UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

TITULO:

Implementación de Machine Learning para predicción de costos de camarón hacia China con exportadores de Guayaquil

Autores:

Acosta Pisco Nicolás Julián

Yepez Jouvin Geanella Alejandra

Trabajo de titulación previo a la obtención del título de

Licenciado en Negocios Internacionales

Tutor:

Ing. Carrera Buri, Félix Miguel, Mgs

Guayaquil, Ecuador

7 de febrero del 2025

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

CERTIFICACIÓN

Certificamos que el trabajo de integración curricular fue realizado en su totalidad por ACOSTA PISCO NICOLÁS JULIÁN y YEPEZ JOUVIN GEANELLA ALEJANDRA, como requerimiento para la obtención del título de Licenciado en Negocios Internacionales.

TUTOR

f._____

Ing. Carrera Buri, Félix Miguel, Mgs.

DIRECTORA DE LA CARRERA

f._____

Ing. Hurtado Cevallos, Gabriela Elizabeth, Mgs.

Guayaquil, 7 de febrero del 2025

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

DECLARACIÓN DE RESPONSABILIDAD

Nosotros**, Acosta Pisco, Nicolás Julián;**

Yepez Jouvin, Geanella Alejandra

DECLARAMOS QUE:

El trabajo de Integración Curricular**, Implementación de Machine Learning para predicción de costos de camarón hacia China con exportadores de Guayaquil**., previo a la obtención del título de **Licenciado en Negocios Internacionales**, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría- En esta virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Integración Curricular referido.

Guayaquil, 7 de febrero del 2025

AUTORES

f._____

Acosta Pisco, Nicolás Julián

f._____

Yepez Jouvin, Geanella Alejandra

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

AUTORIZACIÓN

Nosotros**, Acosta Pisco, Nicolás Julián;**

Yepez Jouvin, Geanella Alejandra

Autorizo a la Universidad Católica de Santiago de Guayaquil a la publicación en la biblioteca de la institución el componente práctico del examen complexivo, "Factores determinantes en el consumo de medicina natural en la ciudad de Guayaquil: Perfil del consumidor", cuyo contenido, ideas y criterios son de mi exclusiva responsabilidad y total autoría.

Guayaquil, 7 de febrero del 2025

AUTORES

f._____

Acosta Pisco, Nicolás Julián

f._____

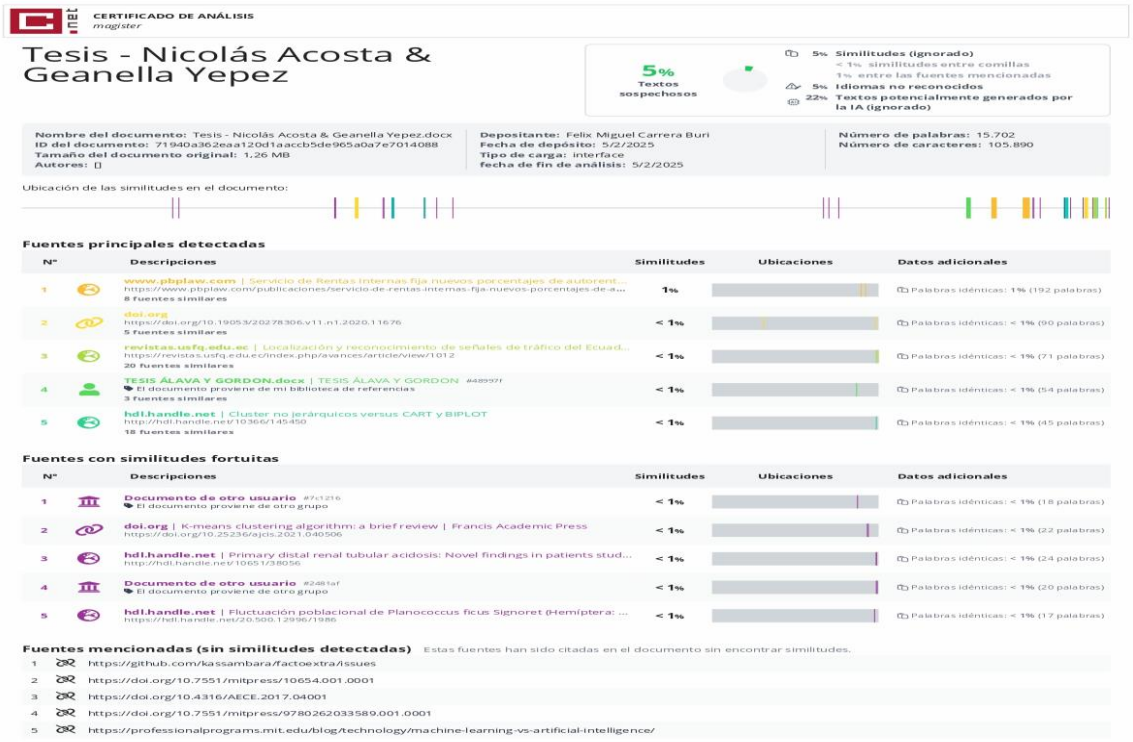
Yepez Jouvin, Geanella Alejandra

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

REPORTE COMPILATIO



CERTIFICADO DE ANÁLISIS
magister

Tesis - Nicolás Acosta & Geanella Yepez

5% Textos sospechosos

5% Similitudes (ignorado)
< 1% similitudes entre comillas
1% entre las fuentes mencionadas
5% Idiomas no reconocidos
22% Textos potencialmente generados por la IA (ignorado)

Nombre del documento: Tesis - Nicolás Acosta & Geanella Yepez.docx
ID del documento: 71940a362eaa120d1aaccb5de965a0a7e7014088
Tamaño del documento original: 1,26 MB
Autores: []

Depositante: Felix Miguel Carrera Buri
Fecha de depósito: 5/2/2025
Tipo de carga: interface
fecha de fin de análisis: 5/2/2025

Número de palabras: 15.702
Número de caracteres: 105.890

Ubicación de las similitudes en el documento:

Fuentes principales detectadas

Nº	Descripciones	Similitudes	Ubicaciones	Datos adicionales
1	www.pbplaw.com | Servicio de Rentas Internas fija nuevos porcentajes de autorent... https://www.pbplaw.com/publicaciones/servicio-de-rentas-internas-fija-nuevos-porcentajes-de-a... 8 fuentes similares	1%		Palabras idénticas: 1% (192 palabras)
2	doi.org https://doi.org/10.19053/20278306.v11.n1.2020.11676 5 fuentes similares	< 1%		Palabras idénticas: < 1% (90 palabras)
3	revistas.usfq.edu.ec | Localización y reconocimiento de señales de tráfico del Ecuad... https://revistas.usfq.edu.ec/index.php/avances/article/view/1012 20 fuentes similares	< 1%		Palabras idénticas: < 1% (71 palabras)
4	TESIS ÁLAVA Y GORDON.docx | TESIS ÁLAVA Y GORDON #489971 El documento proviene de mi biblioteca de referencias 3 fuentes similares	< 1%		Palabras idénticas: < 1% (54 palabras)
5	hdl.handle.net | Cluster no jerárquicos versus CART y BIPLOT http://hdl.handle.net/10366/145450 18 fuentes similares	< 1%		Palabras idénticas: < 1% (45 palabras)

Fuentes con similitudes fortuitas

Nº	Descripciones	Similitudes	Ubicaciones	Datos adicionales
1	Documento de otro usuario #71216 El documento proviene de otro grupo	< 1%		Palabras idénticas: < 1% (18 palabras)
2	doi.org | K-means clustering algorithm: a brief review | Francis Academic Press https://doi.org/10.25236/ajcis.2021.040506	< 1%		Palabras idénticas: < 1% (22 palabras)
3	hdl.handle.net | Primary distal renal tubular acidosis: Novel findings in patients stud... http://hdl.handle.net/10651/38056	< 1%		Palabras idénticas: < 1% (24 palabras)
4	Documento de otro usuario #2481af El documento proviene de otro grupo	< 1%		Palabras idénticas: < 1% (20 palabras)
5	hdl.handle.net | Fluctuación poblacional de Planococcus ficus Signoret (Hemíptera: ... https://hdl.handle.net/20.500.12996/1986	< 1%		Palabras idénticas: < 1% (17 palabras)

Fuentes mencionadas (sin similitudes detectadas) Estas fuentes han sido citadas en el documento sin encontrar similitudes.

1 https://github.com/kassambara/factoextra/issues
2 https://doi.org/10.7551/mitpress/10654.001.0001
3 https://doi.org/10.4316/AECE.2017.04001
4 https://doi.org/10.7551/mitpress/9780262033589.001.0001
5 https://professionalprograms.mit.edu/blog/technology/machine-learning-vs-artificial-intelligence/

f. _____

Ing. Carrera Buri, Félix Miguel, Mgs.

TUTOR

AGRADECIMIENTO

La culminación de esta investigación ha sido posible gracias a la ayuda y la colaboración de algunas personas e instituciones, a quienes agradecemos de todo corazón.

Primero agradezco a la Facultad de Ciencias Económicas y Empresariales, y en especial a la carrera de Negocios Internacionales, por brindarnos las herramientas y conocimientos para la elaboración de esta investigación.

A nuestro tutor Msgc. Félix Miguel Carrera Buri, por su guía, paciencia y asesoramiento durante todo el proceso, por su asesoría y apoyo al mejoramiento de este trabajo, a los exportadores de las empresas camaroneras de Guayaquil por su disposición y por brindar su tiempo para brindarnos la información necesaria para la realización de nuestra investigación.

De una manera especial agradezco a mi familia, por brindarme su apoyo, aliento y confianza, para conseguir mis metas, a la vez reconozco el esfuerzo realizado por cada uno de los autores de esta investigación y el trabajo en equipo, que permitió la culminación de este proyecto.

A todos ustedes nuestro agradecimiento.

- ***Geanella Alejandra Yepez Jouvin***

AGRADECIMIENTO

Quiero empezar agradeciendo a mi familia, que ha sido un pilar importante para mí. Mi madre Mariana, quien ha estado siempre para mí incluso en sus momentos más difíciles. También mi padre Robinson, que ha sabido guiarme y ser un apoyo en mi vida. Y finalmente, mi hermana Doménica la persona con la que he compartido todo este tiempo de estudio.

También agradecer a nuestro tutor y a mi compañera de tesis por ser parte de todo esto. Este momento marca un momento trascendental en mi vida estudiantil y profesional. Su apoyo fue un punto clave para poder realizar esta investigación y siempre estando presentes.

A mis amigos, personas con las que compartí momentos inolvidables que marcaron mi vida y mi personalidad. Me enseñaron a ver el mundo de otra manera y aprender a ser una mejor persona diariamente.

Trato de ser breve, porque aún me queda muchas cosas por vivir y recorrer. Finalmente, solo quiero decir que agradezco a Dios por esto.

Viaje antes que destino, vida antes que muerte y fuerza antes que debilidad.

- ***Nicolás Julián Acosta Pisco***

-

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

TRIBUNAL DE SUSTENTACIÓN

f. _____

Ing. Carrera Buri, Félix Miguel, Mgs.

TUTOR

f. _____

Ing. Hurtado Cevallos, Gabriela Elizabeth Mgs.

DECANO O DIRECTOR DE CARRERA

f. _____

(NOMBRES Y APELLIDOS)

COORDINADOR DEL ÁREA O DOCENTE DE LA CARRERA

Contenido

Resumen

Las exportaciones actualmente son un reto logístico para una gran cantidad de empresas. Los exportadores de camarón deben adaptarse a los nuevos retos que se enfrentan y empresas que manejan estas exportaciones siempre están afrontando variaciones de costos en sus operaciones.

Este aumento de costos obliga a las empresas a subir sus precios e incluso pueden dificultar las exportaciones recortando recursos financieros para los camaroneros. Dentro del estudio de los datos el poder conocer y el poder implementar herramientas de Machine Learning para predecir los costos de exportación de camarón hacia China, se vuelve un recurso muy importante.

Con los datos proporcionados por un forwarder en Guayaquil. Se indago que algoritmo era el más adecuado para su aplicación, para después aplicar los algoritmos de K-Means y KNN con el fin de segmentar a los clientes en grupos con características similares y predecir los costos logísticos asociados a las exportaciones.

Los resultados mostraron una segmentación entre "Clientes Premium" y "Clientes Normales", lo que permitirá desarrollar estrategias diferenciadas. Además, se desarrolló un modelo predictivo con un bajo error cuadrático medio (RMSE), lo que sugiere que el modelo es preciso para estimar los costos de exportación. Con esto poder prevenirse a los costos y saber que se puede reducir los costos de venta para obtener mayor ganancia.

Finalmente, se propusieron mejoras futuras, como la inclusión de variables adicionales y la integración de otros algoritmos de Machine Learning, para optimizar aún más el modelo.

Este estudio contribuye a la optimización de la logística en las exportaciones de camarón y también mejoras logísticas tanto como para el forwarder como el exportador. Donde recibirá un servicio adecuado a los costos que esta pagando al ser considerado al grupo de clientes asignados y el forwarder podrá reducir sus costos a través de distintas estrategias que apliquen al conocer las predicciones del modelo.

Palabras claves: Balanza comercial, recursos financieros, sistema logístico

Abstract

Exports are currently a logistical challenge for many companies. Shrimp exporters must adapt to the new challenges they face and companies that handle these exports are always facing cost variations in their operations.

These cost increases force companies to raise their prices and can even hinder exports by cutting financial resources for shrimp farmers. Within the study of data, being able to know and implement Machine Learning tools to predict the costs of shrimp exports to China becomes a very important resource.

With the data provided by a forwarder in Guayaquil. The investigation has determined which algorithm was the most appropriate for its application, and then applied the K-Means and KNN algorithms in order to segment customers into groups with similar characteristics and predict the logistics costs associated with exports.

The results showed a segmentation between "Premium Customers" and "Normal Customers", which will allow the development of differentiated strategies. In addition, a predictive model with a low root mean square error (RMSE) was developed, suggesting that the model is accurate in estimating export costs. With this, it is possible to anticipate costs and know that sales costs can be reduced to obtain higher profits.

Finally, future improvements were proposed, such as the inclusion of additional variables and the integration of other Machine Learning algorithms, to further optimize the model.

This study contributes to the optimization of logistics in shrimp exports and also logistical improvements for both the forwarder and the exporter. The forwarder will receive a service adequate to the costs he is paying by being considered to the group of assigned clients and the forwarder will be able to reduce his costs through different strategies applied by knowing the predictions of the model.

Keywords: Balance of trade, financial resources, logistics system

1. Introducción

En la actualidad, la industria del camarón ecuatoriano se ha posicionado como una con gran dinamismo en la economía nacional, gracias a su alto nivel de crecimiento (Coello Landires, 2021). Las exportaciones de camarón particularmente emergen como una oportunidad clave para diversificar el comercio y maximizar el potencial de este recurso. Sin embargo, ¿cuáles son los antecedentes y factores que han permitido este crecimiento en las exportaciones hacia el mercado internacional?

Para entender la evolución de este crustáceo como alimento y como ha llegado a impactar la dinámica de la economía mundial es importante comprender sus inicios

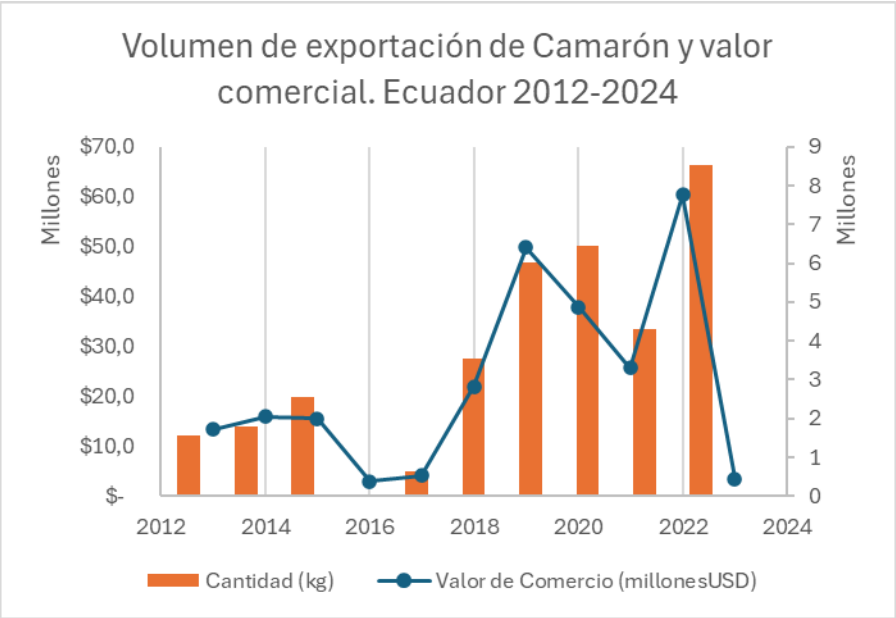
Desde la antigüedad hasta su cultivo intensivo en la actualidad. La industria camaronera se inicia en el Ecuador a finales de la década de los sesenta en las pampas salinas o salitrales (Boletín #10: Los sectores que se favorecen: El camarón, abril 04, 2016).
La acuicultura fue oficialmente introducida en Ecuador en 1968, según Cornejo Moreira y Villegas Guerrero (2018), y desde ese momento se han propuesto métodos novedosos que han logrado mejorar tanto la cosecha como la calidad del camarón. De esta manera, se ha logrado que el camarón ecuatoriano se posicione como un producto relevante en el mercado global.

En las décadas de 1970 y 1980, la acuicultura en Ecuador fue importante a medida que el camarón se consolidó como uno de los principales productos de exportación del país. La expansión de la rentabilidad de camarón ha sido impulsada por políticas gubernamentales que alentaron la inversión en la industria, y Ecuador logro

posesionarse como uno de los mayores exportadores de camarón del mundo a mediados de los años 1980. Ecuador ha adoptado reglas internacionales de acuicultura, lo que le permitió acceder a diversos mercados a nivel global, como: Estados Unidos, Europa y especialmente China, lo que lo convierte en una de las pocas especies de camarón ecuatoriano que se produce.

Año	Valor de Comercio (millones USD)	Cantidad (kg)
2013	$ 13.407.203,0	1560123
2014	$ 15.933.738,0	1798356
2015	$ 15.511.593,0	2557142
2016	$ 3.020.172,0	490,9
2017	$ 4.158.271,8	640625
2018	$ 21.918.970,4	3558557
2019	$ 49.849.962,1	6005793
2020	$ 37.940.626,8	6432879
2021	$ 25.629.605,1	4294977
2022	$ 60.326.775,1	8531066
2023	$ 3.460.166,5	661,4

Gráfico 1: Evolución de la producción y exportación de camarón en Ecuador desde 2012 hasta 2024. Muestra el crecimiento en toneladas y valor exportado a nivel global.



La producción de camarón en Ecuador ha venido creciendo a lo largo de los años, gracias a las nuevas técnicas que los productores de camarón están utilizando, que han hecho que se vean resultados muy buenos en los cultivos, dando énfasis a la forma, a la calidad y el sabor, para que en el mercado se responda de una manera muy favorable.

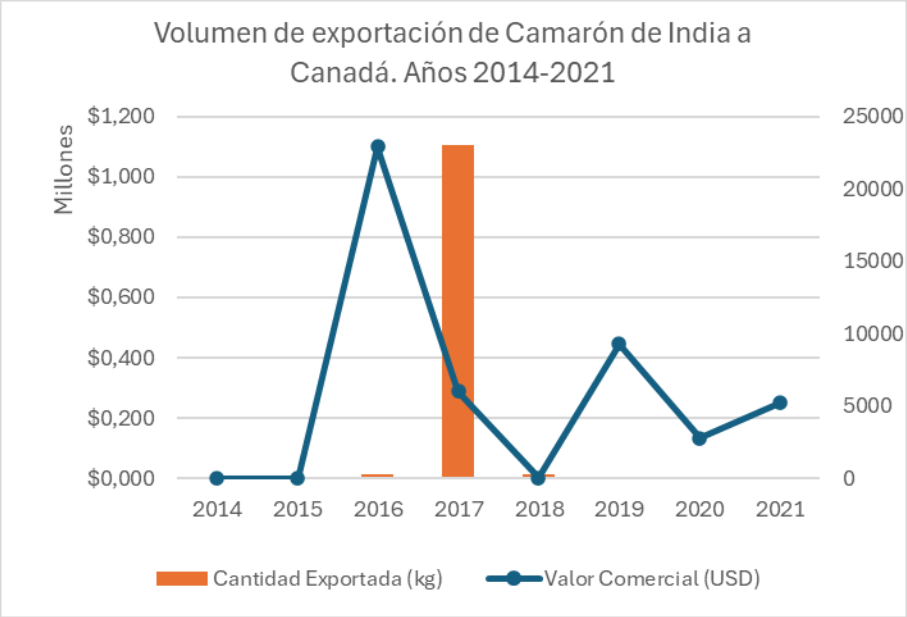
"El comercio internacional se ha visto facilitado por diversos acuerdos comerciales que benefician la entrada del camarón ecuatoriano en mercados con alta demanda. Entre 2015 y 2020, Ecuador aumentó su participación en las exportaciones de camarón a la Unión Europea, comenzando con un 10% y alcanzando un 20% en 2020 (Pol. Con., 2021)"

Entre los principales competidores de Ecuador en la exportación de camarón se encuentran India y Canadá, ambos países comparten varios de los mercados de destino. Sin embargo, a pesar de ser productores de camarón, estos países dependen en cierta

medida de la producción ecuatoriana por razones económicas. "Esto llevó a los emprendedores de camarón en el país a comenzar a exportar temprano, siendo Estados Unidos el principal destino, seguido por Europa y Asia" (Sarmiento et al., 2019).

Año	Valor Comercial (USD)	Cantidad Exportada (kg)	Tendencia del Valor	Tendencia de la Cantidad
2014	$105,000	10	Decrece	Decrece
2015	$439,000	64	Aumenta	Aumenta
2016	$1.100.000	292	Aumenta	Aumenta
2017	$289.086	23007	Decrece	Aumenta
2018	$1.295,000	315	Aumenta	Aumenta
2019	$446.600	111	Decrece	Decrece
2020	$133.100	23	Decrece	Decrece
2021	$250.200	89	Aumenta	Aumenta

Como muestra la Tabla 1, las exportaciones de camarón de India a Canadá entre 2014 y 2021 han experimentado fluctuaciones significativas.



Creada por el autor en base a los datos de las exportaciones de camarón de India a Canadá, 2014-2021.

Desde hace más de tres décadas, se ha intensificado la transformación de los manglares para la creación de estanques camaroneros en las playas y bahías del Ecuador.

Según informes del ex INEFAN, en enero del 2000 se registraron 207,000 hectáreas destinadas a la acuicultura, aunque la Cámara Nacional de Acuicultura sostiene que eran alrededor de 170,000. De estas, únicamente 50,454 hectáreas operan legalmente, mientras que el resto se considera ilegal. En la provincia de Esmeraldas, que alberga los manglares mejor preservados del país, más del 90% de los estanques son ilegales.

Datos del CLIRSEN indican que, en 1984, había 89,368 hectáreas de estanques a pesar de que el gobierno había declarado en 1985 la conservación de los bosques de manglar como de interés público y suspendió licencias para acuicultura en esas áreas. Sin

embargo, entre 1984 y 1999, el país vivió la mayor pérdida de manglares y el mayor crecimiento de estanques camaroneros.

A su vez, el negocio del camarón es rentable, pero no se considera sostenible o ecológico desde un punto de vista económico. De esta manera, la presión constante de este sector ha llevado a los gobiernos a ofrecer subsidios y otras formas de apoyo. Adicionalmente, entre 1997 y 2000, el Banco Mundial otorgó préstamos por 82 millones de dólares para el desarrollo de la acuicultura en América Latina, beneficiando a Ecuador en este proceso.

Los primeros pasos para construir un estanque de camarones consisten en limpiar el manglar y cortar grandes charcas que luego se llenarán con biocidas para matar a otros competidores del camarón. La aplicación de fertilizantes y antibióticos saliniza aún más el suelo y deja el área inutilizable para futuras actividades. Los ciclos de vida del camarón son los que amenazan este ecosistema en los manglares.

Habla por si solo con respecto a los impactos en la biodiversidad; disminución drástica en la diversidad de especies. 256 especies de animales nativos de Ecuador se pueden asociar con los manglares, los cuales están bajo amenaza. Los problemas de sostenibilidad ponen en riesgo a la industria de camarón, ya que la sobrepesca de larvas y el uso prolongado de suelos salinizados limitan la capacidad de la industria a largo plazo.

La industria camaronera, siendo un sector de las exportaciones más importantes de Ecuador, tiene retos en el camino, especialmente en la búsqueda de nuevos mercados.

Este proceso debe incluir los aspectos de sostenibilidad ecológica y la necesidad de proteger los ecosistemas y las comunidades locales.

Esto refleja la importancia del camarón no solo como producto, sino como motor de desarrollo para diversas comunidades, la producción de camarón en la provincia de El Oro actualmente corresponde al 37.68% de las exportaciones no petroleras del país lo que resalta su importancia en el contexto del Tratado Comercial con la Unión Europea (Varela-Véliz et al., 2017).

Los principales mercados corresponden a Asia, Estados Unidos y la Unión Europea. Estas tres áreas consumen la mayor parte de les exportaciones camaroneras ecuatorianas.

Como señala Schwartz, la competitividad del sector ecuatoriano permite al país aumentar su participación sin recurrir a la protección. Sin embargo, los esfuerzos del gobierno, en forma de mejoras en la facilitación aduanera y la promoción de las inversiones, fueron importantes para modernizar la producción de camarón y, por ende, su lugar en el mercado mundial.

Además de los hitos alcanzados por el sector, aún enfrenta desafíos en cuanto al cumplimiento de las regulaciones internacionales y la garantía de la sostenibilidad ambiental, lo cual se ha vuelto crítico porque a veces los consumidores están dispuestos a conocer los impactos ambientales que generan sus decisiones.

El impacto de la empleabilidad a causa de la cadena de valor que aporta el sector camaronero se puede destacar entre la costa, zona donde la actividad camaronera es la principal fuente de vida para muchas de las comunidades, pero la situación del camarón

ecuatoriano también enfrenta retos, entre ellos, la sustentabilidad ecológica y el fortalecimiento económico.

Desde un punto de vista local, la ciudad de Guayaquil es el corazón logístico y comercial para los exportadores de camarón. Aproximadamente el 60% de la producción nacional procede de esta región, gracias a su clima y cercanía con los manglares.

Los ingresos generados por las exportaciones de camarón representan una parte significativa de las divisas en una economía dolarizada, lo que subraya la importancia del sector en el contexto económico del país. Por otro lado, también existen empresas que producen y exportan, lo que les permite manejar el proceso completa y posiblemente tener mayores ingresos.

Este modelo integrado suele ser más sostenible a largo plazo, ya que estas empresas pueden reaccionar con más rapidez a las demandas del mercado y adaptarse a las exigencias internacionales de calidad y regulaciones. Pero desde la perspectiva de la sostenibilidad, el sector del camarón también es presionado a modernizarse y adoptar prácticas más amigables con el medio ambiente.

La creciente presión por parte de consumidores y reguladores para asegurar que el camarón se críe de manera responsable ha llevado a muchas empresas a invertir en tecnologías más limpias y eficientes. No solo es importante para la integridad de nuestros ecosistemas y recursos, sino que puede ser un diferenciador en un mercado cada vez más competitivo.

2. Problemática

Ecuador se destaca como uno de los mejores productores y exportadores de camarón a nivel mundial y uno de los principales consumidores es China, la producción, comercialización y exportaciones del mismo han crecido de maneras constante desde la década de Los 2000s, sin embargo, estas exportaciones se han visto afectadas debido a diversas restricciones e inconvenientes, teniendo esto en cuenta, existen varios factores elementales que han limitado la capacidad del sector camaronero para aprovechar las oportunidades de expandirse en el mercado Chino , y son Los siguientes:

2.1 Fluctuaciones en la demanda del mercado chino

Aunque China es uno de los compradores principales de camarón en ecuador, ha tenido una demanda bastante inestable, Esta fluctuación se ha visto influenciada debido a múltiples factores como los cambios en los indicadores y variaciones económicas en el país, y a su vez, como han ido cambiando los hábitos de consumo en ciudadanos ecuatorianos.

Dentro de la población de clase media en China se ha podido observar tanto un crecimiento crucial como un cambio en los hábitos alimenticios. Los consumidores ahora optan por buscar productos de alta calidad y frescos, lo que ha llegado a favorecer la adquisición del camarón ecuatoriano, no obstante, este nuevo perfil de consumidor es más riguroso y meticuloso con el origen y la calidad de productos que consumen, en consecuencia, las expectativas no se cumplen, lo que llegado a causar fluctuaciones significativas. La economía china ha experimentado periodos de ralentización, pese a su crecimiento constante, como ocurrió tras la pandemia del COVID-19. Esto redujo la compra de bienes de lujo, como el marisco, lo que redujo la demanda de camarón de importación. Ecuador es uno de los principales productores

y exportadores de camarón en el mundo, y uno de los principales compradores de camarón es China. A pesar de que la producción y exportación han estado creciendo desde el año 2000, las exportaciones al mercado de camarón chino enfrentan varios problemas. Estas barreras han limitado el potencial del sector para aprovechar el mercado chino en su totalidad, sobre todo en cuanto a la constancia y estabilidad de las ventas.

Esto ha llevado a una disminución en la compra de artículos de lujo, como los mariscos, lo que ha provocado un descenso en la demanda de camarón importado. Ecuador es uno de los principales productores y exportadores de camarón en el mundo, y China es uno de sus principales compradores.

A pesar de que la producción y las exportaciones se han disparado, especialmente a partir de la década de los años 2000, las exportaciones de camarón a China han enfrentado numerosas barreras. Estas barreras han impedido que el sector aproveche de forma plena y constante el mercado chino. Los mayores desafíos que afronta el sector camaronero Se puede identificar una serie de importantes retos que experimenta el sector camaronero, desde la inestabilidad de la demanda hasta los problemas logísticos y de calidad, así como la creciente competencia internacional.

Objetivos comerciales y logísticos

El comercio de camarón desde Ecuador hasta China presenta numerosos desafíos logísticos, principalmente, por lo que respecta a transporte, aduana y controles de salud. Estas limitaciones, agravadas por la pandemia, han incrementado los costos y los tiempos de tránsito, lo que finalmente afecta el estado del producto.

Costos de transporte

Los precios del transporte internacional se han incrementado por la subida de los precios de los combustibles, la falta de contenedores y las aglomeraciones en los puertos. Al ser

China un destino lejano, se requiere que el camarón llegue en la mejor condición posible. Este aumento en los costos de logística reduce la competitividad del producto ecuatoriano frente a competidores más cercanos.

Inspecciones aduaneras y de salud.

La importación de alimentos a China, como el camarón, está sometida a estrictos controles aduaneros. Con la llegada del COVID-19, estos controles han sido acentuados, lo que ha ocasionado retrasos en los puertos chinos. Estos largos periodos de espera afectan la frescura del camarón y, por tanto, la percepción de calidad entre los consumidores.

Efecto que la duración tiene sobre la calidad del producto

De acuerdo con el tiempo de tránsito y los retrasos, es posible que el camarón sufra una pérdida en su calidad. Los controles extras en los puertos y una mala administración de la cadena de frío generan pérdidas para el exportador y se traducen en una sensación de insatisfacción entre las compras de China.

La percepción de calidad del camarón de Ecuador en China

Aunque el camarón ecuatoriano es reconocido por su calidad, los consumidores chinos piden una mejor consistencia en el producto. En los últimos años se han identificado residuos químicos en algunos lotes de exportación, generando algunas sospechas.

Contaminación y uso de químicos

El uso de químicos en el proceso de camarón es un problema global que se da en la industria del camarón. Al conocer estos casos, los consumidores de China muestran desconfianza con algunos de los proveedores. Adicionalmente, el gobierno chino ha acentuado sus controles, desechando los lotes que no cumplen con los altos estándares de la seguridad alimentaria.

Rivalidad con producto local

El camarón de Ecuador también se enfrenta al camarón de producción china. La producción local ha aumentado en volumen y en calidad. A pesar de que la producción local no cubre la demanda, el avance en su producción ha reducido la necesidad de importaciones.

3. Justificación

En el comercio internacional actual, las exportaciones de camarón se han convertido en un sector clave para Ecuador, debido a que sientan las bases de un enfoque sostenible, así como para China, que no solo es uno de los primeros importadores del producto, sino también uno de sus mayores competidores.

En este trabajo se hace el pronóstico de las exportaciones de camarón de estos dos países, se analizan las tendencias del mercado, las políticas comerciales y las fluctuaciones de la demanda. La relevancia de este tema se debe a que se ha desarrollado una dependencia económica mutua entre Ecuador y China, en la que las decisiones comerciales de uno afectan directamente al otro. En segundo lugar, el estudio de aquellas variables que influyen en las exportaciones, como las condiciones climáticas, las normativas sanitarias y las dinámicas de consumo, es esencial para entender lo que se puede esperar del sector en el futuro.

Según Muñoz y Carvajal (2024), En los últimos años, la exportación de camarón de Ecuador se ha convertido en uno de los sectores más importantes, consolidando una posición privilegiada en el mercado internacional.

De acuerdo con lo anterior, la presente justificación pretende abordar la importancia de realizar un análisis que permita anticipar las oportunidades y desafíos competitivos de la industria del camarón siendo responsable en el desarrollo sostenible de la misma en Ecuador y la consolidación de China en el mercado mundial.

De esta forma, al proporcionar una visión clara y objetiva de lo que puede esperarse para el sector con la debida prudencia en cuanto a futuros imprevisibles, la presente investigación no solo contribuye a la literatura académica existente sino también se

convierte en una herramienta de apoyo para decisiones de política comercial estratégicas de ambos países

En este contexto, Ecuador ha comenzado un proceso de transformación en la forma en que genera y gestiona sus recursos. Según la Vicepresidencia de la República (2013), se estableció el Comité Institucional para el cambio de la matriz productiva, con el objetivo de modernizar la economía ecuatoriana y reducir su dependencia de la exportación de materias primas sin valor agregado.

Este esfuerzo busca incentivar la producción de bienes con valor agregado, promoviendo la industria local y disminuyendo la importación de productos terminados. La propuesta de tesis que se presenta se centra en el análisis de la capacidad de producción de camarón con destino a China, un mercado que se ha vuelto estratégico para Ecuador debido a su creciente demanda de productos alimenticios de alta calidad.

Según Cañares, Jiménez y González (2021), en el año 2019 el camarón destacó como uno de los productos más valorados en los mercados, no solo por ser uno de los más relevantes en términos de exportación, sino porque sustenta a miles de familias y comunidades ecuatorianas en las regiones costeras de la nación. Es una parte vital de la economía nacional y representa una fuente principal de empleo, generación de riquezas y estabilidad para quienes dependen de esta actividad.

El camarón es una fuente fundamental de proteína. La importancia de este sector es evidente, ya que la industria camaronera representa una parte importante de las exportaciones no petroleras, lo cual subraya la relevancia que tiene para el crecimiento económico de la nación.

A pesar de que Ecuador ha logrado posicionarse en el mercado internacional gracias a la calidad de su camarón, la competencia sigue siendo feroz. Según James y Valderrama (2020), los principales competidores en Asia suelen utilizar antibióticos en sus sistemas de producción, lo que coloca al camarón ecuatoriano en una posición de ventaja competitiva. La razón es que se basa en prácticas de cultivo sostenibles y los altos estándares de calidad en la producción. La capacidad de Ecuador de cumplir con las exigencias de los mercados internacionales será un factor de éxito importante si se desea tener éxito en este sector.

Otra de las razones más importantes por las que estamos llevando a cabo esta investigación se debe a que la diversificación de la producción de camarón se convierte en una necesidad. Abrirse a nuevos mercados, como el chino, se vuelve fundamental para evitar el riesgo de dependencia de un solo cliente, hasta fortalecer la producción nacional.

Esto no solo ayudará a aumentar la producción, sino que también mejorará las condiciones económicas de las comunidades que se dedican a la industria camaronera. Según Domínguez (2019), es esencial que Ecuador continúe invirtiendo en el desarrollo de las capacidades productivas de sus industrias camaroneras y en la introducción de tecnologías para el fortalecimiento de la sostenibilidad del producto. La acuicultura, la cría de camarones en particular, tiene efectos fiscales, pero también actúa como un símbolo de resistencia social y cultural de los asentamientos costeros. Esta investigación busca profundizar en cómo la matriz de producción innovadora puede afectar el comercio de crustáceos y, en consecuencia, la fortuna de la gente.

La bioseguridad y el saneamiento son componentes clave que garantizan tanto la calidad del producto como el bienestar de los hábitats acuáticos y la perdurable viabilidad de las actividades. Ecuador ha tenido una historia de altibajos en la acuicultura, pero a

partir de 2014, la producción de camarón experimentó un resurgimiento y una ola de innovación que mejoró el rendimiento del producto y lo ubicó a nivel internacional. Ecuador produce el ciento por ciento de los camarones en el país, lo que da una gran reputación. Este estudio, por lo tanto, no es solo cuestión de intereses fiscales, sino de un desarrollo sostenible y equitativo para las comunidades y naciones. Si se cultiva adecuadamente bajo regulaciones que promuevan la innovación, la excelencia y la integración social, el cultivo de camarón podría ser el catalizador de la innovación para Ecuador.

Esto permitirá a Ecuador posicionarse de la mejor manera posible en la escena internacional y garantizar un futuro próspero para su comercio de camarones, y el bienestar de su gente. Este estudio no solo busca comprender los desafíos y perspectivas del sector, sino también ayudar a idear estrategias que contribuyan a un mejoramiento estable y un progreso comunitario inclusivo en este esfuerzo tan crucial. Además, es un hecho que integrar la cría de camarones en la nueva estructura de producción ayudará a los productores y la economía del país, pero también a la seguridad alimentaria. Dada la amenaza del calentamiento global y el incremento de la población, mejorar la capacidad de Ecuador para generar alimentos de calidad, accesibles, y sostenibles es vital. Este método no solo tendrá un impacto positivo en la diversificación de la economía, sino que también aumentará la resiliencia de las comunidades frente a futuras incertidumbres económicas.

Por último, el éxito de la transición hacia una economía más diversificada y sostenible depende de la colaboración entre el sector público y privado. Se deben establecer alianzas estratégicas, de las cuales se pueda extraer el conocimiento, tecnología e inversiones que la transición requiere. Las comunidades locales deben ser parte de este proceso para asegurarse de que los beneficios de la producción de camarón se compartan

equitativamente y la calidad de vida de quienes dependen de esta industria mejore. Esta investigación, por su parte, también se encargará de identificar oportunidades para fortalecer estas colaboraciones, para garantizar que la industria camaronera no solo sea rentable, sino también inclusiva y sostenible en el futuro.

Por otro lado, es importante mencionar que el camarón ecuatoriano no solo tiene que enfrentar la competencia de precios de la competencia internacional, sino también competencia en materia de calidad y sostenibilidad, una creciente tendencia a nivel global y que hoy en día se le da mucha importancia a la población.

Debido al avance en la conciencia planetaria, muchas personas se inclinan a consumir productos que provengan de prácticas sostenibles y éticas. Por lo tanto, la industria camaronera del ecuador tiene la posibilidad de destacarse como un modelo de este campo, ejecutando estrategias de promoción y publicidad de elementos ecológicos de la industria. Por otra parte, desarrollar una infraestructura adecuada para el sector es esencial para su crecimiento.

Y mejorar los puertos y la logística de transporte son aspectos clave para que los exportadores de camarón tengan competencia en el mercado. La inversión en tecnología, así como el perfeccionamiento de la capacidad de los trabajadores también son necesarios en el trabajo para mejorar la calidad de la producción y asegurar a la vez que los trabajadores están calificados con las mejores prácticas de cultivo y manejo del entorno. La aplicación de políticas públicas para promover la investigación y desarrollo del sector acuícola puede asegurar que las innovaciones de la industria aumentarán la competitividad.

Los programas de apoyo financiero a pequeños y medianos productores, así como la generación de incentivos para utilizar tecnologías limpias, permitirán un crecimiento más equilibrado y sostenible del sector. La colaboración con universidades y centros de

investigación puede abrirse nuevas oportunidades de desarrollo de productos diferenciados, lo que responde a la demanda del mercado. Finalmente, es fundamental tener en cuenta las preocupaciones ambientales y sociales durante el crecimiento de la industria camaronera. Las prácticas de cultivo deben cumplir con los principios de sostenibilidad para proteger los ecosistemas acuáticos y la salud de las comunidades costeras. La promoción de la acuicultura responsable y la adopción de medidas para proteger el medio ambiente serán esenciales para el futuro de la industria.

La industria de camarones de Ecuador, no solo puede ser un motor de la economía, sino que también puede convertirse en un modelo de sostenibilidad y responsabilidad social. Medio la presente tesis se tratará de predecir y analizar el cómo puede exportarse el camarón a China, de manera que el sector crezca, respetando las necesidades de las comunidades locales y las del medio ambiente. Así, la producción de camarón no solo florecerá, sino que también ayudará a construir un futuro sostenible y justo para todas las comunidades.

4. Objetivos

Objetivo General:

Indagar sobre el Machine Learning y sus algoritmos para su implementación en la predicción de los costos para exportaciones de camarón hacia China con exportadores de Guayaquil.

Objetivos Específicos:

- Definir la importancia del Machine Learning y encontrar el modelo más adecuado para nuestro conjunto de datos.
- Implementar algoritmos de Machine Learning para clasificar a los clientes y predecir los costos de exportación de camarón hacia China basándonos en nuestro conjunto de datos.
- Utilizar los resultados del modelo predictivo para optimizar la planificación logística, reduciendo costos y mejorando la eficiencia en las exportaciones.

5. Marco Teórico

5.1 Machine Learning

5.1.1 Introducción al Machine Learning

Varios expertos consideran que el machine learning es una herramienta esencial hoy en día.

Según (Juarez, 2007 como se citó en Rojas, s.f) ML resuelve situaciones por sí solo a partir de un análisis de datos y cuantos más datos tengan mejores resultados, además, para realizar el análisis se utilizan algoritmos que diseñan otros datos según las necesidades.

También explican que el Machine Learning busca:

"El ML involucra la búsqueda en un amplio espacio de posibles hipótesis, con el fin de determinar cuál es la que más encaja con los datos observados y el conocimiento previo del aprendiz." (Mitchell, 1997)

5.1.2 Definición y tipos de aprendizaje: supervisado, no supervisado y semi-supervisado.

- Según Alpaydin (2021), El aprendizaje supervisado se fundamenta en un grupo de datos clasificados que facilitan la capacitación del modelo para vincular entradas con salidas determinadas. Este método es habitual en actividades como la proyección y la categorización, en las que cada entrada se asocia con una salida deseada, denominada etiqueta.

- El aprendizaje no supervisado implica aprender “propiedades de la estructura subyacente de los datos. En lugar de centrarse en una salida específica, este tipo de aprendizaje identifica patrones latentes en el conjunto de datos, como la agrupación de datos en clústeres o la reducción de dimensionalidad" (Goodfellow et al., 2016, p. 97).
- "El aprendizaje semi-supervisado es un enfoque intermedio entre el aprendizaje supervisado y no supervisado. Utiliza una pequeña cantidad de datos etiquetados junto con una gran cantidad de datos no etiquetados para construir modelos que sean robustos y precisos, especialmente en dominios donde el etiquetado es costoso o impráctico" (Chapelle et al., 2010, p. 15).

Los datos comerciales son parte importante de cualquier negocio, por eso, se nos indica que el machine learning puede ser de gran ayuda para estos datos.

> Según Sekeroglu et al. (2022), "las aplicaciones de aprendizaje automático permiten a las empresas extraer información útil de grandes conjuntos de datos, mejorando la toma de decisiones y optimizando procesos comerciales críticos"

> Según Aggarwal (2020) argumenta también que "el machine learning ha transformado el análisis de datos comerciales al permitir que las empresas extraigan información clave de grandes volúmenes de datos"

En el sector logístico, el poder conocer cómo se comportará un producto en el mercado es importante para los exportadores. Según un estudio de Micocci y Rungi (2021), "los algoritmos de machine learning pueden predecir la probabilidad de que una empresa se convierta en exportadora exitosa, basándose en información financiera y de mercado"

5.1.3 Inteligencia Artificial

"El aprendizaje automático es un componente crucial de la IA, ya que proporciona los métodos necesarios para analizar grandes volúmenes de datos y derivar conocimientos útiles." (Malone, 2018, p. 12).

Para Shah (2020) El aprendizaje automático constituye la columna vertebral de numerosas aplicaciones e implementaciones contemporáneas de inteligencia artificial. Aunque la Inteligencia Artificial se centra en la simulación de habilidades humanas, ML facilita el desarrollo de estas habilidades. a través del estudio de datos y la experiencia acumulada a través del análisis de información y la experiencia adquirida.

5.2 Predicción y Clasificación

Según Padilla-Ospina, Medina-Vásquez y Ospina-Holguín (2020), los métodos avanzados de predicción de aprendizaje automático, como la regresión logística, las máquinas de vectores de soporte, las máquinas de gradiente potenciado, los bosques aleatorios y las redes neuronales, son útiles para el desarrollo de estudios prospectivos. Además, se destaca la importancia de asegurar la robustez y validar dichos modelos de predicción.

Según García, Luengo y Herrera (2015), Los modelos de clasificación automática en las máquinas aprendizaje, tales como los árboles de decisión, las máquinas de soporte

por vectores y las redes neuronales, resultan fundamentales para detectar patrones y efectuar proyecciones exactas.

5.3 Algoritmo K-Means

Según MacQueen (1967) define El algoritmo K-means es una técnica de agrupación iterativa que intenta segmentar un conjunto de datos en k grupos basándose en atributos parecidos, reduciendo al mínimo la cantidad de distancias al centroide de cada grupo.

"RStudio facilita el uso de K-means a través de funciones predefinidas como kmeans(), que permiten especificar el número de clústeres y asignar observaciones con base en distancias minimizadas al centroide del grupo." (R Documentation, 2024)

5.3.1 Fundamentos del K-Means

Clustering

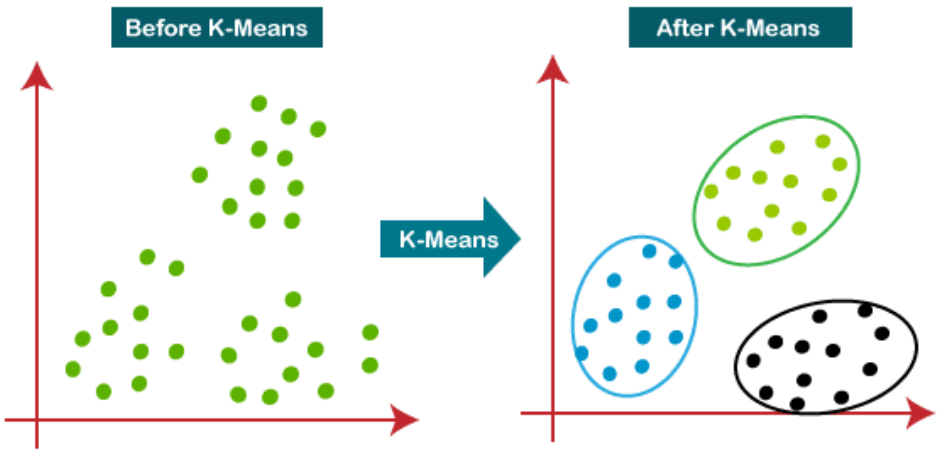
“Es una de las herramientas más importantes en la ciencia de datos. Dado que el algoritmo es sencillo y eficiente, se ha utilizado ampliamente en diversos campos: por ejemplo, en bioinformática, marketing, visión por ordenador, geoestadística, astronomía y horticultura” (Bao Chong, 2021)

Etapas del proceso: inicialización, asignación, actualización y convergencia.

La inicialización según (Harris, 2021) es una partición completamente aleatoria del conjunto de datos, que se denomina Partición aleatoria. También conocidos como

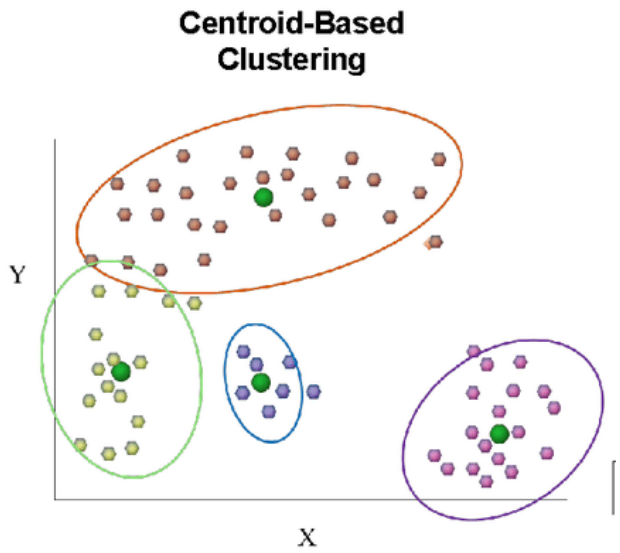
observaciones, del conjunto de datos como centroides iniciales, que denominamos Centroides Aleatorios y que es esencialmente un centroide aleatorio, pero con una modificación por la que los centroides iniciales se recalculan a medida que se les asigna secuencialmente cada punto de datos restantes.

La asignación es donde se suele "encontrar clusters distintos no solapados en los que cada punto se asigna a un grupo. Donde se distribuye cada punto entre los conglomerados o subgrupos más cercanos" ((Zubair et al., 2022)



Según Borlea et al. (2017) el enfoque de actualización del centroide formulado como algoritmo e incluido en el algoritmo k-means reduce el número de iteraciones necesarias para realizar un proceso de agrupación, lo que se traduce en una reducción del tiempo necesario para procesar un conjunto de datos.

También se nos indica que antes de actualizar los centroides es importante cerciorarse de que los nuevos centroides conducirán a una mejor agrupación. "Para ello, se calcula la suma de las distancias al cuadrado entre cada punto y su centroide asignado y se compara con la iteración anterior. Si la suma disminuye, se considera que los nuevos centroides son mejores." (Borlea et al., 2017)



Finalmente, Borlea et al. (2017) habla sobre la repetición, es un proceso donde los pasos de asignación de clusters y actualización de centroides iterativamente hasta que las asignaciones de clusters dejen de cambiar o se alcance el número máximo de iteraciones. Este proceso asegura que el algoritmo converge hacia una solución estable

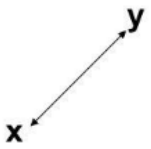
5.3.2 Métrica de distancia y centroides

Según Jain (2010), "el algoritmo K-means utiliza la distancia euclidiana para asignar puntos de datos a los centroides más cercanos"

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Explicación de las métricas más comunes: distancia euclidiana.

Según Lifeder (2019), define la distancia euclidiana de la siguiente manera: es un número positivo que indica la separación que tienen dos puntos en un espacio donde se cumplen los axiomas y teoremas de la geometría de Euclides La distancia entre dos puntos A y B de un espacio euclidiano es la longitud del vector AB perteneciente a la única recta que pasa por dichos puntos.



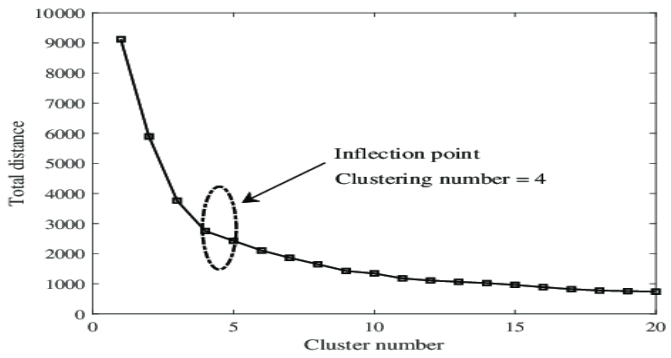
5.3.2.1 Rol de los centroides en la agrupación.

Según Tan et al. (2006) Los centroides son puntos importantes que establecen el núcleo de cada cluster. En el proceso iterativo, el algoritmo ajusta los centroides con el fin de reducir la suma de las distancias entre los puntos de datos y sus centroides asignados, lo que conduce a una definición más precisa de los clústeres.

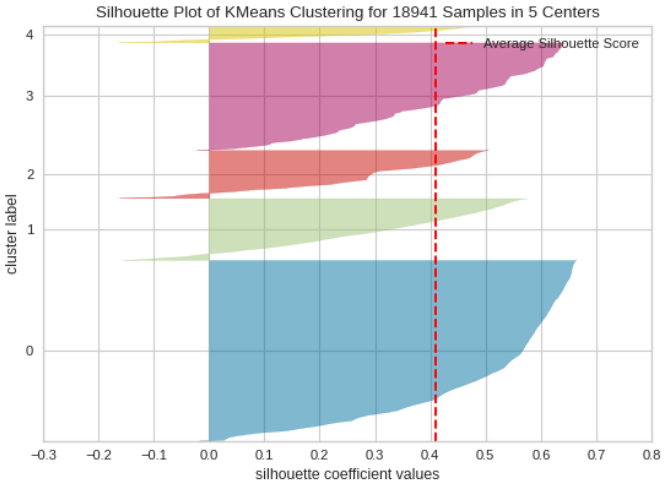
Elección del número óptimo de clusters (K)

Métodos de selección de K: Método del codo (Elbow Method), Silhouette Score y Gap Statistics.

Thorndike (1953) describe la importancia de la clasificación jerárquica y menciona que: "La elección del número de grupos es una decisión crítica en el análisis de agrupamiento, y métodos como el del codo ayudan a determinar un número adecuado observando los cambios en la varianza dentro de los grupos"



Rousseeuw (1987) sostiene que "el coeficiente de silueta proporciona una medida de cómo se ajusta un objeto a su propio clúster en comparación con otros clústeres"



Liu y Wang (2023) argumentan que:

"La combinación de las estadísticas Gap con algoritmos como K-means no solo mejora la precisión del agrupamiento, sino que también proporciona un método robusto para determinar el número óptimo de clústeres en aplicaciones de minería de datos"

5.3.3 Impacto de K en la interpretación de los resultados.

"La selección del número de clústeres K influye significativamente en la interpretación de los resultados, ya que un valor inadecuado puede llevar a conclusiones erróneas sobre la estructura de los datos" (Liu & Wang, 2023)

"Un número incorrecto de clústeres puede no solo distorsionar la representación visual de los datos, sino también afectar las decisiones basadas en esos resultados, lo que subraya la importancia de elegir K adecuadamente" (Tibshirani, Walther, & Hastie, 2001, p. 412).

Según Rousseeuw (1987), La selección del número de clústeres K es crucial ya que establece la forma en que los datos se agrupan y, por consiguiente, influye en las conclusiones que se pueden derivar del análisis.

5.3.3.1 Problemas comunes como sensibilidad a los valores iniciales y datos atípicos.

Cuando se utilizan centroides iniciales seleccionados aleatoriamente, diferentes ejecuciones del K-means pueden producir distintos valores de SSE, lo que indica que la elección de los centroides iniciales es una pieza clave en el rendimiento del algoritmo" (Bravo-Marquez, 2024)

"Los valores atípicos son especialmente problemáticos para K-means, ya que pueden sesgar los resultados al influir desproporcionadamente en la posición de los centroides" (IBM, 2024).

"Es crucial eliminar o manejar adecuadamente los outliers antes de aplicar K-means, ya que estos pueden producir centroides no representativos y llevar a un agrupamiento ineficaz" (Bravo-Marquez, 2024)

5.3.3.2 Limpieza y preparación de datos

Según Dagnino (2023):

"No hay ninguna forma totalmente satisfactoria para el manejo de los datos faltantes, por lo que se debe ser estricto en optimizar la recolección y registro de datos en la etapa de diseño y ejecución. Las alternativas incluyen omitir variables o individuos con datos faltantes, o imputar los datos faltantes utilizando valores predichos"

Dentro del Data cleaning existen variedad de técnicas para poder tener una base de datos funcional.

Como se menciona en el trabajo de Freire Míguez (2023), "la elección del método para tratar los datos faltantes depende del tipo de datos y del contexto del análisis, siendo la imputación una técnica comúnmente utilizada para aproximar estos valores"

"Los valores atípicos pueden distorsionar significativamente los resultados del análisis, por lo que su identificación y tratamiento son esenciales" (Freire Míguez, 2023)

5.3.4 Técnicas de limpieza: tratamiento de datos faltantes y valores atípicos.

5.3.4.1 Tratamiento de Datos Faltantes

"Las opciones para tratar los datos faltantes incluyen omitir registros, imputar valores o eliminar columnas enteras" (Freire Míguez, 2023, p. 5).

5.3.4.2 Identificación y Tratamiento de Valores Atípicos

Según AWS (2024):

"Los valores atípicos repercuten sustancialmente en el rendimiento del modelo, por lo que es importante identificarlos y determinar las medidas apropiadas para su tratamiento"

5.3.4.3 Eliminación de Duplicados

En la guía de Solex (2024), se menciona que:

"Eliminar observaciones duplicadas es uno de los primeros pasos en el proceso de limpieza de datos, ya que los duplicados pueden llevar a conclusiones incorrectas"

5.3.4.4 Normalización y escalado de las variables.

La normalización, transforma los datos para que se encuentren dentro de un rango específico, generalmente entre 0 y 1. Este método es útil cuando se requiere que todas las características contribuyan de manera equitativa al análisis.

La normalización convierte los valores de las características en un rango uniforme, lo que resulta crucial para muchos algoritmos de aprendizaje automático. (Morante, 2024).

La estandarización implica centrar los datos alrededor de la media y escalarlos según la desviación estándar.

" La normalización hace posible que los rasgos posean una media de cero y una desviación estándar de uno, lo que es esencial para numerosos modelos estadísticos." (Morante, 2024).

Jiménez (2023) explica que:

" El robusto método de escalado ajusta los datos empleando la mediana y el rango, lo que facilita la gestión de valores atípicos sin alterar la distribución de los datos." (p. 4).

5.4 Fundamentos del Algoritmo k-means

El algoritmo k-means es un método muy empleado en el aprendizaje no supervisado para agrupar datos en grupos homogéneos. Su meta es reducir la totalidad de las distancias cuadradas entre cada punto de datos y el centroide de su clúster designado, consiguiendo de esta manera una agrupación eficaz que mejora la vairanza intra-clúster. (Arthur & Vassilvitskii, 2007).

El proceso implica la selección inicial de k centroides, la asignación de cada punto al centroide más cercano y la actualización iterativa de las posiciones de los centroides hasta alcanzar la convergencia (Oyelade et al., 2010).

Aunque k-means garantiza la convergencia, los resultados dependen de la inicialización y la elección del número de clusters (k), ya que la solución es local (Arthur & Vassilvitskii, 2007).

Métodos como k-means potencian notablemente el proceso al elegir de forma estratégica los puntos iniciales, lo que disminuye la posibilidad de obtener resultados no óptimos. (Arthur & Vassilvitskii, 2007).

Utilidad del KM para la predicción

El k-means puede tener una gran utilidad en el campo de la predicción, ya que el algoritmo es rápido y eficiente para procesar grandes cantidades de datos, con lo que es posible hacer una segmentación y posterior análisis de los mismos (Arthur & Vassilvitskii, 2007). Además, K-means permite descubrir patrones ocultos en los datos, lo que lo hace muy útil para tareas como segmentar bases de datos de clientes, extrapolar patrones de éxito en una base de datos de exportación y agrupar a los exportadores por comportamiento y datos pasados (Oyelade et al., 2010).

El k-means en la predicción

El éxito de k-means en la predicción depende de muchos factores, como: 4. 5. 2. Elementos necesarios para su correcto uso El preprocesamiento de datos (limpieza, normalización y transformación) es un factor clave para evitar sesgos en los datos por valores atípicos u errores

en los mismos (Chu, 2017), la selección del número de clusters (k) también es un factor, el cual puede ser determinado por una técnica como el codo o la validación cruzada, para garantizar que los clusters generados sean significativos (Oyelade et al., 2010). Sin embargo, la variante k-means ++ es esencial para mejorar la unidad de los clusters, ya que se compensa la falla de k-means en el uso de una inicialización aleatoria (Arthur & Vassilvitskii, 2007).

Interpretabilidad de resultados

Una vez que los clusters han sido calculados, se debe analizar la pertinencia de cada uno de ellos, es decir, el analisis de centroides, los cuales representan el perfil promedio de los puntos dentro de un cluster, de manera que es posible identificar las características dominantes en cada cluster (Oyelade et al., 2010). Las medidas de validación del resultado de K-Means permiten confirmar la calidad de los clusters generados, entre las medidas de validación más comunes se encuentra el índice de silhouette y la suma de errores al cuadrado (SSE) (Arthur & Vassilvitskii, 2007). Los patrones que se obtengan al aplicar el k-means pueden integrarse a un modelo supervisado de aprendizaje automático para tomar decisiones con campañas de marketing, estrategias de negocio, proyecciones de ventas, etcétera (Chu, 2017).

Aplicación del K-Means como herramienta de predicción

En un contexto como el de exportaciones desde Guayaquil a China, k-means puede ayudar a: Segmentar los exportadores, (volumen de exportación-frecuencia de exportación), productos, mercados destino Identificar clusters de alto rendimiento y extraer patrones de éxito Predecir volúmenes de exportación futuras del cluster, en base a los datos historicos de los clusters. Oyelade et al. (2010) se usó el algoritmo k-means para analizar la predicción del rendimiento, agrupando a los estudiantes en clusters en base a sus calificaciones y comportándose de manera parecida a los datos de los clientes. Este enfoque también puede aplicarse en el caso de los exportadores de la industria del

camarón para identificar los patrones de alto rendimiento y extraer de ellos los patrones de éxito para el crecimiento de la industria de exportaciones del camarón de Ecuador

6. Marco Conceptual

El aprendizaje automático (ML) es una parte de la IA que permite a las máquinas aprender patrones de grandes cantidades de datos sin necesidad de un humano que le instruya cómo hacerlo. En vez de programarse para realizar una tarea, las máquinas pueden aprender patrones, tender a hacer predicciones, tomar decisiones y realizar mejoras a lo largo del tiempo, que son las acciones de aprendizaje que ocurren cuando se ingresa más información a la máquina (O'Neil, 2016). Esta habilidad de aprender de los datos se convierte en un elemento fundamental en la economía, la salud, los negocios, las ciencias sociales, entre otras. De modo que permite la realización de tareas de clasificación, predicción, segmentación de tal forma, que la eficiencia de los procesos de toma de decisiones puede ser tanto profesional como científico (Bishop, 2006).

6.1 Aprendizaje automático e inteligencia artificial

En general, la inteligencia artificial intenta imitar los procesos cognitivos humanos en los sistemas informáticos. Esto incluye la capacidad de pensar, aprender y adaptarse a nuevas situaciones y tomar decisiones sin intervención humana directa (Russell & Norvig, 2010). Dentro de este campo, el aprendizaje automático se ha consolidado como un subcampo centrado en desarrollar algoritmos que permitan a las máquinas aprender a partir de datos. Por lo tanto, el aprendizaje automático busca mejorar el rendimiento de la máquina a lo largo del tiempo utilizando los datos disponibles para descubrir patrones y hacer predicciones.

Una de las técnicas de aprendizaje automático más importantes es el aprendizaje supervisado. Si bien el aprendizaje supervisado implica el uso de datos de entrenamiento que contienen tanto entradas como resultados esperados, el aprendizaje supervisado generalmente se usa para descubrir características o patrones relacionados con datos sin etiquetas previas. Por tanto, la agrupación es un proceso en el que los algoritmos clasifican puntos de datos en diferentes grupos de datos. JAIN 45794, requerido para instrucción no supervisada.

El algoritmo K-means es uno de los métodos de agrupación más utilizados. Este algoritmo clasifica el conjunto de datos en k grupos de modo que los puntos de un grupo sean lo más similares posible entre sí y los puntos de los otros grupos sean lo más similares posible entre sí. La similitud normalmente se mide utilizando la distancia euclidiana, una métrica que mide la "distancia" entre dos puntos en un espacio multidimensional (McQueen, 1967). K-means es especialmente útil en situaciones en las que no se dispone de una clasificación previa y es necesario clasificar una gran cantidad de datos para obtener información relevante (Chandra, 2017).

Algoritmo K-Means: concepto y funcionamiento básicos El algoritmo K-Means es un proceso iterativo que consta de dos pasos principales: Asignación de puntos a conglomerados: en este paso, cada punto de datos se asigna a un conglomerado cuyo centroide (centro del conglomerado dentro del conglomerado) es más cercano a es. Esto se hace usando la distancia euclidiana como medida de similitud.

Recalcular centroides: después de asignar todos los puntos, los centroides de cada grupo se recalculan promediando los puntos de ese grupo. Este proceso se repite hasta que los centroides cambien significativamente,El algoritmo indica convergencia (McQueen, 1967).

6.2 Elegir el número de clusters (k)

Una de las decisiones más importantes al implementar un algoritmo de k-medias es decidir cuántos clusters (k) usar. Aunque este valor debe ser elegido por el analista, existen métodos que ayudan a elegir el número óptimo de grupos. El método del codo es uno de los métodos

más utilizados. Este método calcula la suma de errores cuadráticos (SSE) para diferentes valores de k y elige el valor de k donde la disminución en SSE comienza a estabilizarse, denominado en el gráfico la “rodilla” (Thorndike, 1953). Otra técnica utilizada es el índice de silueta, que mide el grado de formación de grupos comparando la distancia promedio entre puntos de un grupo con la distancia promedio entre puntos de otros grupos (Roscio, 1987).

6.3 El Algoritmo K-Means: Fundamentos y Funcionamiento

El algoritmo K-Means es un proceso iterativo que se compone de dos fases principales:

6.3.1 Asignación de puntos a clústeres:

En esta fase, cada punto de datos se asigna al clúster cuyo centroide (el centro de masa del clúster) está más cercano, utilizando la distancia euclidiana como medida de similitud.

6.3.2 **Reajuste de los centroides**:

Una vez que todos los puntos han sido asignados, el centroide de cada clúster se recalcula tomando el promedio de los puntos dentro de ese clúster. Este proceso se repite hasta que los centroides no cambian significativamente, lo que indica que el algoritmo ha convergido (MacQueen, 1967).

6.4 Selección del Número de Clústeres (k)

Una de las decisiones más críticas al aplicar el algoritmo K-Means es determinar cuántos clústeres (k) se deben usar. Aunque el valor de k debe ser elegido por el analista, existen métodos que ayudan a determinar el número óptimo de clústeres. El Método del Codo es uno de los más usados. En él se calcula la Suma de Errores Cuadráticos (SSE) para diferentes valores de k, y se elige el valor de k donde la reducción en la SSE comienza a estabilizarse, lo que se ve como un "codo" en el gráfico (Thorndike, 1953). Otra técnica comúnmente utilizada es el Índice de Silueta, que mide qué tan bien está formado un clúster comparando la distancia promedio de los puntos del clúster con la distancia promedio hasta los puntos del clúster más

cercano (Rousseeuw, 1987). 5.4. 1 Métricas de distancia y centroides El hecho de utilizar la distancia euclidiana es de suma importancia en el algoritmo K-Means. Este permite cuantificar la similitud entre dos puntos de un espacio de dimensión k. La fórmula para calcular la distancia euclidiana entre dos puntos "x" y "y" en un espacio n-dimensional es:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Esta clase de distancia es muy útil cuando los datos son de naturaleza continua, por ejemplo en el caso de las características de un consumidor o del estilo de compra (Bishop, 2006). El centroide de un clúster se obtiene calculando la media de las coordenadas de todos los puntos del clúster. A través de las iteraciones de los algoritmos, cada centroide se renueva para dar minimizar la variabilidad de los puntos dentro de cada clúster, mejorando así la precisión de la segmentación (Chandra, 2017).

5.4. 2 Ventajas y desventajas del K-Means

Algunas de las principales ventajas del algoritmo K-Means son la simplicidad, eficiencia computacional y facilidad. Esto lo hace uno de los métodos más utilizados para el clustering de grandes volúmenes de datos (Jain, 2010). Sin embargo, también tiene limitaciones, como la elección del número de clústeres (k), que debe definirse antes de ejecutar el algoritmo. Además, el rendimiento de K-Means puede verse afectado por la inicialización de los centroides, ya que una mala selección inicial puede llevar a soluciones subóptimas. Para mejorar esta inicialización, se ha propuesto el uso de K-Means++ (Arthur & Vassilvitskii, 2007).

K-Means suele ser sensible a Outliers; también pueden alterar enormemente la ubicación del centroide debido a la estructura de la distancia euclidiana: cuando el conglomerado no esférico,

y los tamaños son muy diferentes, se necesita un algoritmo de agregación, como DBSCAN: Density-Based Spatial Clustering of Applications with Noise sincete al., 1996).

6.4.1 Métricas y Validación de Modelos

La evaluación de un modelo de clustering es una etapa crítica para confirmar que la segmentación final tiene sentido y es válida para tomar decisiones. Se emplean otras medidas además de la a *Suma de Errores Cuadráticos* (SSE) para indicar cuán bien se aplica un modelo en los casos. Otros indicadores, tales como el Índice de Silueta, se utilizan para indicar la calidad de clustering. Un buen coeficiente Silueta es igual o cercano a 1, si el coeficiente es cercano a -1, eso indica un mal coeficiente de clustering. Si el coeficiente Silueta es 0, entonces los clústeres superpuestos se pueden superponer (Rousseeuw, 1987). La validación cruzada también es necesaria en los modelos supervisados, aunque esto no siempre se aplica debido a que los datos no están etiquetados en el caso de la agrupación (Bishop, 2006).

6.4.2 Impacto del Algoritmo K-Means en las Exportaciones

El algoritmo K-Means tiene un gran potencial en el sector de las exportaciones, en especial para la segmentación de mercados. Al agrupar mercados según características similares, como el comportamiento de consumo o la demanda de productos, las empresas pueden personalizar sus estrategias de comercialización y optimizar sus procesos de ventas. En el caso específico de las exportaciones de camarón, este tipo de segmentación puede ser crucial para identificar patrones de demanda en diferentes regiones del mundo y anticipar fluctuaciones en el mercado (Chandra, 2017).

6.4.3 Aplicaciones del Algoritmo K-Means

El K-Means tiene una amplia gama de aplicaciones:

- **Segmentación de Mercado**: Utilizado para agrupar clientes según comportamientos similares, lo que permite a las empresas ofrecer productos adaptados a las necesidades específicas de cada grupo (Jain, 2010).
- **Análisis de Comportamiento del Usuario**: En comercio electrónico y redes sociales, K-Means se usa para personalizar recomendaciones basadas en los intereses y patrones de compra de los usuarios (Revista Tecnológica, 2018).
- **Detección de Anomalías**: En ciberseguridad y transacciones financieras, el algoritmo puede identificar patrones anómalos que podrían indicar fraudes o actividades sospechosas (Chandra, 2017).
- **Agrupamiento Geoespacial**: También se usa para analizar zonas geográficas, como la segmentación de mercados urbanos o rurales, a partir de datos demográficos o infraestructurales (Revista Tecnológica, 2018).

7. Marco legal

La normativa que regula la exportación de alimentos procesados, como el camarón, se encuentra dentro de diversos acuerdos y leyes nacionales e internacionales que buscan asegurar la calidad, seguridad y legalidad de los productos. Tanto en Ecuador, como en otros países, depende de las leyes de salud, comercio exterior y regulaciones fitosanitarias. La Ley Orgánica de Salud establece las bases y directrices para que los productos de seguridad alimenticia procedan a la industria procesados, comercio. Además, se encuentra el cumplimiento de los estándares nacionales e internacionales, como una de sus principales características (González, 2020). La normativa del comercio exterior, mientras tanto, queda reflejada en la Ley de Fomento a la Competitividad y Comercio Exterior.

Esta ley se destaca por el señalamiento del marco legal para la facilitación del comercio exterior, regulando las condiciones para las exportaciones. En consecuencia, esta ley también promueve la competitividad en los mercados internacionales (Cevallos, 2019). Sin embargo, también es indispensable la promulgación de acuerdos internacionales, como la Organización Mundial del Comercio, que busca la armonización de las normativas sanitarias comerciales entre los países (Rodríguez, 2018). 6. 1 Ley Orgánica de Protección de Datos Personales Dentro del análisis de los datos para la predicción de exportación de camarón a China a través del uso de Machine Learning es importante considerar la Ley Orgánica de Protección de Datos Personales (LOPDP) que obliga el uso, tratamiento y protección de los datos personales en Ecuador. Dicha ley que data del año 2021, estipula que toda persona tiene derecho a la protección de sus datos personales, la recolección y el tratamiento de sus datos (Asamblea Nacional del Ecuador, 2021).A su vez, es necesario asegurarse de que los datos personales utilizados, como los relacionados con clientes, proveedores o transacciones comerciales, sean

manejados conforme a los principios de transparencia, legalidad y seguridad establecidos en la LOPDP.

Es importante que, si el análisis de datos involucra información personal, se obtenga el consentimiento explícito de los titulares de esos datos o bien se justifique el uso de estos bajo el mandato de la ley (Asamblea Nacional del Ecuador, 2021). La Ley también establece que los titulares de los datos tienen derecho a acceder, rectificar, cancelar y oponerse al tratamiento de sus datos personales, lo cual debe ser respetado en todos los procesos relacionados con la predicción de exportaciones. La implementación de estas normativas garantiza no solo el cumplimiento de la ley, sino también la confianza y transparencia en las relaciones comerciales.

7.1 Finalidad y Uso de los Datos Personales

En cuanto al uso de datos personales dentro de los Machine Learning para la predicción de exportación, el artículo 27 del Código Orgánico de la Economía Social de los Conocimientos (Cámara de Comercio de Quito, 2016) establece que en el territorio de la República de Ecuador, el tratamiento de datos personales será informado y utilizado para fines estadísticos o científicos. En ese sentido, las entidades que manejan los datos deben garantizarse que los usuarios estén debidamente informados de cómo se utilizarán sus datos, sobre todo si tales datos forman parte de bases de datos utilizadas para entrenar los modelos de predicción. Por ejemplo, las transacciones comerciales de camarón de la industria ecuatoriana a China pueden ser utilizados para identificar patrones de consumo, hacer predicciones de variaciones en la demanda y optimizar las estrategias de exportación. No obstante, los datos únicamente pueden ser utilizados para razón de tales finales y no se pueden compartir con terceros sin el consentimiento de los afectados, según lo estipulado en la LOPDP. 6. 3 Derechos de los Titulares de los Datos Los titulares de los datos personales dentro del análisis de exportación de camarón

tienen garantizados los derechos de acceder, rectificar y eliminar los datos conforme la actual normativa ecuatoriana. El artículo 15 de la Ley Orgánica de Protección de Datos Personales sirve para garantizar el derecho de los titulares de los datos de acceder a sus datos, corregir datos erróneos y eliminarlos si ya no son necesarios para el propósito para el cual fueron recopilados (Ministerio de Telecomunicaciones, 2019). Esto es particularmente relevante cuando se manejan datos en tiempo real que pueden cambiar con frecuencia, como en el caso de las exportaciones.

Por otro lado, los titulares deben ser informados sobre el uso de técnicas de Machine Learning que puedan implicar la toma de decisiones automatizadas sobre los datos personales, lo cual se alinea con las directrices internacionales del Reglamento General de Protección de Datos de la Unión Europea (GDPR), en donde se establece la transparencia en cuanto al uso de algoritmos para decisiones automatizadas (European Commission, 2020).

7.2 Buenas prácticas de manufactura (BPM)

La ley también destaca la importancia de cumplir con las BPM en la producción de alimentos, que es el conjunto de normas que los productores deben seguir para garantizar la adecuada higiene y seguridad en las instalaciones y procesos de fabricación. Estas BPM son los principios y directrices que deben seguirse para prevenir la contaminación de los productos y se deben mantener las condiciones que se requieren para su consumo. son requisitos tanto para la venta en el mercado local como para exportar a otros países(Asamblea Nacional del Ecuador, 2015). Según la ley, los productos alimenticios procesados que quieren exportar deben obtener una certificación que indique que cumplen con las BPM y otras normativas de calidad de la autoridad sanitaria. Este la certificación es parte del registro y notificación sanitaria, y es un

requisito para la autorización de los productos para exportar. Es más, el incumplimiento de las BPM puede resultar en la cancelación de los registros sanitarios, lo que significa que los productos no pueden ser vendidos, ni en el mercado nacional ni en el extranjero.

Bases legales del tratamiento de datos personales

El tratamiento de datos personales en el contexto de ML para pronóstico de las exportaciones de camarón debe estar exento en una base legal clara, en el caso de ecuador la base legal es el Código Orgánico de la Economía Social de los Conocimientos que determina que las instituciones públicas y privadas deben cumplir con los principios de protección de datos personales cuando los datos, son usados para fines científicos o estadísticos. (Cámara de Comercio de Quito, 2016). En este caso, los datos sobre las exportaciones y los patrones de consumo en China pueden ser utilizados para desarrollar modelos predictivos que beneficien a los exportadores, siempre bajo la premisa de que se respete la confidencialidad y el consentimiento de los implicados.

El **acuerdo ministerial 012-2019** del Ministerio de Telecomunicaciones, que proporciona directrices sobre la **protección de datos personales en la administración pública**, también debe ser tenido en cuenta si los datos son recopilados por una entidad pública que maneje este tipo de información (Ministerio de Telecomunicaciones, 2019).

8. Metodología

8.1 Diseño Metodológico

Para esta investigación se lleva a cabo la implementación de modelos de ML, en este caso el algoritmo k-means, para el pronóstico de las exportaciones de camarón de Guayaquil a China. se toma una perspectiva de enfoque cuantitativo, los datos son los históricos de exportación de los que se utilizarán técnicas de aprendizaje no supervisado. El diseño metodológico se basa en las siguientes etapas:

8.1.1 Criterio de la Suma de Cuadrados (SSQ)

El criterio principal de K-Means es la minimización de la varianza dentro de los clústeres, que se puede definir como la suma de cuadrados de las distancias de cada punto de datos a su centroide correspondiente.

8.1.2 Criterio Discreto SSQ (Suma de Cuadrados Discreta)

Dado un conjunto de datos X = $\{x_1, x_2, \ldots\ x_n\}$ en un espacio $\mathbb{R}^p$(donde p es la dimensión de los datos), y una partición C = $\{C_1, C_2, \ldots\ C_k\}$de los datos en K clústeres, el objetivo es minimizar el siguiente criterio de la suma de cuadrados:

$$J(C) = \sum_{i=1}^{k} \sum_{x_j \epsilon c_i} \left\| x_j - c_i \right\|^2$$

Aquí, c_i es el centroide del clúster C_i y $\left\| x_j - c_i \right\|$ es la distancia euclidiana al cuadrado entre el punto de datos x_i y el centroide c_i. Este criterio mide la "inercia" o variabilidad dentro de los clústeres, y el objetivo es minimizarlo para obtener una partición óptima de los datos.

Este mismo criterio puede ser formulado como un problema de optimización con respecto tanto a la partición C como a los centroides Z, es decir:

$$J(C,Z) = \sum_{i=1}^{k} \sum_{x_{je}c_i} \|x_j - Z_i\|^2$$

Donde Z = $\{z_1, z_2, \ldots\ z_k\}$ son los K centroides a optimizar.

Según Bock (2020), el problema de optimización es equivalente a buscar los centroides Z_i tal que minimicen la distancia dentro de cada clúster c_i . Esto se puede expresar de la siguiente manera (Bock, 2020) ^[2]:

$$Min_{c,z} = \sum_{i=1}^{k} \sum_{x_{je}c_i} \|x_j - Z_i\|^2$$

Este es el principio subyacente del algoritmo K-Means: asignar los puntos de datos a clústeres de manera que la suma de las distancias cuadradas de los puntos al centroide de su clúster sea mínima.

Recalculo de Centroides

El centroide $c_i$de un clúster $C_i$es el promedio de los puntos en ese clúster. Matemáticamente, esto se expresa como:

$$ci = \mid Ci \mid 1xj \in Ci\sum xj$$

Donde Ci es el número de elementos en el clúster.

8.2 Algoritmo K-Means: Desarrollo Iterativo

El algoritmo K-Means se desarrolla en un ciclo iterativo que alterna entre dos pasos fundamentales: **asignación de puntos a clústeres** y **actualización de centroides**. Estos pasos son los siguientes:

1. **Asignación de puntos a clústeres**: Para cada punto de datos x_j se asigna al clúster cuyo centroide $c_i$sea el más cercano. Es decir, el punto x_j se asigna al clúster $c_i$que minimiza la distancia $\| x_j - c_k \|^2$

$$C_i = \{x_j : \| x_j - c_i \|^2 \leq \| x_j - c_k \|^2, \forall \mathrm{k} \neq \mathrm{i}\}$$

   En este paso, la función de asignación es responsable de dividir el espacio de datos en K clústeres, de acuerdo con los centroides actuales.

2. **Recalculo de los centroides**: Una vez que los puntos han sido asignados a sus respectivos clústeres, el siguiente paso es actualizar los centroides. Como se mencionó anteriormente, el centroide c_i de cada clúster se calcula como el promedio de los puntos dentro del clúster:

$$c_i = | C_i | 1 x_j \in C_i \mathrm{i} \sum x_j$$

   Este proceso se repite iterativamente hasta que los centroides ya no cambian significativamente entre dos iteraciones consecutivas.

8.3 Convergencia del Algoritmo

El algoritmo K-Means converge cuando las asignaciones de los puntos a los clústeres no cambian entre iteraciones. Matemáticamente, esto se puede expresar como:

$$C^{(t+1)} = C^t$$

Donde C^t es la partición de los puntos en clústeres en la iteración t, y la convergencia se alcanza cuando las asignaciones de los puntos a los clústeres no varían entre dos iteraciones consecutivas.

Sin embargo, como señalan **Ahmed et al. (2020)**, el algoritmo K-Means no garantiza una convergencia global óptima debido a su dependencia de la inicialización de los centroides, lo que puede llevar a una convergencia a un mínimo local (Ahmed et al., 2020)

El algoritmo **K-means** busca particionar un conjunto de datos en KKK clusters, minimizando la suma total de las distancias cuadradas entre los puntos de datos y sus respectivos centroides. El objetivo es minimizar la **suma de los errores cuadrados** (SSE):

$$SSE = \sum_{\{k=1\}}^{\{K\}\sum_{\{i=1\}}^{\{n_k\}|}} x_i - c_k|^2$$

donde:

- x_j es el punto de datos i,
- c_k es el centroide del cluster **K**,
- k_n es el número de puntos asignados al cluster **K**,

- **K** es el número total de clusters.

Inicialización de los Centroides

La inicialización de los centroides c_k se realiza generalmente de manera aleatoria. Algunas estrategias incluyen:

1. Selección aleatoria de **k** puntos del conjunto de datos.
2. Selección de puntos distantes entre sí.

Paso 1: Asignación de Puntos al Cluster

Para cada punto x_i , se asigna al cluster **k** cuya distancia al centroide c_k es mínima. Esto se hace utilizando alguna métrica de distancia, como la **distancia Euclidiana**:

$$d_{E(x_i,c_k)} = \sqrt{\left\{\sum_{\{j=1\}}^{\{d\}} \left(x_{\{ij\}} - c_{\{kj\}}\right)^2\right\}}$$

donde:

- $x_{(ij)}$es la j-ésima característica del punto x_i
- $c_{(kj)}$ es la j-ésima característica del centroide $c_{(k)}$
- d es el número de dimensiones.

Paso 2: Actualización de los Centroides

Una vez asignados los puntos a los clusters, el centroide $c_{(k)}$ de cada clúster se actualiza como la media de los puntos asignados:

$$c_k = \frac{1}{n_k} \sum_{i \in S_k} x_i$$

donde $s_{(k)}$ es el conjunto de puntos asignados al clúster **k**.

Iteración

El algoritmo alterna entre asignar puntos a los centroides más cercanos y recalcular los centroides hasta que la suma de las distancias no cambie significativamente, o se alcance el número máximo de iteraciones.

Fórmula General del Proceso

La iteración del algoritmo puede describirse como:

1. **Asignación de puntos** a los centroides más cercanos:

$$x_i \in \backslash text\{argmin\}_k | \, x_i - \, c_k|^2$$

Actualización de los centroides:

$$c_k^{\{(new)\}} = \backslash frac\{1\}\{n_k\} \sum_{\{i \in S_k\}} x_i$$

Variantes de K-means

Algoritmo de Lloyd (1957)

El algoritmo de Lloyd, distingue puntos asignando a los centroides y sus correspondientes actualizaciones de esta manera como se mencionó anteriormente.

Algoritmo de MacQueen (1967)

El algoritmo de **MacQueen** realiza actualizaciones incrementales de los centroides en cada paso. Si un punto cambia de cluster, el centroide de ambos clusters se actualiza de inmediato:

$$c_k^{\{(new)\}} = \backslash frac\{1\}\{n_k\} \sum_{\{i \in S_k\}} x_i$$

Algoritmo de Hartigan & Wong (1979)

Este algoritmo intenta minimizar el **SSE** dentro de cada cluster, permitiendo que un punto cambie de cluster si reduce el error cuadrático total:

El algoritmo evalúa si mover un punto a un nuevo cluster reduce el **SSE** total.

Métricas de Distancia

Distancia Euclidiana:

$$d_{E(x_i,c_k)} = \sqrt{\left\{ \backslash sum_{\{j=1\}}^{\{d\}} (x_{\{ij\}} - c_{\{kj\}})^2 \right\}}$$

La distancia euclidiana, un concepto originado en las matemáticas griegas y popularizado por Euclides, es necesaria para unir la geometría clásica y los métodos algorítmicos modernos, consiste en una medida de distancia de línea recta entre dos puntos en un espacio, ya sea 2D o de múltiples dimensiones. Este método tiene una amplia aplicación, desde la ciencia de datos hasta el aprendizaje automático y el análisis espacial.

El mundo real de ejemplo en Python y R te ayudará a entender cómo puedes utilizar esta idea para resolver problemas y mejorar tu eficiencia en una variedad de campos. La distancia euclidiana es la distancia más corta entre dos puntos en el espacio. Su base es el teorema de Pitágoras, que establece en un triángulo rectángulo, la suma de los cuadrados de los dos catetos es igual al cuadrado de la hipotenusa. El álgebra lineal es el estudio de funciones

vectoriales y medidas de distancia. Esto nos permite abordar eficazmente la geografía espacial y dimensional.

La distancia euclidiana en el álgebra lineal está relacionada con la longitud de un vector, o lo que más fácilmente se puede considerar como la longitud de un camino recto en el espacio entre dos puntos. El producto escalar también es útil para encontrar el ángulo entre dos vectores y descomponer la distancia en segmentos más pequeños. datos Es esencial para el algoritmo.

El agrupamiento en clústeres k-means ayuda a organizar los datos en clústeres en función de la proximidad. El escalado multidimensional (MDS) también simplifica la visualización de datos complejos utilizando distancias euclidianas, lo que facilita la identificación de patrones y tendencias.

Distancia de Mahalanobis:

$$d_{M(x_i,c_k)} = \sqrt{\{(x_i - c_k)^T \backslash text\{Cov\}^{\{-1\}(x_i - c_k)}\}}$$

Distancia Manhattan

$$d_{M(x_i,c_k)} = \sum_{\{j=1\}}^{\{d\}|x_{\{ij\}} - c_{\{kj\}}|}$$

Convergencia del Algoritmo

El algoritmo **K-means** siempre converge, pero puede llegar a un **mínimo local** en lugar del óptimo global. La convergencia se verifica cuando la posición de los centroides ya no cambia significativamente entre iteraciones:

$$| c_k^{\{(t+1)\}} - c_k^{\{(t)\}|} < \epsilon$$

donde ϵ\epsilonϵ es un umbral pequeño que determina la convergencia.

Evaluación de la Calidad

Una forma común de evaluar la calidad de los clusters es mediante el **Índice Dunn** (Dunn, 1979), que mide la relación entre la dispersión intra-cluster y la separación entre clusters. El índice Dunn se define como:

$$D = \frac{\min_{\{k \neq j\}\text{dist}(c_k,c_j)}}{\max_k \text{diam}(k)}$$

Donde $\{k \neq j\}\text{dist}$ es la distancia entre los centroides de los clusters k y j, y $\{\max_k\text{diam}(k)\}$es el diámetro del cluster k.

Aplicación del modelo para la predicción.

Una vez que los exportadores se han dividido en clústeres de exportadores, los patrones identificados se analizarán para predecir las exportaciones de volumen. Los clústeres resultantes se interpretarán en términos de:

- Tamaño y frecuencia de exportación.
- Contribución al valor total de las exportaciones.
- Capacidad de los exportadores de mantenerse atractivos para el mercado chino.

Estos hallazgos serán utilizados para desarrollar estrategias diferenciadas para cada clúster, lo que optimizará la participación de Ecuador en el mercado chino.

Herramientas utilizadas

El desarrollo del modelo se realizará utilizando: Lenguaje de programación R o Python (bibliotecas como Scikit-learn y pandas).

Métricas y visualizaciones:

Gráficos de SSE y silueta para evaluar el modelo. Conclusión metodológica La metodología aplicada combina técnicas de preprocesamiento, algoritmos de agrupamiento y métodos de validación, teniendo como resultado el pronóstico y clasificación de las exportaciones de camarón.

La implementación del algoritmo k-means permitirá identificar patrones clave en el historial de datos y mejorar la toma de decisiones estratégicas y la competitividad del sector camaronero en el mercado chino.

8.4 Desarrollo R:

Como paso inicial, debemos acceder al dataset mediante la función (read.csv) o (read.csv2), todo depende del formato de este dataset. Hay que considerar que es muy importante seleccionar correctamente la función para que el modelo pueda ser integrado de manera óptima.

```
Flete <- read.csv2("../DATA/Datos Tesis - NA - GY.csv")
```

Una vez que defimos la variable con el nombre de **Flete**, procedemos a crear una nueva variable con la base de datos eliminando la columna 1 y 2 (Consignor Name y ETD).

```
Flete_data_cleaned <- Flete[,c(-1,-2)]
```

Una vez que la base de datos esté lista, podemos comenzar a codificar las funciones esenciales para generar el algoritmo de K-means. En esta etapa, es fundamental descargar una serie de paquetes. Para utilizar estos paquetes en RStudio, llamamos la librerias que vamos a usar mediante la funcion **library**, como se muestra a continuación:

```
#Llamamos a las librerias a usar

library(class)

library(caret)

library(tidyverse)

library(cluster)

library(factoextra)

library(NbClust)
```

Los paquetes necesarios variarán según el algoritmo y la base de datos que se elijan. A continuación, se detallarán los paquetes instalados para este proyecto:

class El paquete class proporciona varias funciones para la clasificación, incluyendo el algoritmo de k-vecinos más cercanos (k-NN), la cuantificación vectorial de aprendizaje y los mapas auto-organizativos (SOM). Estas herramientas son esenciales para tareas de clasificación en análisis de datos (Ripley, 2025)

caret El paquete caret (Classification And Regression Training) contiene funciones para simplificar el proceso de entrenamiento de modelos para problemas complejos de regresión y clasificación. Incluye herramientas para la selección de características, la evaluación de modelos y la optimización de parámetros (Kuhn, 2024)

tidyverse El tidyverse es un conjunto de paquetes que comparten una filosofía de diseño y estructuras de datos comunes. Incluye herramientas para la manipulación de datos (dplyr), visualización (ggplot2), importación de datos (readr), y más. Estos paquetes están diseñados para trabajar juntos de manera armoniosa (Wickham, 2023)

cluster El paquete cluster proporciona métodos para el análisis de conglomerados, incluyendo técnicas jerárquicas y de partición. Es una extensión del trabajo original de Kaufman y Rousseeuw (1990) y ofrece herramientas para encontrar grupos en datos (Maechler et al., 2024)

factoextra El paquete factoextra facilita la extracción y visualización de los resultados de análisis de datos multivariados, como el análisis de componentes principales (PCA) y el análisis

de correspondencias múltiples (MCA). Proporciona visualizaciones elegantes basadas en ggplot2 para interpretar los resultados de manera efectiva (Kassambara & Mundt, 2020)

NbClust El paquete NbClust ofrece 30 índices para determinar el número óptimo de conglomerados en un conjunto de datos y propone el mejor esquema de agrupamiento basado en diferentes combinaciones de métodos y medidas de distancia (Charrad et al., 2022)

```
#Plantamos semilla

set.seed(123)
```

Definimos la semilla, la cual marca el comienzo de la implementación de procedimientos relacionados con árboles. Aunque hay varias opciones de semillas disponibles, en este contexto de clasificación, la elección más adecuada es la semilla (123).

```
#Dividimos los datos en entrenamiento y prueba

trainIndex <- createDataPartition(Flete_data_cleaned$Costo,

                                  p=.7, list = FALSE)

train_data <- Flete_data_cleaned[trainIndex,]

test_data <- Flete_data_cleaned[-trainIndex]
```

createDataPartition: Esta función crea particiones de prueba y entrenamiento en un conjunto de datos. Es útil para dividir los datos de manera estratificada, asegurando una distribución equitativa de las clases entre los conjuntos.

p: Este parámetro indica el porcentaje de datos que se destinará al conjunto de entrenamiento. Determina la proporción de datos que se utilizará para entrenar un modelo en comparación con el conjunto total de datos.

list: Es un parámetro lógico que determina el formato de los resultados. Si se utiliza "FALSE", los resultados se presentarán en una matriz donde el número de filas es igual a floor(p*length(y)) y el número de columnas es el número de veces que se realiza la partición

Usamos un valor de p de 0.7 porque asigna una cantidad adecuada de datos al conjunto de entrenamiento, mientras que reserva una porción significativa para la evaluación del rendimiento. Esto ayuda a garantizar que el modelo no solo se ajuste bien a los datos de entrenamiento, sino que también tenga una buena capacidad de generalización a datos nuevos.

```
#Escalamos los datos

SKtraindata <- scale(train_data)

SKtestdata <- scale(test_data)
```

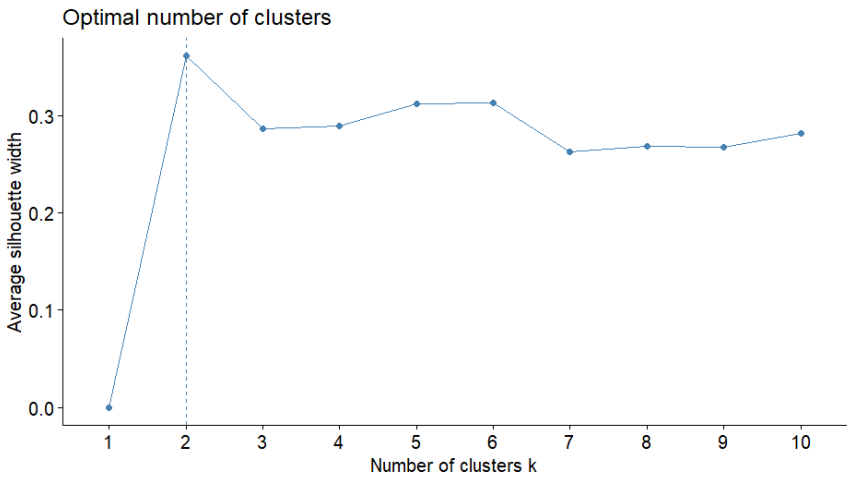
La función **scale** estandariza los datos de entrenamiento. Esta centra y escala las variables, es decir, resta la media y divide por la desviación estándar de cada variable. Esto es útil para asegurar que todas las variables tengan la misma escala y evitar que algunas variables dominen el modelo debido a sus magnitudes.

Lo hace para el **train_data** que contiene el 70% de los datos originales. Mientra que **test_data** contiene el 30% de los datos restantes.

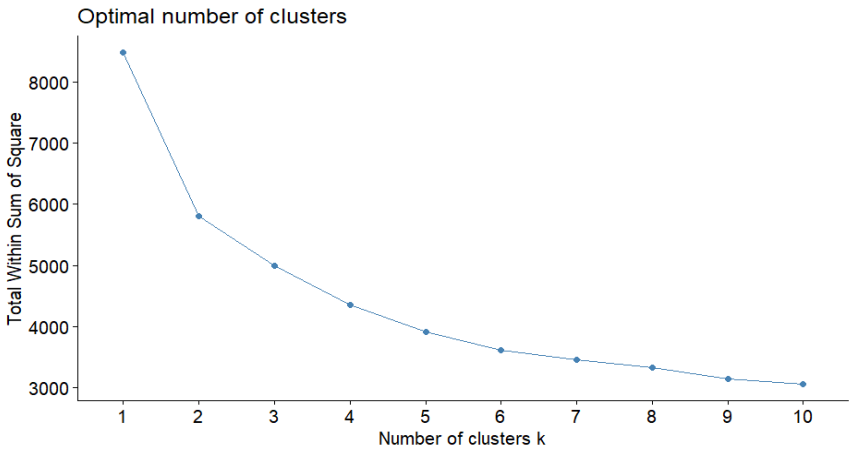
```
#TRABAJANDO CON LOS DATOS DE ENTRENAMIENTO
```

La función **fviz_nbclust** del paquete **factoextra** en R se utiliza para determinar y visualizar el número óptimo de clusters en un análisis de clustering. Esta función es especialmente útil cuando se utilizan métodos de particionamiento como k-means, donde es necesario especificar el número de clusters a generar. (Kassambara & Mundt, 2020).

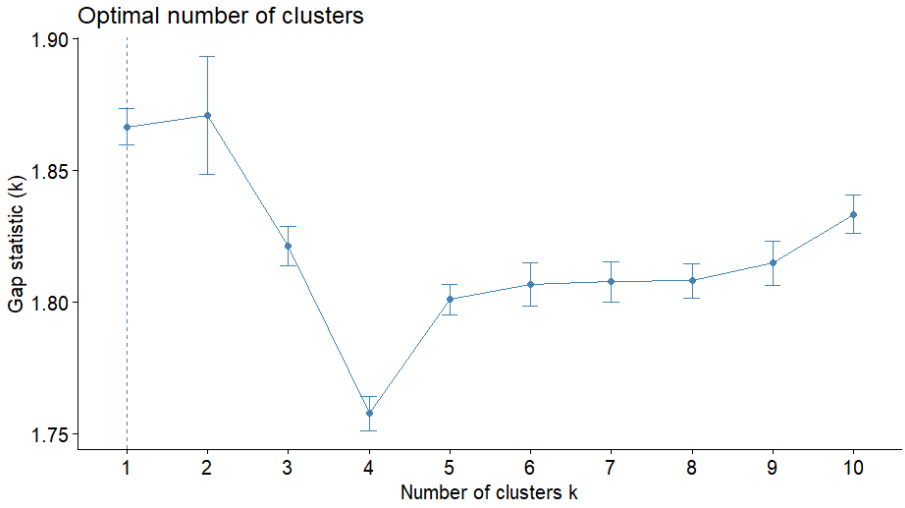
fviz_nbclust(SKtraindata, kmeans, method = "silhouette")



fviz_nbclust(SKtraindata, kmeans, method = "wss")



fviz_nbclust(SKtraindata, kmeans, method = "gap_stat")



Silhouette: Evalúa la calidad de los clusters midiendo qué tan similares son los puntos dentro de un cluster en comparación con los puntos de otros clusters.

Within Sum of Squares (WSS): Calcula la suma de las distancias cuadradas dentro de cada cluster, buscando el punto donde la WSS disminuye significativamente.

Gap Statistic: Compara la variación total dentro de los clusters con la variación esperada bajo una distribución nula de referencia

Después de evaluar los 3 metodos mencionados, Silhouette y WSS indican que el número ideal de clusters son 2, a diferencia de Gap Statistic que indica que solo 1 cluster es lo correcto. Por lo tanto, debido a la mayoría vamos a trabajar con 2 clusters.

```
#Calculamos los clústers


K2 <- kmeans(SKtraindata, centers = 2, nstart = 25)
```

La función kmeans en R se utiliza para realizar el análisis de clustering k-means, que es una técnica de aprendizaje no supervisado que agrupa observaciones en K clusters. El objetivo es que las observaciones dentro de cada cluster sean lo más similares posible entre sí y lo más diferentes posible de las observaciones en otros clusters (R Documentation, 2014).

centers: Número de clusters deseados o un conjunto de centros de clusters iniciales.

nstart: Número de configuraciones iniciales aleatorias a probar.

```
#Plot

PTrainOC <- fviz_cluster(K2, data = SKtraindata,repel = TRUE)

PTrainOC
```

```
PTrainMC <- fviz_cluster(K2, data = SKtraindata,
                         ellipse.type = "euclid",
                         repel = FALSE, star.plot= TRUE)
PTrainMC


PTrainIC <- fviz_cluster(K2, data = SKtraindata,
                          ellipse.type = "norm",repel = FALSE)
PTrainIC
```

Creamos graficos para visualizar los clusters con nuestra variable K2

data = SKtraindata: Especifica los datos estandarizados que se utilizaron para el clustering

repel = TRUE: Utiliza la función ggrepel para evitar la superposición de etiquetas en el gráfico, mejorando la legibilidad. Donde el resultado sera almacenado en PTrainOC.

ellipse.type = "euclid": Dibuja elipses euclidianas alrededor de cada cluster, lo que ayuda a visualizar la dispersión de los datos dentro de cada cluster.

repel = FALSE: No utiliza ggrepel, por lo que las etiquetas pueden superponerse.

star.plot = TRUE: Dibuja líneas desde los centros de los clusters hasta cada punto de datos, lo que ayuda a visualizar la estructura del cluster.

```
#TRABAJANDO CON LOS DATOS DE PRUEBA
```

#Calculamos los clústers

Kt2 <- kmeans(SKtestdata, centers = 2, nstart = 25)

Ahora calculamos los clusters pero con los datos de prueba del dataset.

#agregamos las clasificaciones al conjunto de datos

Flete_data_cleaned$class <- NA

Flete_data_cleaned$class[trainIndex] <- as.factor(K2$cluster)

Flete_data_cleaned$class[-trainIndex] <- as.factor(Kt2$cluster)

Creamos una nueva variable, donde dentro de **Flete_data_cleaned** se encuentran los clientes clasificados a través de los procesos que realizamos anteriormente.

#Pasamos las clasificaciones a la base original

Flete$clasificacion <- Flete_data_cleaned$class

Dentro del dataset **flete** agregamos la variable clasificación, que contendra a **Flete_data_cleaned**

 # REALIZAMOS REGRESIÓN POR KNN

library(FNN)

library(scales)

library(car)

FNN: Fast Nearest Neighbor Search Algorithms and Applications o **FNN** nos da funciones para realizar búsquedas rápidas de vecinos más cercanos y estimaciones de densidad de kernel.Según Ripley y Knorr-Held (2020), esta librería implementa algoritmos eficaces para la búsqueda de vecinos más cercanos, lo que permite un procesamiento rápido ya se con grandes conjuntos de datos.

scales: La librería **scales** ofrece herramientas para escalar datos y mapearlos a estéticas visuales. Wickham (2020) destaca que **scales** ayuda al manejo de datos para que se ajusten adecuadamente a los gráficos.

Car: Según Fox y Weisberg (2019) explican que la función car es una herramienta importante para los analistas de datos que trabajan con modelos de regresión. Ya que ofrece una amplia gama de funciones para mejorar la interpretación y la presentación de los resultados.

```
#Paso 1. Generamos las variables ficticias para poder validar cada uno de los clusteres


Rflete <- Flete[,c(-1,-2,-7)]


Rflete$cliente1 <- recode(Rflete$clasificacion, "1=1; 2=0")

Rflete$cliente2 <- recode(Rflete$clasificacion, "1=0; 2=1")


#Paso 2. Estandarizar las variables por medio del rescalamiento

Rflete$PesoEs <-  rescale(Rflete$Peso)

Rflete$ResidualES <-  rescale(Rflete$Residual)
```

```
Rflete$VentaES <-  rescale(Rflete$Venta)
```

La función **rescale** viene del paquete **scale** la cual permite transformar los datos de manera que se ajusten al nuevo rango que definimos.

```
#Paso 3. Hago las particiones correspondientes, entrenamiento, validaciones y prueba

set.seed(2018)

id.entrenamiento <-  createDataPartition(Rflete$Costo, p=0.6, list = F)

entrenamiento <- Rflete[id.entrenamiento,]

Con.temp <-  Rflete[-id.entrenamiento,]

Validacion <-  createDataPartition(Con.temp$Costo, p=0.5 ,list = F)

V.entrenamiento <-  Con.temp[Validacion,]

Pruebas <-  Con.temp[-Validacion,]
```

Creamos las nuevas variables desde el dataset **Rflete**.

```
#Paso 4. Extraer el mejor modelo de regresión con el mejor K-vecinos


mod1 <-  knn.reg(entrenamiento[,6:10], V.entrenamiento[,6:10], entrenamiento$Costo,

            k=1, algorithm = "brute")
```

Generamos el primer modelo utilizando la función **knn.reg** del paquete **FNN.** Con esta linea de código realiza un regresión donde definimos el número de clusters con **K.**

```
mod1
```

Después sacamos el error cuadratico para ir comparando cada modelo que realizamos.

```
rmse1 = sqrt(mean(mod1$pred - V.entrenamiento$Costo)^2)

Rmse1
```

Error Cuadratico 1= > rmse1

[1] 12.65174

```
mod2 <-  knn.reg(entrenamiento[,6:10], V.entrenamiento[,6:10], entrenamiento$Costo,
                 k=2, algorithm = "brute")

mod2
```

```
rmse2 = sqrt(mean(mod2$pred - V.entrenamiento$Costo)^2)

rmse2
```

Error Cuadratico 2= rmse2

[1] 22.54097

```
mod3 <-  knn.reg(entrenamiento[,6:10], V.entrenamiento[,6:10], entrenamiento$Costo,
                 k=3, algorithm = "brute")

mod3
```

```
rmse3 = sqrt(mean(mod3$pred - V.entrenamiento$Costo)^2)

Rmse3
```

Error Cuadratico 3= rmse3

[1] 24.10476

```
mod4 <-  knn.reg(entrenamiento[,6:10], V.entrenamiento[,6:10], entrenamiento$Costo,
                 k=4, algorithm = "brute")
mod4
```

```
rmse4 = sqrt(mean(mod4$pred - V.entrenamiento$Costo)^2)
rmse4
```

Error Cuadratico 4= rmse4

[1] 31.38464

```
#El mejor modelo de predicción de ventas es el modelo 1 debido a que tiene menor error cuadrático. Utilizaremos 1 cluster para la regresión.
```

```
#Predicciones para todo el conjunto de datos
mod1_all <- knn.reg(entrenamiento[,6:10], Rflete[,6:10], entrenamiento$Venta, k=1, algorithm = "brute")
```

```
# Asignar las predicciones al conjunto completo de datos
Rflete$CostoPred <- mod1_all$pred
```

```
library(ggplot2)
```

Según Wickham et al. (2024), **ggplot2** ayuda con la creación de gráficos mediante la combinación de elementos componibles, lo que permite personalización en la visualización de datos.

```
# Crear el gráfico de dispersión entre las ventas reales y las predicciones
ggplot(Rflete, aes(x = Rflete$Costo, y = Rflete$CostoPred)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) + # Línea de regresión (sin intervalo de confianza)
  labs(title = "Dispersión entre Costo Reales y Predicciones de costo",
       x = "Costos Reales",
       y = "Predicciones de Costos") +
  theme_minimal()
```

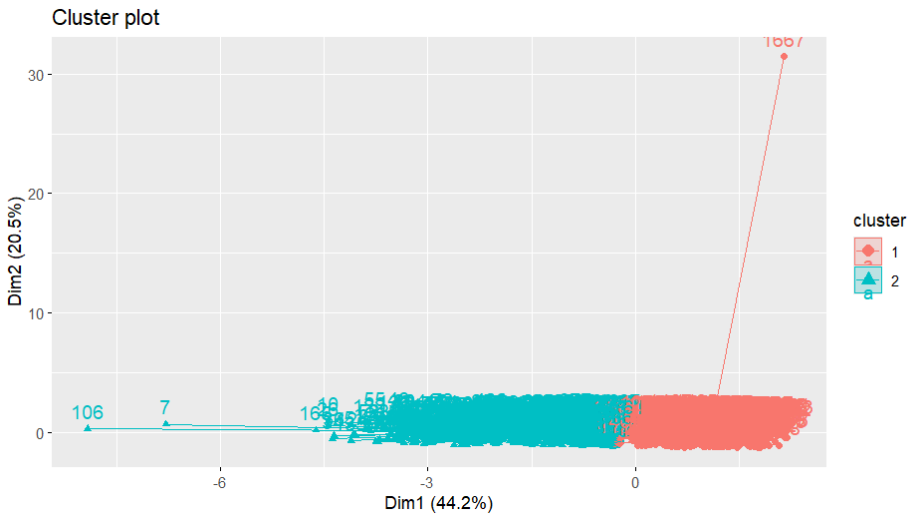
Creamos un gráfico con ggplot, donde se observamos la comparación entre costo reales y predicciones del costo. Esto para mejorar la visualización de como ayudo la clasificación a la predicción de los costos.

```
#Pasamos los resultados a la base original
Flete$Costo_Predicho <- Rflete$CostoPred
```

```
#Validamos mediante comparación dos gráficos de costos reales y costos predichos por tipo de clasificación.

Flete$clasificacion <- factor(Flete$clasificacion)


CReales <- ggplot(data = Flete, aes(x=Flete$clasificacion,
                                    y=Flete$Costo))+
  geom_boxplot(fill= "#27F4B8", colour="black")+
  labs(title = "Costos reales por tipo de clasificación de clientes")
CReales


 CPred <- ggplot(data = Flete, aes(x=Flete$clasificacion,
                                   y=Flete$Costo_Predicho))+
  geom_boxplot(fill= "#27F4B8", colour="black")+
  labs(title = "Costos predichos por tipo de clasificación de clientes")
CPred


library(gridExtra)
grid.arrange(CReales, CPred)
```

9. Resultados

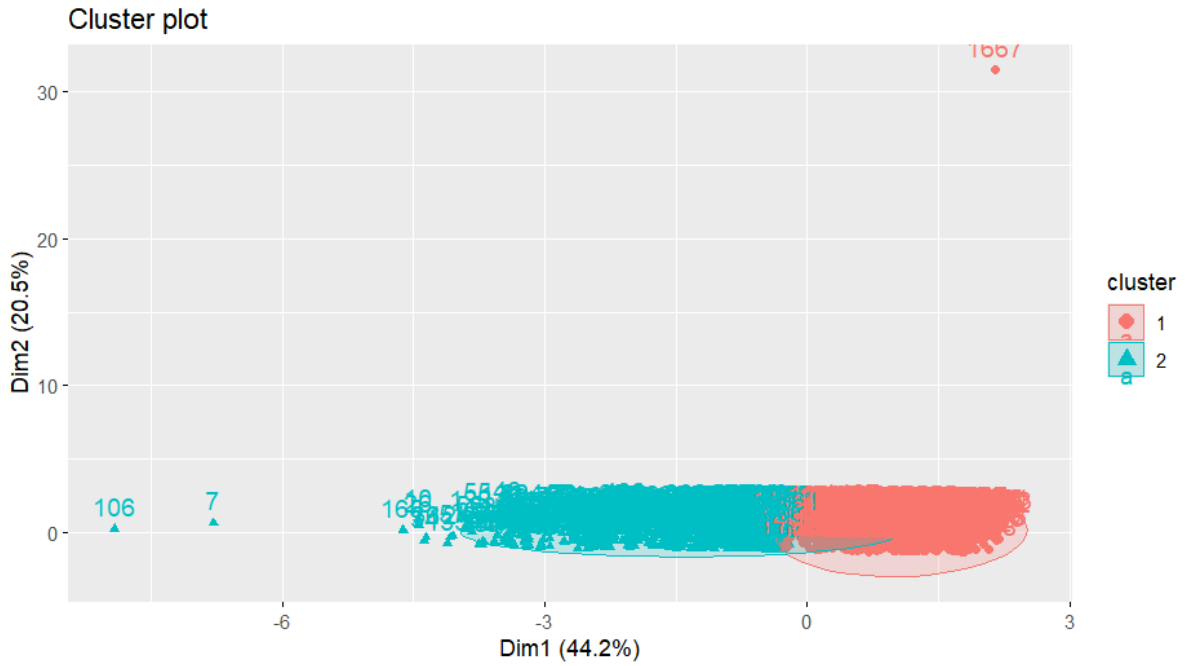
En la **Figura 1** evaluamos como están siendo clasificados los costos por "Clientes Normales" y "Clientes Premium" donde observamos que la mayoría son clasificados como "Clientes Premium". El grupo 1 (rojo) representa a los clientes premium mientras que el grupo 2 (turquesas). A pesar de obtener una buena clasificación, notamos que ciertos clientes pueden pertenecer a ambos grupos.



En la **Figura 2** podemos visualizar que no tenemos demasiados outliers, por lo tanto, sabemos que el dataset que tenemos es bastante confiable. Tener pocos outliers mejora la precisión de

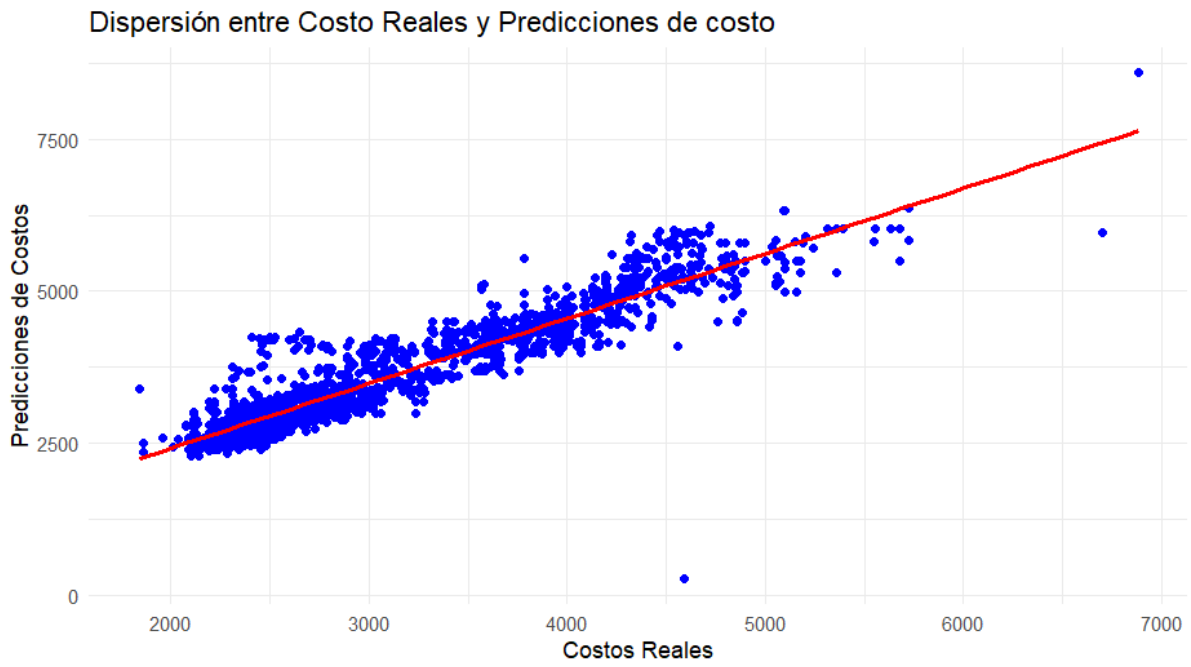
las medidas estadísticas, hace que los datos sean más consistentes y predecibles, y asegura que los modelos predictivos sean más robustos y fiables.

Estos outliers pueden representar clientes que solo trabajaron una vez y el valor del flete dejo una ganancia considerable, pero estos después del costo tan elevado decidieron no trabajar. Por otra parte, podemos ver que hay clientes que dejaron un profit muy bajo. O sea, que la operación pudo haber tenido errores o se realizó un embarque con bajo profit para atraer un cliente.



En la **Figura 3** visualizamos las siluetas de los clusters. Donde podemos visualizar como se agrupan los clientes y su concentración dentro de la figura. Pero, también debemos considerar que clientes del grupo 2 también podrían ser clasificados dentro del grupo 1.

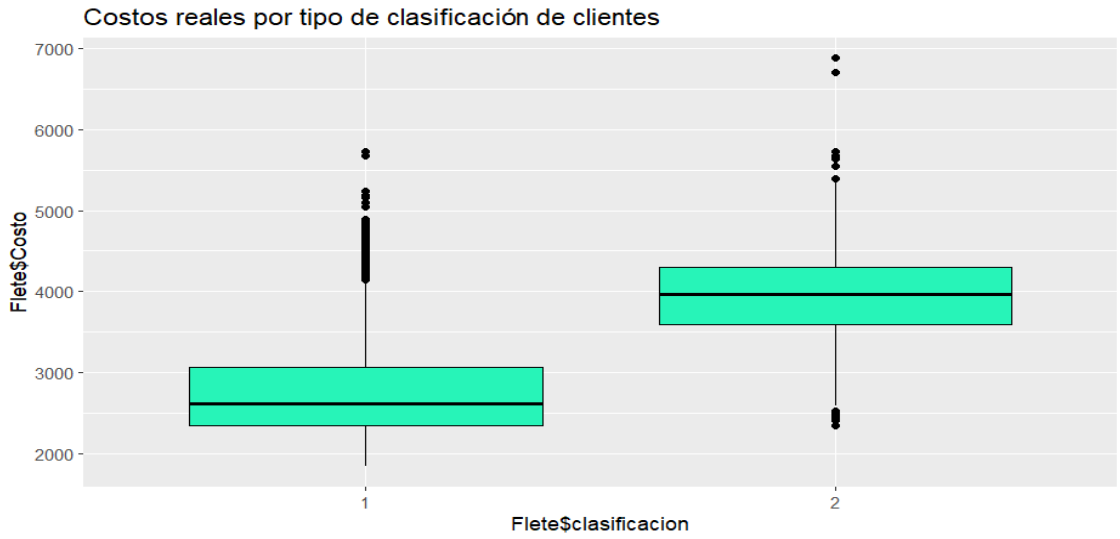
Esto puede ser debido a varios factores desde los costos hasta la atención personalizada para el servicio al cliente. Clientes que son considerados como normales, pueden pertenecer al grupo de clientes premium y poder cobrarles más por un mejor servicio.



En la **Figura 4** vemos la tendencia ascendente de los puntos indica una relación positiva entre el costo predicho y el costo real. Esto sugiere que, en general, el modelo de predicción es preciso, ya que los costos predichos tienden a alinearse con los costos reales.

Además, la cercanía de los puntos a la línea de identidad (donde el costo predicho es igual al costo real) refleja la precisión del modelo. Cuanto más cerca estén los puntos de esta línea, más precisas son las predicciones del modelo.

Es crucial para evaluar la efectividad del modelo predictivo de costos de fletes. La alineación de los costos predichos con los costos reales demuestra que el modelo puede ser utilizado de manera confiable para estimar los costos de futuros envíos. Además, la identificación de clusters mediante k-means permite una mejor comprensión de las variaciones en los costos de fletes, lo que puede llevar a estrategias de optimización más efectivas.



En la **Figura 5** se visualizan ambos grupos de clientes. Donde podemos ver que los clientes sin clasificar presentan costos se encuentra bastante elevados.

Finalmente, en la **Figura 6** tras la clasificación la reducción de outliers es significativa. Además, está segmentación disminuye significativamente los costos. A pesar de que, el grupo 1 sigue teniendo ciertos outliers. Podemos visualizar como la mediana tuvo cierto ajuste.

10. Conclusiones

El presente estudio muestra la eficacia del uso del algoritmo k-means para la clasificación y predicción de las exportaciones de camarón desde Guayaquil hacia China. A través de la implementación de técnicas de machine learning, se ha logrado clasificar a los exportadores en clústeres con características similares, lo que permite una mejor comprensión del comportamiento de exportaciones y una predicción más precisa de las ventas.

Los resultados obtenidos indican que el algoritmo k-means es una herramienta útil para identificar grupos de alto rendimiento y extrapolar patrones de éxito. La segmentación de los exportadores ha permitido desarrollar estrategias diferenciadas para cada clúster, lo que optimiza la participación de Ecuador en el mercado chino. Además, la metodología aplicada fue robusta y consistente. Al comparar nuestros gráficos.

En conclusión, la implementación de modelos de ML, en este caso el algoritmo k-means representa una excelente oportunidad para mejorar la competitividad del sector camaronero ecuatoriano en el mercado internacional. La capacidad de predecir los costos ayuda a la toma de decisiones estratégicas. Se debe considerar para estudios futuros la integración de otro algoritmo de ML y la exploración de nuevas variables que puedan afectar las exportaciones, con el objetivo de seguir optimizando el modelo pronóstico y adaptarlo a la empresa.

El poder segmentar los clientes en grupos premium y normales ayuda a reducir los costos de la empresa. Brindando un mejor servicio a su grupo de clientes y ayudando a los exportadores a mantener un costo operativo manejable, mientras el profit aumenta debido a la reducción.

Finalmente, podemos inferir que poder clasificar y predecir a través del machine learning puede ayudar de manera positiva a las empresas. Siendo la inteligencia artificial parte importante del

machine learning y contribuyendo a las exportaciones de camarón de nuestro país. Siendo uno de nuestros productos lideres en ventas.

11. Anexos

Anexo 1

Script Completo

```
setwd("../BI/DATA/")
Flete <- read.csv2("../DATA/Datos Tesis - NA - GY.csv")

Flete_data_cleaned <- Flete[,c(-1,-2)]

#Llamamos a las librerias a usar
library(class)
library(caret)
library(tidyverse)
library(cluster)
library(factoextra)
library(NbClust)

#Plantamos semilla
set.seed(123)

#Dividimos los datos en entrenamiento y prueba
trainIndex <- createDataPartition(Flete_data_cleaned$Costo,
                                  p=.7, list = FALSE)
train_data <- Flete_data_cleaned[trainIndex,]
test_data <- Flete_data_cleaned[-trainIndex]

#Escalamos los datos
SKtraindata <- scale(train_data)
SKtestdata <- scale(test_data)

################TRABAJANDO CON LOS DATOS DE ENTRENAMIENTO######

fviz_nbclust(SKtraindata, kmeans, method = "silhouette")
fviz_nbclust(SKtraindata, kmeans, method = "wss")
fviz_nbclust(SKtraindata, kmeans, method = "gap_stat")

#Calculamos los clústers

K2 <- kmeans(SKtraindata, centers = 2, nstart = 25)

```

```r
#Plot
PTrainOC <- fviz_cluster(K2, data = SKtraindata,repel = TRUE)
PTrainOC
PTrainMC <- fviz_cluster(K2, data = SKtraindata,
                         ellipse.type = "euclid",
                         repel = FALSE, star.plot= TRUE)
PTrainMC
PTrainIC <- fviz_cluster(K2, data = SKtraindata,
                          ellipse.type = "norm",repel = FALSE)
PTrainIC

################TRABAJANDO CON LOS DATOS DE PRUEBA######

#Calculamos los clústers

Kt2 <- kmeans(SKtestdata, centers = 2, nstart = 25)

#agregamos las clasificaciones al conjunto de datos
Flete_data_cleaned$class <- NA
Flete_data_cleaned$class[trainIndex] <- as.factor(K2$cluster)
Flete_data_cleaned$class[-trainIndex] <- as.factor(Kt2$cluster)

#Pasamos las clasificaciones a la base original
Flete$clasificacion <- Flete_data_cleaned$class

###HACER UN BOXPLOT ENTRE EL COSTO Y LA CLASIFICACIÓN
PARA VER SEGMENTADO EL GRUPO DE CLIENTES

############ REALIZAMOS REGRESIÓN POR KNN

library(FNN)
library(scales)
library(car)

#Paso 1. Generamos las variables ficticias para poder validar cada uno de
los clusteres

Rflete <- Flete[,c(-1,-2,-7)]
```

```r
Rflete$cliente1 <- recode(Rflete$clasificacion, "1=1; 2=0")
Rflete$cliente2 <- recode(Rflete$clasificacion, "1=0; 2=1")


#Paso 2. Estandarizar las variables por medio del rescalamiento
Rflete$PesoEs <-  rescale(Rflete$Peso)
Rflete$ResidualES <-  rescale(Rflete$Residual)
Rflete$VentaES <-  rescale(Rflete$Venta)

#Paso 3. Hago las particiones correspondientes, entrenamiento, validaciones y prueba
set.seed(2018)

id.entrenamiento <-  createDataPartition(Rflete$Costo, p=0.6, list = F)
entrenamiento <- Rflete[id.entrenamiento,]
Con.temp <-  Rflete[-id.entrenamiento,]
Validacion <-  createDataPartition(Con.temp$Costo, p=0.5 ,list = F)
V.entrenamiento <-  Con.temp[Validacion,]
Pruebas <-  Con.temp[-Validacion,]

#Paso 4. Extraer el mejor modelo de regresi?n con el mejor K-vecinos

mod1 <-  knn.reg(entrenamiento[,6:10], V.entrenamiento[,6:10], entrenamiento$Costo,
               k=1, algorithm = "brute")
mod1

rmse1 = sqrt(mean(mod1$pred - V.entrenamiento$Costo)^2)
rmse1

mod2 <-  knn.reg(entrenamiento[,6:10], V.entrenamiento[,6:10], entrenamiento$Costo,
                 k=2, algorithm = "brute")
mod2

rmse2 = sqrt(mean(mod2$pred - V.entrenamiento$Costo)^2)
rmse2

mod3 <-  knn.reg(entrenamiento[,6:10], V.entrenamiento[,6:10], entrenamiento$Costo,
                 k=3, algorithm = "brute")
mod3
```

```r
rmse3 = sqrt(mean(mod3$pred - V.entrenamiento$Costo)^2)
rmse3

mod4 <-  knn.reg(entrenamiento[,6:10], V.entrenamiento[,6:10], entrenamiento$Costo,
                 k=4, algorithm = "brute")
mod4

rmse4 = sqrt(mean(mod4$pred - V.entrenamiento$Costo)^2)
rmse4


#El mejor modelo de predicción de ventas es el modelo 1

#Predicciones para todo el conjunto de datos
mod1_all <- knn.reg(entrenamiento[,6:10], Rflete[,6:10], entrenamiento$Venta, k=1,

# Asignar las predicciones al conjunto completo de datos
Rflete$CostoPred <- mod1_all$pred

library(ggplot2)

# Crear el gráfico de dispersión entre las ventas reales y las predicciones
ggplot(Rflete, aes(x = Rflete$Costo, y = Rflete$CostoPred)) +
  geom_point(color = "blue") + # Puntos en azul
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Dispersión entre Costo Reales y Predicciones de costo",
       x = "Costos Reales",
       y = "Predicciones de Costos") +
  theme_minimal() # Tema limpio para el gráfico


#Pasamos los resultados a la base original
Flete$Costo_Predicho <- Rflete$CostoPred

#Validamos mediante comparación
Flete$clasificacion <- factor(Flete$clasificacion)

```

```r
CReales <- ggplot(data = Flete, aes(x=Flete$clasificacion,
                                    y=Flete$Costo))+
  geom_boxplot(fill= "#27F4B8", colour="black")+
  labs(title = "Costos reales por tipo de clasificación de clientes")
CReales



CPred <- ggplot(data = Flete, aes(x=Flete$clasificacion,
                                  y=Flete$Costo_Predicho))+
  geom_boxplot(fill= "#27F4B8", colour="black")+
  labs(title = "Costos predichos por tipo de clasificación de clientes")
CPred

library(gridExtra)
grid.arrange(CReales, CPred)


plot(Flete$Costo,Flete$Costo_Predicho)
# Crear el gráfico de dispersión y línea de tendencia
ggplot(Flete) +
  geom_point(aes(x = Costo, y = Costo_Predicho), color = "red", shape = 1) +
  geom_smooth(aes(x = Costo, y = Costo_Predicho),
  method = "lm", se = FALSE, color = "black") +
  scale_x_continuous(breaks = seq(0, max(Flete$Costo), by = 1000)) +
  scale_y_continuous(breaks = seq(0, max(Flete$Costo_Predicho), by = 500)) +
  labs(title = "Gráfico de Dispersión de Costo vs Costo Predicho",
       x = "Costo",
       y = "Costo Predicho") +
  theme_minimal()

```

```r
ggplot(Flete) +
  geom_point(aes(x = Costo, y = Costo_Predicho, color = "Predicho"), shape = 1) +
  geom_point(aes(x = Costo, y = Costo, color = "Real"), shape = 1) +
  geom_smooth(aes(x = Costo, y = Costo_Predicho),
              method = "lm", se = FALSE, color = "black") +
  scale_x_continuous(breaks = seq(0, max(Flete$Costo), by = 1000)) +
  scale_y_continuous(breaks = seq(0, max(Flete$Costo_Predicho), by = 500)) +
  labs(title = "Costo vs Costo Predicho",
       x = "Costo",
       y = "Costo Predicho") +
  scale_color_manual(values = c("Predicho" = "blue", "Real" = "red")) +
  theme_minimal()

```

Anexo 2

Análisis descriptivo con la función Summary

```
> summary(Flete$Peso)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20000   22335   24409   24430   26629   28799
>   summary(Flete$Venta)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  277.9  2888.0  3426.2  3665.4  4307.3  8588.0
>   summary(Flete$Costo)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1842    2465    2891    3165    3860    6882
```

Anexo 3

Base de Datos

Consignor Name	Origin ETD	Peso	Venta	Costo	Residual	TEU
OCEANTREASURE S.A.	04-Jan-23	23099	6367,07	5244,07	1123	2
FRIGOLANDIA S.A.	04-Jan-23	26239	6298,98	4623,98	1675	2
COFIMAR S.A.	04-Jan-23	23455	5800	4897	903	2
COFIMAR S.A.	04-Jan-23	26740	5800	4872	903	2
COFIMAR S.A.	04-Jan-23	27088	5800	4872	928	2
COFIMAR S.A.	04-Jan-23	24443	5800	4797	1003	2
COFIMAR S.A.	20-Jan-23	23146	8012	6704	1308	2
CEAEXPORT	04-Jan-23	24548	5384,53	4562,14	822,39	2
CEAEXPORT	04-Jan-23	20912	5346,03	4523,64	822,39	2
FRIGOLANDIA S.A.	04-Jan-23	22236	6237,31	4562,31	1675	2
COFIMAR S.A.	04-Jan-23	21878	5800	4797	1003	2
COFIMAR S.A.	04-Jan-23	21359	5800	4797	1003	2
COFIMAR S.A.	04-Jan-23	26362	5800	4797	1003	2
EXPOTUNA S.A	13-Jan-23	26540	5214,37	4623,37	591	2
EXPOTUNA S.A	13-Jan-23	24999	5214,37	4623,37	591	2

EXPOTUNA S.A	13-Jan-23	26281	5320,13	4724,13	596	2
EXPOTUNA S.A	13-Jan-23	23546	5320,13	4724,13	596	2
EXPOTUNA S.A	13-Jan-23	21869	5320,13	4724,13	596	2
EXPOTUNA S.A	13-Jan-23	27402	5210,61	4619,61	591	2
EXPOTUNA S.A	13-Jan-23	23346	5210,61	4619,61	591	2
EXPOTUNA S.A	13-Jan-23	21042	5210,61	4619,61	591	2
EXPOTUNA S.A	13-Jan-23	27128	5210,61	4619,61	591	2
EXPOTUNA S.A	13-Jan-23	27451	5210,61	4619,61	591	2
EXPOTUNA S.A	13-Jan-23	23055	5210,61	4619,61	591	2
EXPOTUNA S.A	13-Jan-23	20172	5211,81	4620,81	591	2
FRIGOLANDIA S.A.	04-Jan-23	23657	6212,61	4537,61	1675	2
COFIMAR S.A.	04-Jan-23	25225	5800	4772	1003	2
COFIMAR S.A.	04-Jan-23	20623	5800	4797	1003	2
COFIMAR S.A.	19-Jan-23	27178	5490	5160	330	2
COFIMAR S.A.	11-Jan-23	23351	5489,5	5359,5	130	2
COFIMAR S.A.	11-Jan-23	27035	5489,5	5359,5	130	2
COFIMAR S.A.	11-Jan-23	24227	5310	5175	135	2
COFIMAR S.A.	08-Jan-23	22003	5290	5160	130	2

FRIGOLANDIA S.A.	19-Jan-23	27412	6317,36	5089,6	1227,76	2
FRIGOLANDIA S.A.	19-Jan-23	23757	6328,64	5100,88	1227,76	2
FRIGOLANDIA S.A.	19-Jan-23	26442	6317,36	5089,6	1227,76	2
EXPORTQUILSA	13-Jan-23	27980	5583,12	4802,73	780,39	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	20-Jan-23	23544	5161,26	4288,26	873	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	20-Jan-23	24512	5257,61	4364,61	893	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	20-Jan-23	22892	5157,61	4284,61	873	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	20-Jan-23	24759	5157,61	4284,61	873	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	20-Jan-23	28789	5157,61	4284,61	873	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	20-Jan-23	20088	5157,61	4284,61	873	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	20-Jan-23	25737	5157,61	4284,61	873	2
EXPORTQUILSA	13-Jan-23	20305	5588,46	4808,07	780,39	2
PROCESADORA DEL RIO S.A. PRORIOSA	13-Jan-23	20031	5795,61	4634,61	1161	2
PROCESADORA DEL RIO S.A. PRORIOSA	13-Jan-23	27922	5795,61	5634,61	161	2
EXPORTQUILSA	13-Jan-23	23749	5440	4659,61	780,39	2
EXPORTQUILSA	13-Jan-23	24911	5360	4579,61	780,39	2
EXPORTQUILSA	13-Jan-23	26957	5360	4579,61	780,39	2
EXPORTQUILSA	13-Jan-23	25768	6375	5722	653	2

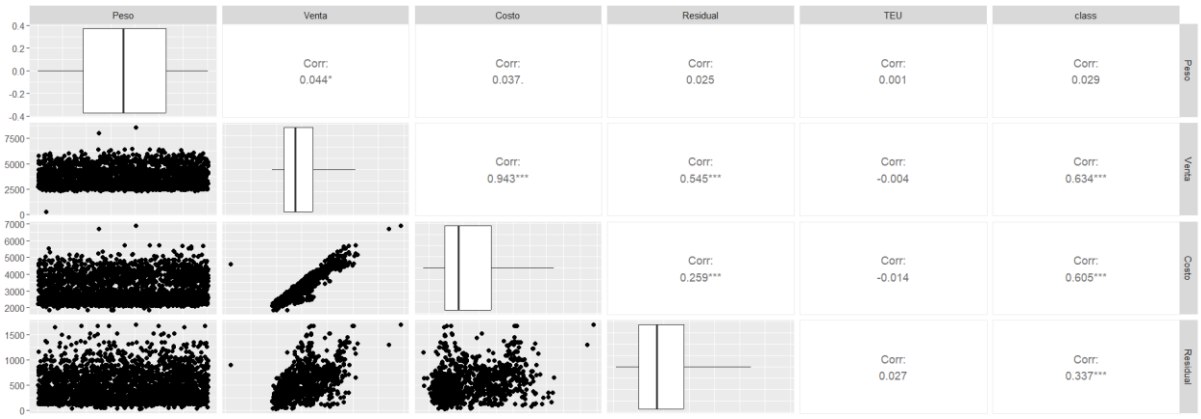
EXPORTQUILSA	13-Jan-23	26250	6375	5722	653	2
EXPORTQUILSA	13-Jan-23	24472	6375	5722	653	2
EXPORTQUILSA	13-Jan-23	24154	6086,69	5188,69	898	2
FRIGOLANDIA S.A.	13-Jan-23	20370	5925,61	4542,61	1383	2
FRIGOLANDIA S.A.	15-Jan-23	23925	5951,62	4602,62	1349	2
FRIGOLANDIA S.A.	15-Jan-23	23390	5994,84	4645,84	1349	2
FRIGOLANDIA S.A.	15-Jan-23	25647	5884,99	4535,99	1349	2
EXPOTUNA S.A	13-Jan-23	27921	5903	5552	351	2
EXPOTUNA S.A	13-Jan-23	27788	5903	5552	351	2
EXPOTUNA S.A	13-Jan-23	27465	6038	5682	356	2
EXPOTUNA S.A	20-Jan-23	28059	5614,67	4323,67	1291	2
EXPOTUNA S.A	20-Jan-23	23697	5610,61	4319,61	1291	2
PROCESADORA Y EXPORTADORA DE CAMARON PROCAMARONEX	15-Jan-23	26881	4720	4354	366	2
CEAEXPORT	17-Jan-23	26629	5575	4452,61	1122,39	2
EXPOTUNA S.A	13-Jan-23	28525	6038	5682	356	2
PROCESADORA DEL RIO S.A. PRORIOSA	20-Jan-23	21190	5395,61	4304,61	1091	2
FRIGOLANDIA S.A.	20-Jan-23	25860	5720,61	4404,61	1316	2
FRIGOLANDIA S.A.	20-Jan-23	24661	5720,61	4404,61	1316	2

FRIGOLANDIA S.A.	20-Jan-23	27686	5720,61	4404,61	1316	2
PROCESADORA DEL RIO S.A. PRORIOSA	20-Jan-23	26556	5395,61	4304,61	1091	2
PROCESADORA DEL RIO S.A. PRORIOSA	20-Jan-23	20380	5395,61	4284,61	1091	2
EXPOTUNA S.A	20-Jan-23	24244	5540	4782	758	2
EXPOTUNA S.A	20-Jan-23	24966	5405	4652	753	2
EXPOTUNA S.A	20-Jan-23	23689	5405	4652	753	2
EXPOTUNA S.A	20-Jan-23	21550	5405	4652	753	2
PROCESADORA DEL RIO S.A. PRORIOSA	20-Jan-23	23320	5728	5032	696	2
PROCESADORA DEL RIO S.A. PRORIOSA	20-Jan-23	24348	5395,61	4304,61	1091	2
PROCESADORA DEL RIO S.A. PRORIOSA	20-Jan-23	22349	5432,43	4341,43	1091	2
COFIMAR S.A.	19-Jan-23	27386	5510	5175	335	2
COFIMAR S.A.	19-Jan-23	21626	5490	5160	330	2
COFIMAR S.A.	19-Jan-23	28488	5490	5160	330	2
PROCESADORA DEL RIO S.A. PRORIOSA	20-Jan-23	22768	5208	4517	691	2
PROCESADORA DEL RIO S.A. PRORIOSA	20-Jan-23	21999	5208	4517	691	2
PROCESADORA DEL RIO S.A. PRORIOSA	20-Jan-23	27340	5246,65	4535,65	691	2
EXPORTQUILSA	13-Jan-23	23197	5440	4659,61	780,39	2
EXPORTQUILSA	13-Jan-23	23563	5440	4659,61	780,39	2

EXPORTQUILSA	13-Jan-23	21114	5440	4659,61	780,39	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	6-feb-23	23426	5067,66	4294,66	773	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	9-feb-23	24349	5060,7	4287,7	773	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	6-feb-23	23558	5140,7	4367,7	773	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	9-feb-23	27441	5242,66	4469,66	773	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	9-feb-23	21994	5242,66	4469,66	773	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	6-feb-23	20655	5067,66	4294,66	773	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	9-feb-23	23216	5067,66	4294,66	773	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	9-feb-23	27966	5067,66	4294,66	773	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	9-feb-23	23172	5140,7	4367,7	773	2
EXPORTADORA TOTAL SEAFOOD TOTALSEAFOOD S. A.	9-feb-23	25485	5060,7	4287,7	773	2
PROCESADORA Y EXPORTADORA DE CAMARON PROCAMARONEX	24-Jan-23	24566	4720	4319	401	2

Anexo 3

Análisis descriptivo.



12. Referencias

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. In *Electronics (Switzerland)* (Vol. 9, Issue 8, pp. 1–12). MDPI AG. https://doi.org/10.3390/electronics9081295

Alboukadel Kassambara, & Fabian Mundt. (2022). *Package "factoextra" Type Package Title Extract and Visualize the Results of Multivariate Data Analyses*. https://github.com/kassambara/factoextra/issues

Arthur, D., & Vassilvitskii, S. (n.d.). *k-means++: The Advantages of Careful Seeding*.

AWS. (2024). *¿En qué consiste la limpieza de los datos?* Amazon. https://aws.amazon.com/es/what-is/data-cleansing/

Bao Chong. (2021). K-means clustering algorithm: a brief review. *Academic Journal of Computing & Information Science*, *4*(5). https://doi.org/10.25236/ajcis.2021.040506

Beygelzimer Alina, Kakadet Sham, & Langford John. (2025). *Package "FNN" Title Fast Nearest Neighbor Search Algorithms and Applications*.

Bifet, A., Gavaldà, R., Holmes, G., & Pfahringer, B. (2018). *Machine Learning for Data Streams*. The MIT Press. https://doi.org/10.7551/mitpress/10654.001.0001

Borlea, I. D., Precup, R. E., Dragan, F., & Borlea, A. B. (2017). Centroid update approach to K-means clustering. *Advances in Electrical and Computer Engineering*, *17*(4), 3–10. https://doi.org/10.4316/AECE.2017.04001

Chapelle, O., Schölkopf, B., & Zien, A. (Eds.). (2006). *Semi-Supervised Learning*. The MIT Press. https://doi.org/10.7551/mitpress/9780262033589.001.0001

Charrad, M., Ghazzali, N., Boiteau, V., & Maintainer, A. N. (2022). *Title Determining the Best Number of Clusters in a Data Set*.

Chu, X. (n.d.). *Data Cleaning*.

Devavrat Shah. (2024). *From Data to Decisions at MIT Professional Education*. https://professionalprograms.mit.edu/blog/technology/machine-learning-vs-artificial-intelligence/

Effendy, D. A., Kusrini, K., & Sudarmawan, S. (2017). Classification of intrusion detection system (IDS) based on computer network. *2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 90–94. https://doi.org/10.1109/ICITISEE.2017.8285566

Ethem Alpaydin. (2010). Introduction to Machine Learning. *MIT PRESS*, 9–9.

Fox Jhon, Weisberg Sanford, Price Brad, Adler Daniel, Baud-Bovy Gabriel, & Bolker Ben. (2024). *Package "car."* https://r-forge.r-project.org/projects/car/,

Freire Míguez, L. (2022). *Tratamiento de falta de información en técnicas de minería de datos*.

Goodfellow, I., Bengio, Y., & Courville, A. (n.d.). *Deep Learning*.

Hadley Wickham. (2023). *Package "tidyverse" Title Easily Install and Load the "Tidyverse."* https://github.com/tidyverse/tidyverse

Hadley Wickham, & Thomas Lin Pedersen. (2023). *Package "scales" Title Scale Functions for Visualization*.

Hadley Wickham, Winston Chang, & Lionel Henry. (2024). *Package "ggplot2" Title Create Elegant Data Visualisations Using the Grammar of Graphics*.

Harris, S. (2021). *k-means initialisation algorithms: an extensive comparative study*.

Huang, S., Cao, H., Liu, J., Yan, R., & Zhang, H. (2022). Study on the Application of Elastic Wave CT Technique to Detect the Effect of Post-Grouting of Pile Foundation. *Applied Sciences*, *13*(1), 456. https://doi.org/10.3390/app13010456

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

James MacQueen. (1967). *Some Methods for Classification and Analysis of Multivariate Observations*.

Jorge Dagnino. (2014). Datos Faltantes (Missing Values). *Revista Chilena de Anestesia*.

Khiem, N. M., Takahashi, Y., Dong, K. T. P., Yasuma, H., & Kimura, N. (2021). Predicting the price of Vietnamese shrimp products exported to the US market using machine learning. *Fisheries Science*, *87*(3), 411–423. https://doi.org/10.1007/s12562-021-01498-6

Malone, T. W., & Laubacher, R. (n.d.). *ARTIFICIAL INTELLIGENCE AND THE FUTURE OF WORK SCIENCE MIT TASK FORCE ON THE WORK OF THE FUTURE MEMBER*.

Martin Maechler, Mia Hubert, Anja Struyf, & Peter Rousseeuw. (2024). *Package "cluster."* https://orcid.org/0000-0001-9143-4880

Max Kuhn, Steve Weston, Jed Wing, & Chris Keefer. (2024). *Package "caret" Title Classification and Regression Training*.

Micocci, F., & Rungi, A. (2021). *Predicting Exporters with Machine Learning*. https://doi.org/10.1017/S1474745623000265

Microsoft. (2024). *Las diez formas principales de limpiar los datos*. Microsoft. https://support.microsoft.com/es-es/office/las-diez-formas-principales-de-limpiar-los-datos-2844b620-677c-47a7-ac3e-c2e157d1db19

Miguel Jimenez. (2024). *Escalado de variables*. https://migueljimenezg.github.io/cursos/Machine%20Learning/Deep%20Learning/Escalado%20de%20variables/Escalado%20de%20variables.html

Mitchell, T. M. . (1997). *Machine learning*. McGraw-Hill.

Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C. (2010). Application of k-Means Clustering algorithm for prediction of Students' Academic Performance. In *IJCSIS) International Journal of Computer Science and Information Security* (Vol. 7, Issue 1). http://sites.google.com/site/ijcsis/

Padilla-Ospina, A. M., Medina-Vásquez, J. E., & Ospina-Holguín, J. H. (2020). Métodos de aprendizaje automático en los estudios prospectivos desde un ejemplo de la financiación de la innovación en Colombia. *Revista de Investigación, Desarrollo e Innovación*, *11*(1), 9–21. https://doi.org/10.19053/20278306.v11.n1.2020.11676

Pang Ning Tan, Michael Steinbach, & Vipin Kumar. (2006). *Introduction to data mining*. Pearson Addison Wesley.

R Documentation. (2014). *kmeans: K-Means Clustering*. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kmeans

R Documentation. (2024). *Clustering con K-means en R*.

Ripley, B., Venables, W., & Maintainer,]. (2025). *Title Functions for Classification NeedsCompilation yes*.

Rojas, E. M. (n.d.). *Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo*.

Rousseeuw, P. J. (1987a). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rousseeuw, P. J. (1987b). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Santiago Morante. (2018, November 1). *Precauciones a la hora de normalizar datos en Data Science*. https://telefonicatech.com/blog/precauciones-la-hora-de-normalizar

Sekeroglu, B., Ever, Y. K., Dimililer, K., & Al-Turjman, F. (2022). Comparative Evaluation and Comprehensive Analysis of Machine Learning Models for Regression Problems. *Data Intelligence*, *4*(3), 620–652. https://doi.org/10.1162/dint_a_00155

Solex. (2024). *Guía para la limpieza de datos: definición, beneficios, componentes y cómo limpiar*. https://support.microsoft.com/es-es/office/las-diez-formas-principales-de-limpiar-los-datos-2844b620-677c-47a7-ac3e-c2e157d1db19

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267–276. https://doi.org/10.1007/BF02289263

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the Number of Clusters in a Data Set Via the Gap Statistic. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293

Zubair, M., Iqbal, M. A., Shil, A., Chowdhury, M. J. M., Moni, M. A., & Sarker, I. H. (2022). An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling. *Annals of Data Science*. https://doi.org/10.1007/s40745-022-00428-2

DECLARACIÓN Y AUTORIZACIÓN

Nosotros, **Acosta Pisco Nicolás Julián** con C.C: # 0951184456 y **Yepez Jouvin Geanella Alejandra** con C.C: # 0943941062 autores del trabajo de integración curricular: **Implementación de Machine Learning para predicción de costos de camarón hacia China con exportadores de Guayaquil,** previo a la obtención del título de **Licenciado en Negocios Internacionales** en la Universidad Católica de Santiago de Guayaquil.

1.- Declaro tener pleno conocimiento de la obligación que tienen las instituciones de educación superior, de conformidad con el Artículo 144 de la Ley Orgánica de Educación Superior, de entregar a la SENESCYT en formato digital una copia del referido trabajo de integración curricular para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.

2.- Autorizo a la SENESCYT a tener una copia del referido trabajo de integración curricular, con el propósito de generar un repositorio que democratice la información, respetando las políticas de propiedad intelectual vigentes.

Guayaquil, 07 de febrero de 2025

AUTORES

f._____

Acosta Pisco, Nicolás Julián

C.C: # 0951184456

f._____

Yepez Jouvin, Geanella Alejandra

C.C: # 0943941062

REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA			
FICHA DE REGISTRO DE TRABAJO DE INTEGRACIÓN CURRICULAR			
TEMA Y SUBTEMA:	Implementación de Machine Learning para predicción de costos de camarón hacia China con exportadores de Guayaquil		
AUTOR(ES)	Acosta Pisco Nicolás Julián Yepez Jouvin Geanella Alejandra		
REVISOR(ES)/TUTOR(ES)	Ing. Carrera Buri, Félix Miguel, Mgs		
INSTITUCIÓN:	Universidad Católica de Santiago de Guayaquil		
FACULTAD:	Facultad de Economía y Empresa		
CARRERA:	Negocios Internacionales		
TITULO OBTENIDO:	Licenciado en Negocios Internacionales		
FECHA DE PUBLICACIÓN:	07 de febrero de 2025	**No. DE PÁGINAS:**	96
ÁREAS TEMÁTICAS:	Balanza comercial, recursos financieros		
PALABRAS CLAVES/ KEYWORDS:	Sistema logístico, exportaciones		

RESUMEN/ABSTRACT:

Las exportaciones actualmente son un reto logístico para una gran cantidad de empresas. Los exportadores de camarón deben adaptarse a los nuevos retos que se enfrentan y empresas que manejan estas exportaciones siempre están afrontando variaciones de costos en sus operaciones.

Este aumento de costos obliga a las empresas a subir sus precios e incluso pueden dificultar las exportaciones recortando recursos financieros para los camaroneros. Dentro del estudio de los datos el poder conocer y el poder implementar herramientas de Machine Learning para predecir los costos de exportación de camarón hacia China, se vuelve un recurso muy importante.

Con los datos proporcionados por un forwarder en Guayaquil. Se indago que algoritmo era el más adecuado para su aplicación, para después aplicar los algoritmos de K-Means y KNN con el fin de segmentar a los clientes en grupos con características similares y predecir los costos logísticos asociados a las exportaciones.

Los resultados mostraron una segmentación entre "Clientes Premium" y "Clientes Normales", lo que permitirá desarrollar estrategias diferenciadas. Además, se desarrolló un modelo predictivo con un bajo error cuadrático medio (RMSE), lo que sugiere que el modelo es preciso para estimar los costos de exportación. Con esto poder prevenirse a los costos y saber que se puede reducir los costos de venta para obtener mayor ganancia.

Finalmente, se propusieron mejoras futuras, como la inclusión de variables adicionales y la integración de otros algoritmos de Machine Learning, para optimizar aún más el modelo.

Este estudio contribuye a la optimización de la logística en las exportaciones de camarón y también mejoras logísticas tanto como para el forwarder como el exportador. Donde recibirá un servicio adecuado a los costos que esta pagando al ser considerado al grupo de clientes asignados y el forwarder podrá reducir sus costos a través de distintas estrategias que apliquen al conocer las predicciones del modelo.

ADJUNTO PDF:	☒ SI	☐ NO
CONTACTO CON AUTOR/ES:	**Teléfono:**	E-mail: nicolas.acosta@cu.ucsg.edu.ec geanella.yepez@cu.ucsg.edu.ec
CONTACTO CON LA INSTITUCIÓN (COORDINADOR DEL PROCESO UIC):	**Nombre: Freire Quintero Cesar enrique**	
	Teléfono: +593-990090702	
	E-mail: cesar.freire@cu.ucsg.edu.ec	

SECCIÓN PARA USO DE BIBLIOTECA	
Nº. DE REGISTRO (en base a datos):	
Nº. DE CLASIFICACIÓN:	
DIRECCIÓN URL (tesis en la web):	