farr, D.F. (1983). Mushroom industry: diversification with additional species in the United States. *Mycology*, *75*, 351-360.

Febles Rodríguez, J.P. & González Pérez, A. (2002). Aplicación de la minería de datos en la bioinformática. *ACIMED*, *10*, 69-76.

Flury, B. (1997). A first course in multivariate statistics. Springer-Verlag, New York.

Forgy, E.W. (1965). Cluster Analysis of Multivariate Data: Efficiency vs Interpretability of Classifications. *Biometrics*, *21*, 768-780.

Fultz, S.A. (1998). Fruiting at high temperature and its genetic control in the basidiomycete *Flammulina velutipes*. *Applied and Environmental Microbiology*, *54*, 2460-2463.

Gabriel, K.R. (1971). The biplot-graphical display of matrices with applications to principal component analysis. *Biometrica*, *58*, 453-67.

Galindo, M.P. (1986) Una alternativa de representación simultánea: HJ-Biplot. *Qüestiió*, *10*, 13–23.

García-Escudero, L.A. & Gordaliza, A. (1999). Robustness properties of k-means and trimmed k-means. *Journal of the American Statistical Association*, *94*, 956–9,69.

Gardner-Lubbe, S., Le Roux, N.J. & Gower, J.C. (2008). Measures of Fit in Principal Component and Canonical Variate Analyses. *Journal of Applied Statistics*, *35*, 947-965.

Gil-Núñez, J.C., Martínez-Córdova, L.R., Servín-Villegas, R.S., Magallon-Barajas, F.J., Bórques-López, R.A., Gonzalez-Galaviz, J.R. & Casillas-Hernández, R. (2020). Production of *Penaeus vannamei* in low salinity, using diets formulated with different protein sources and percentages. *Latin American Journal of Aquatic Research*, *48*, 396–405.

Govender, P. & Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research*, *11*, 40-56.

Gower, J.C., Lubbe, S. & Le Roux, N.J. (2011). *Understanding Biplots*. John

Wiley & Sons: Chicester, UK.

Guan, Y., Ghorbani, A.A. & Belacel, N. (2003). Y-means: a clustering method for intrusion Detection. In: IEEE Canadian conference on electrical and computer engineering, proceedings, 1083–1086.

Guartatanga, R., Schwartz, L., Wigglesworth, J.M. & Griffith, D.R.W. (1993). Experimental intensive rearing of red drum (*Sciaenops ocellatus*) in Ecuador.

Guevara-Viejó, F., Valenzuela-Cobos, J.D., Vicente-Galindo, P. & Galindo-Villardón, P. (2021). Application of K-Means Clustering Algorithm to Commercial Parameters of *Pleurotus* spp. Cultivated on Representative Agricultural Wastes from Province of Guayas. *Journal of Fungi*, *7*, 537.

Guler, C., Thyne, G.D., McCray, J.E. & Turner, A.K. (2002). Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. *Hydrogeology Journal*, *10*, 455–474.

Gulzar, S., Raju, N., Nagarajarao, R.C. & Benjakul, S. (2020). Oil and pigments from shrimp processing by-products: Extraction, composition, bioactivities and its application—A review. *Trends in Food Science & Technology*, *100*, 307–319.

Halim, N.R., Yusof, H.M. & Sarbon, N.M. (2016). Functional and bioactive properties of fish protein hydolysates and peptides: A comprehensive review. *Trends in Food Science & Technology*, *51*, 24–33.

Hamilton, S.E. & Stankwitz, C. (2012). Examining the relationship between international aid and mangrove deforestation in coastal Ecuador from 1970 to 2006. *Journal of Land Use Science*, 7, 177–202.

Han, J. & Kamber, M. (2001) Data mining: concepts and techniques. Massachusetts: Morgan Kaufmann Publishers.

Hansen, P. & Mladenovic, N. (2001). J-means: A new local search heuristic for minimum sum-of-squares clustering. *European Journal of Operational Research*, *130*, 449-467.

Hartigan, J. (1975). Clustering algorithms. John Wiles & Sons, New York.

Hayes-Roth, F., Waterman, D. A., & Lenat, D. B. (1983). *Building expert systems*. Reading, MA: Addison-Wesley.

Holmgren, P. & Thuresson, T. (1998). Satellite remote sensing for forestry planning: a review. *Scandinavian Journal of Forest Research*, *13*, 90–110.

Hornik, K., Feinerer, I., Kober, M. & Buchta, C. (2012). Spherical k-Means Clustering. *Journal of Statistical Software*, *50*, 1–22.

Hot, E. & Popović-Bugarin, V. (2016). Soil data clustering by using K-means and fuzzy K-means algorithm. In: Telecommunications Forum Telfor (TELFOR), pp. 890-893.

Huang, F., Wang, L., Zhang, C. & Song, K. (2017). Replacement of fishmeal with soybean meal and mineral supplements in diets of *Litopenaeus vannamei* reared in low-salinity water. *Aquaculture*, *473*, 172–180.

Jong, S.C. & Peng, J.T. (1975). Identity and cultivation of a new commercial mushroom in Taiwan. *Mycology*, *67*, 1235-1240.

Jorquera, H., Perez, R., Cipriano, A. & Acuna, G. (2001). Short term forecasting of air pollution episodes. In: Zannetti P (eds) Environmental modeling 4. WIT Press, UK.

Kashangura, C., Hallsworth, J.E. & Mswaka, A.Y. (2006). Phenotypic diversity amongst strains of *Pleurotus sajor-caju*: Implications for cultivation in arid environments. *Mycological Research*, *110*, 312-317.

Kaufman, L. & Rousseeuw, P.J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, New York.

Kaul, T.N. (1983). Cuftivated edible mushrooms. Regional Research Laboratory, Jammu.

Kaul, T.N. & Kapur, Y.B.M. (eds.). (1987) Indian Mushroom Science, 11. Procceding Intern. Conference on Science and Cultivation Technol. Edible Fungi. Regional Research Laboratory, Jammu.

Kodinariya, T.M. & Makwana, P.R. (2013). Review on determining number of cluster in K-means clustering. *International Journal of Advance Research in Computer Science and Management Studies*, *1*, 90–95.

Kostic, M., Smiljkovic, M., Petrovic, J., Glamocilija, J., Barros, L., Ferreira, I.C.F.R., Ciric, A. & Sokovic, M. (2017). Chemical, nutritive composition and wide-broad bioactive properties of honey mushroom *Armillaria mellea* (Vahl: Fr.) Kummer. *Food & Function*, *8*, 3239–3249.

Krishna, K. & Murty, M. (1999) Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *29*, 433–439.

Kurgan, L.A. & Musilek P. (2006). A Survey of Knowledge Discovery and

Data Mining Process Models. *The Knowledge Engineering Review*, *21*, 1-24.

Lakzian, A., Berenji, A.R., Karimi, E. & Razavi, S. (2008). Adsorption capability of lead, nickel and zinc by exopolysaccharide and dried cell of *Ensifer meliloti*. *Asian Journal of Chemistry*, *20*, 6075-6080.

Lawrence, A.L. & Lee, P.G. (1997). Research in the Americas. In Crustacean Nutrition. Advances in World Aquaculture; D'Abramo, L.R., Conklin, D.E., Akiyama, D.M., Eds.; World Aquaculture Society: Baton Rouge, LA, USA, Volume 6, pp. 566–587.

Liberty, E., Sriharsha, R. & Sviridenko, M. (2014). An algorithm for online k-means clustering. CoRR, abs/1412.5721.

Liu, B., Hsu, W., & Ma, Y. (1999). Mining association rules with multiple minimum supports. *Proceedings of the ACM-SIGMOD international conference on knowledge discovery and data mining (KDD'99)* (pp. 15–18).

Liu, H., Chen, N., Feng, C., Tong, S. & Li, R. (2017). Impact of electrostimulation on denitrifying bacterial growth and analysis of bacterial growth kinetics using a modified Gompertz model in a bio-electrochemical denitrification reactor. *Bioresource Technology*, *232*, 344-353.

Liu, Z., Liu, Q., Zhang, D., Wei, S., Sun, Q., Xia, Q., Shi, W., Ji, H. & Liu, S. (2021). Comparison of the Proximate Composition and Nutritional Profile of Byproducts and Edible Parts of Five Species of Shrimp. *Foods*, *10*, 2603.

Lloyd, S.P. (1982) Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*, 129–137.

Madeiro, F., Galvão, R.R.A., Ferreira, F.A.B.S. & Cunha, D.C. (2012). Uma alternativa de aceleração do algoritmo fuzzy k-means aplicado à quantização vetorial. *TEMA Trends in Computational and Applied Mathematics*, *13*, 193–206.

Majumdar. J. & Ankalaki, S. (2016). Comparison of clustering algorithms using quality metrics with invariant features extracted from plant leaves. In: Paper presented at international conference on computational science and engineering.

Majumdar, J., Naraseeyappa, S. & Ankalaki, S. (2017). Analysis of agriculture data using data mining techniques: Application of big data. *Journal of Big Data*, *4*, 20.

Manzi, P., Aguzzi, A. & Pizzoferrato, L. (2001). Nutritional value of mushrooms widely consumed in Italy. *Food Chemistry*, *73*, 321-325.

McQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Computer and Chemistry*, *4*, 257-272.

Meyer, G.E., Neto, J.C., Jones, D.D. & Hindman, T.W. (2004). Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images. *Computers and Electronics in Agriculture, 42*, 161–180.

Mmanda, F.P., Lindberg, J.E., Halden, A.N., Mtolera, M.S.P., Kitula, R. & Lundh, T. (2020). Digestibility of local feed ingredients in tilapia *Oreochromis niloticus* juveniles, determined on faeces collected by siphoning or stripping. *Fishes*, *5*, 32.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM, 43*, 142–151.

Mocan, A., Fernandes, A., Barros, L., Crişan, G., Smiljković, M., Soković, M. & Ferreira, I.C.F. (2018). Chemical composition and bio-active properties of the wild mushroom *Polyporus squamosus* (Huds.) Fr: A study with samples from Romania. *Food Science & Nutrition*, *9*, 160-170.

Mohanty, R.K. (1999). Growth performance of *Penaeus monodon* at different stocking densities. *Inland Fisheries Society of India*, *31*, 53–59.

Mora, E. (1988). Auge y crisis de una economía agroexportadora: el período cacaotero. Quito: Corporación Editora Nacional-Editorial Grijalbo Ecuatoriana.

Mori, K., Fukai, S. & Zennyoji, A. (1974). Hybridization of shiitake (*Lentinus edodes*) between cultivated strains of Japan and wild strains grown in Taiwan and New Guinea. *Mushroom Science*, *9*, 391-403.

Naylor, R.L.; Goldberg, R.J.; Mooney, H.; Beveridge, M.; Clay, J.; Folke, C.; Kautsky, N.; Lubchenco, J.; Primavera, J. & Williams, M. (1998). Nature's subsidies to shrimp and salmon farming. *Science*, 282, 883–884.

Naylor, R.L., Goldberg, R.J., Primavera, J.H., Kautsky, N., Beveridge, M.C.M., Clay, J., Folke, C., Lubchenco, J., Mooney, H. & Troell, M. (2000). Effect

of aquaculture on world fish supplies. *Nature*, 405, 1017–1024.

Naylor, R.L. (2011). Expanding the Boundaries of Agricultural Development. *Food Security*, 3, 233–251.

Nirmal, N.P., Santivarangkna, C., Rajput, M.S. & Benjakul, S. (2020). Trends in shrimp processing waste utilization: An industrial prospective. *Trends in Food Science & Technology*, *103*, 20–35.

Oliva, G., Setola, R. & Hadjicostis, C.N. (2013). Distributed K-means algorithm, arXiv:1312-4176.

Oyedele, O. F., & Lubbe, S. (2015). The Construction of a Partial Least Squares Biplot Opeoluwa. *Journal of Applied Statistics*, *42*(11), 2449–2460.

Palma, J., Bureau, D.P. & Andrade, J.P. (2008). Effects of binder type and binder addition on the growth of juvenile *Palaemonetes varians* and *Palaemon elegans* (Crustacea: Palaemonidae). *Aquaculture International*, 16, 427–436.

Partridge, G.J. & Southgate, P.C. (1999). The effect of binder composition on ingestion and assimilation of microbound diets MBD by barramundi Lates calcarifer Bloch larvae. *Aquaculture Research*, 30, 879–886.

Pasqualoto, K.F., Teófilo, R.F., Guterres, M., Pereira, F.S. & Ferreira, M. (2007). A study of physicochemical and biopharmaceutical properties of Amoxicillin tablets using full factorial design and PCA biplot. *Analytica Chimica Acta*, *595*, 216–220.

Patel, V.C., McClendon, R.W. & Goodrum, J.W. (1994). Crack detection in eggs using computer vision and neural networks. *Artif Intell Appl.*, *8*, 21–31.

Pineo, R. (2008). Guayaquil and coastal Ecuador during the cacao era. In C. De La Torre & S. Striffler (Eds.), The Ecuador reader (pp. 136–147). Durham and London: Duke University Press.

PROECUADOR. (2013). Instituto de Promoción de Exportaciones e Inversiones. El camarón congelado es el tercer producto de exportación.

Rajagopalan, B. & Lall, U. (1999) A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water Resources Research*, *35*, 3089–3101.

Rasulov, B.A., Yili, A. & Aisa, H.A. (2013). Biosorption of metal ions by exopolysaccharide produced by *Azotobacter chroococcum* XU1. *Journal*

of Environmental Protection, *4*, 989-993.

Razavi Zadegan, S.M., Mirzaie, M. & Sadoughi, F. (2013). Ranked k-medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets. *Knowledge-based Systems*, *39*, 133–143.

Reis, F.S., Barros, L., Martins, A. & Ferreira, I. (2012). Chemical composition and nutritional value of the most widely appreciated cultivated mushrooms: An inter-species comparative study. *Food and Chemical Toxicology*, *50*, 191-197.

Ringner, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*, 303–304.

Riul A Jr, de Sousa, H.C., Malmegrim, R.R., dos Santos, D.S. Jr., Carvalho, A.C.P.L.F., Fonseca, F.J., Oliveira, Jr. O.N. & Mattoso, L.H.C. (2004). Wine classification by taste sensors made from ultra-thin films and using neural networks. *Sensors and Actuators B*, 77–82.

Rivera, L.M., Trujillo, L.E., Pais-Chanfrau, J.M., Núñez, J., Pineda, J., Romero, H., Tinococo, O., Cabrera, C. & Dimitrov, V. (2018). Functional foods as stimulators of the immune system of *Litopenaeus vannamei* cultivated in Machala, Province of El Oro, Ecuador. *Italian Journal of Food Science*, 227–232.

Rodríguez Suárez, Y. & Díaz Amador, A. (2011) Herramientas de minería de datos. *Revista Cubana de Ciencias Informáticas*, *3*, 3-4.

Salmones, D., Gaitán-Hernández, R., Pérez, R. & Guzmán, G. (1997). Estudios sobre el género Pleurotus. VIII. Interacción entre crecimiento micelial y productividad. *Revista Iberoamericana de Micología*, *14*, 173-176.

Sánchez-Hernández, A., Valenzuela Cobos, J.D., Herrera Martínez, J., Arce, R.V., Gómez y Gómez, Y.M., Segura, P.B.Z., Aguilar, M.E.G., Lara, H.L. & Valencia del Toro, G. (2019). Characterization of *Pleurotus djamor* neohaplonts recovered by production of protoplasts and chemical dedikaryotization. *3 Biotech*, *9*, 24.

Sculley, D. (2010). Web-scale k-means clustering. In Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010, 1177–1178.

Singh, R.S., Bhari, R. & Kaur, H.P. (2010). Mushroom lectins: current status and future perspectives. *Critical Reviews in Biotechnology*, *30*, 99–126.

Shahin, M.A., Tollner, E.W. & McClendon, R.W. (2001). Artificial intelligence classifiers for sorting apples based on watercore. *Journal of Agricultural Engineering Research*, *79*, 265–274.

Shahkar, E., Yun, H., Park, G., Jang, I.K., Kyoung Kim, S., Katya, K. & Bai, S.C. (2014). Evaluation of optimum dietary protein level for juvenile whiteleg shrimp (*Litopenaeus vannamei*). *Journal of Crustacean Biology*, *34*, 552–558.

Smith, D.M., Burford, M.A., Tabrett, S.J., Irvin, S.J. & Ward, L. (2002). The effect of feeding frequency on water quality and growth of the black tiger shrimp (*Penaeus monodon*). *Aquaculture*, 207, 125–136.

Spath, H. (1980). Cluster analysis algorithms for data reduction and classification of objects. Ellis Horwood, Chichester.

Striffer, S. (2008). The united fruit company's legacy in Ecuador. In C. De La Torre & S. Striffler (Eds.), The Ecuador reader (pp. 239–249). Durhan and London: Duke University Press.

Stolz, T., Huertas, M.E. & Mendoza A. (2020) Assessment of air quality monitoring networks using an ensemble clustering method in the three major metropolitan areas of Mexico. *Atmospheric Pollution Research*, *11*, 1271-1280.

Sung, K.K. & Poggio, T. (2009) Example-based learning for view-based human face detection. A.I. Memo 1521, MIT.

Swanson, K. (2007). Revanchist urbanism heads south: The regulation of indigenous beggars and street vendors in Ecuador. Anti-pode journal. University Of Glasgow, Glasgow, UK: Department of Geography and Earth Sciences.

Tacon, A.G.J., Cody, J.J., Conquest, L.D., Divakaran, S., Forster, I.P. & Decamp, O.E. (2002). Effect of culture system on the nutrition and growth performance of Pacific white shrimp *Litopenaeus vannamei* (Boone) fed different diets. *Aquaculture Nutrition*, *8*, 121–139.

Tan, P. N., & Kumar, V. (2000). Interestingness measures for association patterns: A perspective. *KDD 2000 Workshop on Postprocessing in Machine Learning and Data Mining, Boston, MA, August.*

Thongsook, T. & Kongbangkerd, T. (2011). Influence of calcium and silicon supplementation into *Pleurotus ostreatus* substrates on quality of fresh and

canned mushrooms. *Food Science and Technology International*, *17*, 351–365.

Tripathi, S., Srinivas, V.V. & Nanjundiah, R.S. (2006). Downscaling of precipitation for climate change scenarios: a support vector machine approach. *Journal of Hydrology*, *330*, 621–640.

Tsiptsis, K. & Chorianopoulos, A. (2009). Data mining techniques in CRM: Inside Customer Segmentation, UK, John Wiley and Sons Ltd.

Tsukatani, T., Suenaga, H., Shiga, M., Noguchi, K., Ishiyama, M., Ezoe, T. & Matsumoto, K. (2012). Comparison of the WST-8 colorimetric method and the CLSI broth microdilution method for susceptibility testing against drug-resistant bacteria. *Journal of Microbiological Methods*, *90*, 160–166.

Valencia del Toro, G., Ramírez-Ortiz, M.E., Flores-Ramírez, G., Costa-Manzano, M.R., Robles-Martínez, F., Garín Aguilar, M.E. & Leal-Lara, H. (2018). Effect of Yucca *schidigera bagasseas* substrate for Oyster mushroom on cultivation parameters and fruit body quality. *Revista Mexicana de Ingeniería Química*, *17*, 835-846.

Valenzuela-Cobos, J.D., Páramo, E.D., Arce, R.V., Sánchez-Hernández, A., Aguilar, M.E.G., Lara, H.L. & Valencia del Toro., G. (2017). Production of hybrid strains among *Pleutorus* and *Lentinula* and evaluation of their mycelial growth kinetics on malt ex-tract agar and wheat grain using the Gompertz and Hill models. *Emirates Journal of Food and Agriculture*, *29*, 927-935.

Valenzuela-Cobos, J.D., Vásquez-Véliz, G., Zied, D.C., Franco-Hernández,O.M., SánchezHernández, A., Garín Aguilar, M.E., Leal Lara, H. & Valencia del Toro. G. (2019). Bioconversion of agricultural wastes using parental, hybrid and reconstituted strains of *Pleurotus* and *Lentinula*. *Revista Mexicana de Ingeniería Química*, *18*, 647–657.

Valenzuela-Cobos, J.D., Rodríguez-Grimón, R.O., Jara-Bastidas, M.L., Grijalva-Endara, A., Zied, D.C., Garín-Aguilar, M.E. & del Toro, G.V. (2020). Modeling of micelial growth of parental, hybrid and reconstituted strains of *Pleurotus* and *Lentinula*. *Revista Mexicana de Ingeniería Química*, *19*, 165-174.

Valenzuela-Cobos, J.D. & Vargas-Farias, C.J. (2020). Study about the use of aquaculture binder with tuna attractant in the feeding of white shrimp

(*Litopenaeus vannamei*). *Revista Mexicana de Ingeniería Química*, *19*, 355-361.

Velasco, M., Lawrence, A.L. & Castille, F.L. (1999). Effect of variations in daily feeding frequency and ration size on growth of shrimp, *Litopenaeus vannamei* (Boone), in zero-water exchange culture tanks. *Aquaculture*, *179*, 141–148.

Verheyen, K., Adriaens, D., Hermy, M. & Deckers, S. (2001). High-resolution continuous soil classification using morphological soil profile descriptions. *Geoderma*, *101*, 31–48.

Vicente-Villardón, J.L. (2010a). MULTBIPLOT: A package for Multivariate Analysis using Biplots; Departamento de Estadística, Universidad de Salmanca: Salmanca, Spain; Available online: http://biplot.usal.es/multbiplot.

Vicente-Villardón, J.L. (2010b). MultBiplotR: Multivariate Analysis Using Biplots; Version 0.1.0.; Departamento de Estadística. Universidad de Salmanca: Salmanca, Spain; Available online: https://CRAN.R-project.org/package=MultBiplotR.

Xie, S.W., Li, Y.T., Zhou, W.W., Tian, L.X., Li, Y.M., Zeng, S.L. & Liu, Y.J. (2017). Effect of γ-aminobutyric acid supplementation on growth performance, endocrine hormone and stress tolerance of juvenile Pacific white shrimp, *Litopenaeus vannamei*, fed low fishmeal diet. *Aquaculture Nutrition*, 23, 54–62.

Wagner, D., Mitchell, A., Sassaki, G.L. & de Almeida Amazonas., MA.L. (2004). Links between morphology and physiology of *Ganoderma lucidum* in submerged culture for the production of exopolysaccharide. *Journal of Biotechnology*, *114*, 153-164.

Wang, Q., Wang, C., Feng, Z. & Ye, J. (2012). Review of K-means clustering algorithm. *Electronic design engineering*, *20*, 21–24.

Wang, K., He, Y., & Han, J. (2000). Mining frequent itemsets using support constraints. *Proceedings of the international conference on very large data bases (VLDB'00).*

Witten, I., Frank, E. & Hall, M. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Amsterdam: Morgan Kaufman.

Yan, W., Hunt, L. A., Sheng, Q. & Szlavnics, Z. (2000). Cultivar evaluation and mega-environment investigation based on GGE biplot. *Crop Science, 40*, 597-605.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265–286