UNIVERSIDAD PRIVADA DEL NORTE

FACULTAD DE INGENIERÍA

Carrera de Ingeniería de Sistemas Computacionales

“IMPLEMENTACIÓN DE UN MODELO DE CLUSTERIZACIÓN PARA LA SEGMENTACIÓN DEL PERFIL DEL CLIENTE EN EL ÁREA COMERCIAL DE SUPERMERCADOS”

Tesis para optar el título profesional de:

INGENIERO DE SISTEMAS COMPUTACIONALES

Autor:
Gustavo Alfonso Arrelucea Zapata

Asesor:
Mg. Carlos Ramos Gonzales

Lima - Perú

2020

DEDICATORIA

A mi familia:

Mis padres, hermana y mi hija. Por apoyarme en todo momento,

y permitir mi crecimiento personal como profesional

a pesar de las distintas dificultades presentadas.

AGRADECIMIENTO

A mis profesores, con los cuales me permití ampliar

mis conocimientos y empeñarme en sobresalir en cada momento.

A mi familia, por siempre inculcarme valores

e impulsarme siempre a lograr mis metas.

TABLA DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ECUACIONES

RESUMEN

Actualmente, la abundante cantidad de datos que se tienen de clientes en las distintas empresas y el incremento en el uso de la tecnología han generado interés por profundizar la investigación sobre ello, además de desarrollar algoritmos y modelos para análisis de agrupamiento. Los modelos de clustering dirigidos al agrupamiento de clientes permiten a las organizaciones encontrar perfiles y patrones de servicios o compra, los cuales permiten generar estrategias para tomar mejores decisiones en la publicidad y canales con sus clientes. En la presente investigación, se realiza el análisis RFM (Recencia, Frecuencia y Dinero) para identificar los perfiles de clientes en un supermercado en base a sus iteraciones; luego, se emplea el algoritmo k-means para obtener los clústers adecuados y así identificar a los clientes potenciales. El objetivo de esta investigación es desarrollar un modelo de clusterización y generar la segmentación de clientes para el área comercial en un supermercado. Como resultado se tuvo que, a la hora de determinar qué tan leal es un cliente se obtuvo los siguientes perfiles: mejores clientes, clientes leales, los más gastadores, los casi muertos, los perdidos y los perdidos que son baratos; mientras que, el número óptimo de clústers o segmentaciones de clientes son dos, ya que gracias al Coeficiente de Silueta se determinó como valor de predicción un 84%.

Palabras claves: Análisis RFM**,** Perfiles de clientes, Segmentación de clientes, Coeficiente de Silueta.

CAPÍTULO I. INTRODUCCIÓN

1.1. Realidad problemática

En los últimos años, las grandes organizaciones vienen manejando como principal problema el no saber qué hacer con tanta información que manejan, ya sea a nivel interno o las fuentes de datos externas a las que puedan acceder y esto lleva a la pérdida de oportunidades de generar valor agregado a los diversos procesos que tengan y no tener el respaldo de tomar decisiones en base a datos o estadísticas, las cuales permitan ya sea describir la situación actual o analizar de cara al futuro (Mamani, Del Pino & Cortez, 2017).

Además, las áreas comerciales a nivel internacional buscan identificar y definir perfiles de clientes en base a resultados de análisis de sus comportamientos para que de esa forma se pueda identificar patrones, por consecuente, reglas y finalmente estructuras de asociación para clasificar los elementos no tan conocidos; lo cual lleva a la gestión del conocimiento empezando por la generación de todos los datos que se pueda obtener, luego, el descubrir de que se trata lo obtenido y como se relaciona, para finalmente llegar a la recolección de información que es la transformación de simples datos a lo que realmente genera valor (Marulanda, López & Mejía, 2017).

Actualmente, en nuestro país, distintos supermercados emplean como fuente de datos las iteraciones de clientes, pero se observa que en muchos casos la data no está integrada y gobernada, lo cual dificulta la mejora en la segmentación y perfilamiento de los clientes. Asimismo, la poca confianza o interés por parte de la parte ejecutiva dentro de los supermercados no ayuda a mejorar los procesos de explotación de datos mediante tecnología (Mamani, Del Pino & Cortez, 2017).

1.1.1 Antecedentes Internacionales

Las técnicas de minería de datos ya se ven y usan en diferentes países alrededor del mundo. Estas técnicas ya son parte de una estrategia dentro de las organizaciones para lograr sus objetivos a mediano y largo plazo. Se identificaron documentos donde estas técnicas se adecuaron a determinados procesos sin importar el rubro de la empresa.

Martínez & Hernández (2018) en su artículo de Técnicas de minería de datos con software libre para la detección de factores asociados al rendimiento, explican cómo realizaron un análisis de datos en base a técnicas de minería de datos con el objetivo de detectar factores que estaban asociados al rendimiento para así generar buenas prácticas educativas en España. Para ello, utilizaron como muestra, algunos modelos de predicción en una escuela específica, incorporando variables de entrada que hacían referencia al nivel socio económico de las familias, los recursos de la institución o el índice de repetición de grado por parte de los estudiantes; mientras que las variables dependientes se basaban en ciencias, matemática y lectura. Estos modelos estaban basados en un análisis de regresión, permitiendo reconocer la relación entre las variables independientes y dependientes, además de determinar el grado de influencia que se tiene con respecto a la variable dependiente. Así, bajo un enfoque de una metodología de investigación cuantitativa, llevaron la experimentación del modelo implementado, logrando identificar si el rendimiento medio obtenido en los resultados estaba por encima o debajo del objetivo planteado en un inicio.

Arcila, Marina & García (2018) en su artículo sobre "*Los enfoques del Big Data para la comunicación en salud: análisis de redes y análisis de sentimiento a gran escala*", indican como emplearon las dos técnicas más comunes de big data que se

utilizan en el análisis de sentimiento en la comunicación del sector salud en España. Señalan que la primera, se basa en un análisis automático de sentimientos usando diccionarios marcando palabras para generar un valor negativo o positivo, mientras que la segunda técnica, realiza un análisis supervisado, que se basa en el aprendizaje automático. Precisan, que en esta última técnica emplearon algoritmos de machine learning, para así crear modelos predictivos por cada sentimiento específico, los cuales pueden ser positivos, neutros, negativos, etc. Tuvieron una metodología de investigación cuantitativa, cuyo objetivo general era apreciar el impacto en el sector salud monitoreando las redes sociales a través de todo el proceso descrito anteriormente, tomando como muestra palabras que se relacionen con enfermedades o tratamientos médicos determinados. Como resultado, identificaron que el aprendizaje automático en base a un análisis supervisado logró aumentar la precisión en los sentimientos ya que se adaptó con mayor calidad al contexto de mensajes de salud gracias a que aborda el lenguaje lingüístico específico.

Ruiz & Romero (2018) en su artículo sobre los resultados obtenidos en un proceso de minería de datos aplicado a una base de datos que contienen información bibliográfica en Cuba, referida a cuatro segmentos de la ciencia; afirmaban que los objetivos de aplicar técnicas de minería de datos son mejorar la calidad de las diferentes fuentes de información, para así optimizar su funcionamiento, además de encontrar patrones ocultos que aporten mejoras en los diferentes procesos y productos de una organización; por último, el mejorar la gestión de datos y del conocimiento que se tenga del negocio. La metodología aplicada en su investigación estuvo basada en árboles de decisión y matriz de correlación, utilizando como herramienta digital Rapid Miner buscando asegurar la calidad de resultados y gráficos. El proceso de la aplicación consistió en la preparación

de los datos, identificando información externa e interna para seleccionar una muestra de datos necesaria; luego se validó la calidad de los datos recolectados para así llevarlos a la fase de transformación y preparación para el modelo analítico, utilizando el algoritmo apropiado. En la fase final se tuvo que llevar acabo la interpretación de los resultados obtenidos durante el proceso, mediante la ayuda de técnicas de visualización. En conclusión, los autores indican que, como resultado, lograron mejorar la información que contenía la base de datos bibliográfica, además de identificar patrones útiles que permitieron proponer nuevos productos y servicios para la biblioteca.

El análisis envolvente de datos es un complemento a la minería de datos que se utiliza muy a menudo para realizar estudios cuantitativos en Colombia, por lo que Visbal, Mendoza & Orjuela (2017) llevaron a cabo una investigación con esta característica, sobre la eficiencia técnica de las universidades públicas en este país, con el objetivo de generar variables que clasifiquen las instituciones en ineficientes o eficientes en base a los modelos predictivos formulados. Para este estudio emplearon modelos como Vecino más próximo, que consiste en clasificar casos basándose en su parecido a casos similares; Máquina de soporte vectorial, la cual representa a los puntos de muestra en el espacio relacionándose con la clasificación y regresión; Árbol de clasificación, donde la variable dependiente puede tomar un conjunto infinito de valores y Random Forest, que consta de muchos árboles de decisión; con los cuales se puede aplicar a datos sin clasificar el índice de error generado. Para validar estas técnicas se utilizaron Boostrapping y Boldout, los cuales se basan en el re muestreo y la validación simple con un conjunto de universidades que sirvió de muestra y entrenamiento; donde los resultados para el modelo de Árbol de clasificación fue que clasifica de manera eficiente 15 de 24 observaciones de validación, logrando una tasa de correcta clasificación de 62.5%; mientras que gracias a Random

Forest se logró detectar que las variables más importantes para medir la eficiencia de una institución son los alumnos matriculados en pregrado y postgrado, además de obtener solo un 16.67% de observaciones mal clasificadas; por otro lado, para el modelo de Vecino más próximo se clasificó de manera eficiente 13 de 24 observaciones dando una tasa de correcta clasificación de 54.17%; por último, en el modelo de Máquina de soporte vectorial se obtuvo que 15 de 24 observaciones se clasifican correctamente logrando una tasa de correcta clasificación de 83.33%.

Arce, Lima & Orellana (2018) en su investigación cuantitativa para descubrir patrones de comportamiento entre contaminantes del aire en la ciudad de Cuenca, emplearon diferentes algoritmos para poder así conocer los niveles de concentración en los contaminantes y como se relacionan estos. Dos de estas herramientas empleadas para clusterizar fueron X-medias y K-medias, cuyo objetivo era encontrar patrones de comportamiento en base a la contaminación del aire. Se tuvo que obtener datos por intervalos de minutos o segundos en diferentes partes de la ciudad, los cuales se utilizaron como muestra de estudio presentándose un monitoreo muy cambiante ya que cada localización tenía características particulares. Esta base de información pasó una limpieza para así tener una muestra clara para su estudio, logrando identificar correlaciones e incidencia entre cinco contaminantes atmosféricos nocivos en la región andina, además de lograr detectar que el O3 fue el contaminante más relevante en la Ciudad de Cuenca, con lo cual significa que el O3 tuvo una mayor repercusión que otros contaminantes. El resultado final de la evaluación con referencia a los algoritmos da muestra el positivo uso que se les dio para obtener los patrones adecuados entre varios contaminantes de aire en la investigación, con el objetivo de verificar el comportamiento continuo de los contaminantes a lo largo del día.

1.1.2 Antecedentes Nacionales

Del Pino & Cortez (2017) emplearon una metodología cuantitativa de minería de datos distribuida, para la predictibilidad en un proceso petitorio, teniendo como muestra las organizaciones públicas en el Perú, empleando como herramienta de clustering, el algoritmo k-means, el cual permite en base a patrones de afinidad, agrupar los diferentes indicadores que se tengan como fuente de datos. El objetivo del uso de esta técnica era generar una reducción en el volumen procesal, nivel de litigiosidad y plazos procesales, por lo que se brindará una mejor calidad en los servicios brindados a los ciudadanos. La metodología del desarrollo consistió en cinco fases: el diseño del modelo dimensional, donde se establece la arquitectura de los datos; el diseño del algoritmo de minería de datos distribuida, donde se eligió la técnica de clustering K-means, permitiendo agrupar los datos del Datamart según patrones de afinidad; la arquitectura del prototipo, se basó en un arquitectura de tres capas, aplicando el patrón arquitectónico Modelo Vista Controlador (MVC) y por último el análisis de los resultados. Como efecto de este desarrollo se identificó que el 42.8% de la carga procesal corresponden a trámites casatorios indiciados hace o años donde se solicita una bonificación especial por parte del demandante; también se tiene un 18.65% de casos iniciados hace ocho años donde se solicita una pensión de jubilación; además de un 16.51% de pedidos de pago de intereses penales; por último, un 22.04% de solicitudes de nulidad de resoluciones administrativas.

En la tesis "*Segmentación de clientes de un casino utilizando el algoritmo partición alrededor de medoides (PAM) con datos mixtos*" el autor indica que plantea como objetivo, el aplicar un algoritmo de partición alrededor de medoides para segmentar los clientes con los datos de sus tarjetas usadas en un casino a través de una metodología

cuantitativa. En su investigación, explica en que consiste la técnica de minería de datos que utilizará, indicando que este algoritmo busca particionar el conjunto de datos disponible en una cantidad de grupos ya definidos, que en este caso son tres; además del minimizar el porcentaje de disimilitudes entre un objeto y el centro del grupo (Medoide). El uso de este algoritmo en específico para su trabajo está argumentado en base a que está considerado como el más robusto ante datos atípicos. Luego, en el análisis del clustering, señala que, gracias a la medida de distancia, Gower, le arrojó 49.4%, 11.3% y 39.4% dentro de cada grupo para sus tres clústers respectivamente; gracias a esto logró generar su árbol de clasificación C5.0 logrando un 99.5% de precisión. Por último, indica que, como resultados para el proceso ya descrito, en el clúster 1 identificó que se tenían los promedios intermedios, mostrando que el 100% de uso de tarjeta es Classic, el 67% de clientes son hombres y un promedio de S/. 4,500 jugados al mes; mientras que para el clúster 2 se tenían los mejores promedios, observando que el 100% de clientes usan la tarjeta Silver, el 59% de ellos son hombres y llegan aproximadamente a un promedio de S/. 30,000 jugados al mes; además, en el clúster 3 se encontraron los promedios más bajos, llegando a tener que el 64% prefiere el uso de la tarjeta Classic, el 64% son hombres y un promedio de S/. 2,600 soles jugados al mes (Elguera, 2018).

Pacco (2015) se planteó el objetivo de mostrar los riesgos de morosidad por parte de los alumnos en la Universidad Peruana Unión a través de una investigación descriptiva desarrollada en base a redes neuronales con una metodología CRISP-DM, donde consideró sus cinco fases, la comprensión del negocio, la comprensión de datos, la preparación de datos, el modelado, la evaluación y el despliegue. Para esto empleó una solución de Business Analytics, comenzando por el análisis de los datos construyendo un datawarehouse para explotar la data operacional. Luego, construyó un ETL (Extract

Transform and Load) para extraer la información de los sistemas de producción de la empresa a un ambiente donde pueda hacer el desarrollo del modelo de árboles de clasificación con herramientas de Microsoft. Por último, muestra sus resultados en dos grupos, los morosos y no morosos; en el primero indica que hay una probabilidad de 84.21% de que el alumno sea moroso si es que estudia en otra institución, también, una probabilidad del 69.66% de que si el apoderado del alumno es Empleado este sea calificado como moroso. Mientras que, en el otro grupo, hay una probabilidad del 100% de que, si el alumno trabaja, no sea moroso; además de un 70% de probabilidad de que, si el alumno cuenta con apoyo financiero de sus padres, no sea moroso.

Mamani (2015) desarrolló una propuesta de aplicación usando técnicas de minería de datos con el objetivo de mejorar la calidad de los servicios en el poder judicial del Perú. En su investigación, aplicando una metodología no experimental, se enfocó en cuatro fases para el proceso del desarrollo del modelo de clustering. Empezó con el análisis del caso de estudio, donde pudo conocer la organización y el proceso de negocio del análisis. Luego, elaboró un modelo dimensional basándose en la metodología Kimball donde tuvo que definir el nivel de granularidad para así generar sus dimensiones y tabla de hechos. Una vez ya estructurados los datos recolectados del negocio, continuó con la aplicación del algoritmo k-medias como técnica de clustering con la finalidad de agrupar la información almacenada en su tabla de hechos según su afinidad, para ello definió cinco clústers como número de agrupaciones a obtener. En sus resultados señala que, en el primer clúster un aproximado del 50% fundamenta su caso procesal en base al D.U 037-94; en el segundo clúster, el 20% de casos se basan en la ley 23908 del código procesal civil; para el tercer clúster, un 17% sostiene su caso en la ley 24041; en el cuarto clúster un 15% tiene su caso como un proceso especial, pero basándose también en la ley

24041 y por último el quinto clúster presenta un 10% tiene su caso en base al Decreto de Ley 276 del Código Civil. Mamani señala que los resultados obtenidos permitirán un análisis en base los procesos resueltos y enfocados en el fallo.

En la tesis "*Aplicación de minería de datos para pronosticar el riesgo de morosidad de los estudiantes de la Universidad Autónoma del Perú*" de define como objetivo desarrollar una herramienta tecnológica que permita detectar a un posible alumno moroso. Para ello, se empleó un tipo de investigación experimental buscando evaluar la toma de decisiones de una mejor manera, además de correlacional ya que se midió el grade de relación entre los datos. Ya en el desarrollo mediante la metodología CRISP-DM, se tuvo una muestra de 1200 alumnos con un nivel de confianza del 95%, con lo cual se pudo generar pruebas con distintos modelos hasta hallar una solución óptima, esto ajustando algunas variables para ir midiendo el tiempo de respuesta de cada modelo. Como resultado, muestra la media de tiempo para diversos casos, por ejemplo, para identificar un alumno moroso, lleva 2 minutos; para predecir un alumno moroso, 3 minutos; para generar un reporte de los alumnos morosos, 4 minutos; para determinar qué cualidades tiene un alumno moroso, 5 minutos. En el trabajo presentado se puede validar los beneficios logrados gracias a un modelo predictivo, como es el caso del asertividad para los alumnos morosos y la reducción de tiempo en la detección de ello (Córdova & Torres, 2018).

1.1.3 Bases Teóricas

1.1.3.1 Clusterización

Flores (2016) precisa que el clustering divide una fuente de datos en agrupaciones diferentes, pero que no se debe confundir con la segmentación, ya que la segunda se basa al identificar grupos con las mismas características, mientras que el primero, segmenta datos en grupos no predefinidos, por lo que la clusterización es identificar grupos que son distintos a los otros, pero sus componentes son similares entre sí.

Según Prati & Baldoceda (2017), la principal ventaja del análisis por clúster es que se puede reconocer consumidores homogéneos, para que una vez agrupados, se pueda llevar un análisis por separado para identificar patrones o comportamientos de conducta.

Milla (2017) indica que, a diferencia de la clasificación, el clusterizar permite identificar símiles entre los objetos analizados, llegando a agruparlos en base a sus características en común y que los diferencian entre otros grupos.

Según Atalaya, Flores & Flores (2019) un método clúster es un desarrollo estadístico que empieza con una fuente de datos que almacena información en base a una muestra de objetos cuya finalidad es reorganizarlas en agrupaciones relativamente homogéneos.

1.1.3.1.1 Coeficiente de Silueta

Para reconocer que tanto se parecen o diferencian dos segmentaciones, se pueden tomar dos indicadores como es la cohesión y separación de clústers, por lo que a continuación se explica que es el coeficiente de silueta.

El coeficiente de Silueta es un indicador para evaluar la calidad de los clústers obtenidos con un algoritmo de clustering, con el fin de identificar el número óptimo de segmentaciones.

$$s(i) = \frac{b - a}{\max(a, b)}$$

Ecuación 1: Coeficiente de Silueta
Fuente: ingenieria.bogota.unal.edu.co

Donde:

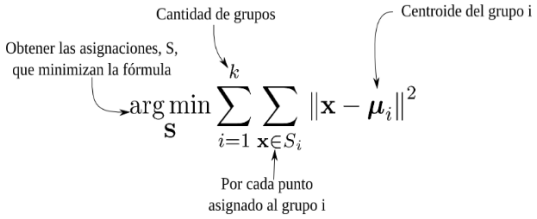
a = Es el promedio de las disimilitudes (o distancias) de la observación i con las demás observaciones del clúster al que pertenece i.

b = Es la distancia mínima a otro clúster que no es el mismo en el que está la observación i. Ese clúster es la segunda mejor opción para i y se lo denomina vecindad de i.

El coeficiente de Silueta [s(i)] es un valor comprendido entre -1 y 1.

1.1.3.1.2 Algoritmo k-means

Tipo de clustering donde se precisan variables cuantitativas u ordinales con un gran número de categorías para determinar número de segmentaciones, número máximo de iteraciones y número de ejecuciones.



Ecuación 2: Algoritmo k-means
Fuente: ingenieria.bogota.unal.edu.co

1.1.3.2 Análisis RFM

Es un método de segmentación de clientes donde ubica a estos en cada escalón de una pirámide de valor. Su análisis consiste en clasificar a los clientes según tres variables, que son, la recencia, la frecuencia y su valor monetario (Martínez & Hernández, 2018).

1.1.3.2.1 Diagrama de Cajas

Es una representación visual que permite describir ciertas características importantes al mismo tiempo, como son la dispersión y simetría. Esto se genera a través de tres cuartiles y los valores mínimos y máximos de los datos, los cuales genera un scoring de clientes para así asignarles un determinado perfil (Martínez & Hernández, 2018).

$$f_i = \frac{i-1}{n-1}$$

Ecuación 3: Cuartil de un indicador
Fuente: ingenieria.bogota.unal.edu.co

1.1.3.3 Minería de datos

Atalaya, Flores & Flores (2019) señalan que, para obtener información procesable dentro de un conjunto grande de datos, es necesario utilizar minería de datos, ya que permite descubrir tendencias y patrones que existen en la información.

Arce, Lima & Orellana (2018) indica que el datamining surge para ayudar a entender el contenido de una fuente de datos, usando métodos estadísticos o también algoritmos en base a inteligencia a artificial y redes neuronales. Resaltan que los datos son la materia prima, por lo que cuando se elabora un modelo hace que la interpretación de ello con la información represente un valor agregado, por lo que brinda conocimiento.

Es el proceso de encontrar nuevos patrones y adquisición de conocimiento a raíz de una gran cantidad de datos. Las fuentes para esta información pueden ser: Base de datos, internet u otros sistemas o repositorios que ingresen al sistema. La minería de datos se enfoca en cubrir el requerimiento de analizar de manera automática e inteligente los datos generados día a día. Esta cantidad de datos no puede ser analizado por una sola persona, ya que la toma de decisiones no se puede basar en la intuición, por lo que debe tener sustento en alguna información (Yamao, 2018).

Para Visbal, Mendoza & Orjuega (2017) las técnicas de minería de datos se agrupan en dos conjuntos según su objetivo de análisis; las técnicas de aprendizaje supervisado, donde una variable debe ser contextualizada por las otras y técnicas de aprendizaje no supervisada, donde no existe una variable principal o específica, por lo que no están previamente clasificados.

Para Prati & Baldoceda (2017), el aplicar técnicas para encontrar predicciones de cara al futuro tomando como base datos pasados define el análisis predictivo. Este, toma resultados conocidos para predecir valores entrenando los diferentes modelos empleados. También, indica que algunas de estas técnicas incluyen reconocimiento de patrones y minería de datos.

Atalaya, Flores & Flores (2019) indican que un análisis predictivo siempre comienza como un objetivo comercial, ya que se pretende ahorrar tiempo, consolidar la información en solo lo necesario y reducir costos. Señalan que, este proceso utiliza fuentes de datos heterogéneos para así generar resultados transparentes y concisos para lograr el propósito inicial, ya sea fabricar productos o reducir data inservible.

Según Flores (2016), el análisis predictivo reside en extraer datos históricos para llevarlo a un modelo analítico para así predecir un comportamiento a futuro o también estimar resultados no conocidos. Resaltan que allí se emplean diversas técnicas de estadística, como minería de datos, modelización o aprendizaje automático con la finalidad de reunir todos los datos para generar predicciones.

Arce, Lima & Orellana (2018) afirman que la aplicación de algunas consultas analíticas y algoritmos automáticos a algún grupo de datos permiten generar modelos predictivos que determinan un valor numérico o una probabilidad de que suceda un evento particular.

1.1.3.4 Medición

1.1.3.4.1 Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson, tiene como finalidad el indicar el nivel de asociación que tienen dos variables entre sí.

$$r = \frac{\alpha xy}{\alpha_x * \alpha_y}$$

Ecuación 4: Coeficiente de correlación de Pearson
Fuente: ingenieria.bogota.unal.edu.co

Donde:

r = Es un coeficiente

Si r ≈ 1, existe una correlación directa fuerte

Si r ≈ -1, existe una correlación indirecta fuerte

Si r = 1 o r = -1, hay una correlación funcional

Si r ≈ 0, no existe una correlación lineal

1.1.3.4.2 Kolmogorov-Smirnov

El test de Kolmogorov-Smirnov se utiliza para validar si un conjunto de datos sigue una distribución normal o no, siempre y cuando la cantidad de datos supere a un número de cincuenta.

$$D = \sup_{1 \le i \le n} |\hat{F}_n(x_i) - F_0(x_i)$$

Ecuación 5: Fórmula de Kolmogorov-Smirnov
Fuente: ingenieria.bogota.unal.edu.co

1.1.3.5 Glosario

- **Machine Learning**

Disciplina que permite generar máquinas que aprendan de forma automática, esto a través de algoritmos que pueden deducir y predecir comportamientos en base a la historia (Díaz, 2017).

- **Aprendizaje supervisado**

También conocido como método predictivo, ya que en base a varias variables se debe explicar una referencial, por lo que ya se debe tener una clasificación establecida de los datos a estudiar (Visbal, Mendoza & Orjuela, 2017).

- **Aprendizaje no supervisado**

Llamado también método descriptivo, ya que es en donde no se cuenta con una variable referencial explicada por otras, por lo que no se cuenta con una clasificación establecida de los datos a analizar (Visbal, Mendoza & Orjuela, 2017).

- **Data Governance**

La gobernanza de los datos ayuda a una correcta gestión de la información para que así se genere confianza entre los usuarios de cada a la autenticidad, credibilidad e integridad de los datos (García, 2016).

- **Data Training y Data Testing**

Entrenar o experimentar modelos con datos para que así cada vez tenga resultados más precisos e identifique relaciones que permitan tomar decisiones en base a la confianza que genera (Rienties, Cross, Marsh & Thomas, 2017).

- **Dataset**

Es una colección o representación de datos en memoria con un modelo relacional que cuentan con una estructura definida (Rienties, Cross, Marsh & Thomas, 2017).

1.1.3.6 Definición de términos

1.1.3.6.1 Jupyter Notebook

Es un entorno de trabajo de código abierto que permite generar archivos en formato JSON siguiendo una estructura versionada y una lista de celdas de entradas y salida, en las cuales se pueden escribir, documentar y ejecutar código desde cualquier navegador estándar. Esta aplicación contiene dos elementos principales, los cuales son, un conjunto de núcleos que sirven de motor para recepcionar solicitudes y devolver las respuestas apropiadas y también el dashboard que es la interfaz de cara al usuario. Para esta investigación se utilizará el núcleo que permite trabajar con Python y aprovechar las principales funcionalidades que son la depuración de datos, modelización estadística, creación y entrenamiento de modelos de aprendizaje automático y la visualización de datos (Martínez & Hernández, 2018).

1.1.3.6.2 Perfil de Cliente

La perfilación de clientes se basa en identificar las características de un cliente para un apropiado tratamiento y asignación de una determinada clasificación (Simonato, 2018).

1.2. Formulación del problema

1.2.1 Pregunta general

¿De qué manera un modelo de clusterización segmenta el perfil del cliente para el área comercial en supermercados?

1.2.2 Preguntas específicas

a) ¿Cómo se puede identificar los indicadores de correlación para el área comercial en supermercados?

b) ¿Cómo se puede identificar el perfil de lealtad en base a los indicadores de cliente para el área comercial en supermercados?

c) ¿Cómo se puede medir la relación entre el modelo de clusterización y los indicadores de predicción para el área comercial en supermercados?

1.3. Objetivos

1.3.1 Objetivo general

Implementar un modelo de clusterización en la segmentación del perfil del cliente para el área comercial de supermercados.

1.3.2 Objetivos específicos

a) Identificar los indicadores de correlación para el área comercial en supermercados.

b) Identificar el perfil de lealtad en base a los indicadores de cliente para el área comercial en supermercados.

c) Evaluar la relación entre el modelo de clusterización y los indicadores de predicción para el área comercial en supermercados.

1.4. Hipótesis

1.4.1 Hipótesis general

El modelo de clusterización segmenta correctamente el perfil del cliente en base a los indicadores de correlación, cliente y predicción.

1.4.2 Hipótesis específicas

a) Los indicadores de correlación y el modelo de clusterización está entre -1 y 1.

b) El perfil de lealtad se basa en los puntajes entre 1 y 4 de los indicadores de cliente.

c) El nivel de relación entre los indicadores de predicción y el modelo de clusterización supera el 70%.

1.5. Justificación

Del análisis realizado en el presente trabajo, se observa que existen técnicas de minería de datos comunes y reutilizables, los cuales pueden ayudar a empezar a encontrar patrones de comportamiento y que se pueden ir adaptando a la necesidad de cada proceso de negocio, ya que la idea es llegar a la mayor eficiencia al momento de segmentar la información (Atalaya, Flores & Flores, 2019).

Ante ello, una de estas técnicas fue empleada por Del Pino & Cortez (2017) en su investigación sobre procesos petitorios, ya que el algoritmo k-means les permitió identificar en base a un nivel de relación entre 0.84 y 0.99 de los indicadores de predicción a través de un modelo de clusterización y la segmentación de clientes en los procesos petitorios.

Además, tomando como referencia la investigación de Elguera (2018) sobre la clusterización de clientes para un casino, se puede ir midiendo la eficiencia de un modelo de clusterización en base al número de clústers, hasta encontrar el óptimo, ya que él menciona que esto depende mucho de qué tipo de variables se manejen en la base de datos que se tenga, por lo que tuvo que evaluar el nivel de relación entre los indicadores de correlación de las variables y la segmentación de clientes para así definir, justamente que variables debía considerar en el modelo.

Teniendo en cuenta las desventajas que tienen los supermercados para predecir aquellos clientes que son potenciales re compradores, es importante investigar algún modelo que permita mejorar este proceso de identificación.

Por consiguiente, la presente investigación se enfocará en la implementación de un modelo de clusterización en la segmentación del perfil del cliente para el área comercial de supermercados, ya que según lo que indica Simonato (2018) en su libro sobre la innovación en un área comercial a través de la gestión de experiencias, en las áreas comerciales para los distintos análisis de clientes, normalmente solo se explota entre el 50% y 60% de los datos disponibles, no permitiendo identificar un perfil para cada cliente; por lo que, la implementación que se ve en esta investigación, permitirá utilizar el 100% de los datos disponibles y además, a obtener un perfilamiento de clientes según sus características basadas en comportamientos y tendencias.

CAPÍTULO II. METODOLOGÍA

2.1 Operacionalización de la variable

En la tabla Nro. 1 se muestra el detalle de la operacionalización de las variables de esta investigación.

Tabla 1:

Operacionalización de la variable

Variables	Dimensiones	Indicadores	Unidad de medida	Fórmula	Instrumentos
Perfil de cliente	Indicadores de correlación	Correlación entre variables Correlación con respecto a la predicción	Nivel de correlación	$r = \frac{\alpha xy}{\alpha_x * \alpha_y}$	Guía de observación
	Indicadores de cliente	Cuartil de Recencia Cuartil de Frecuencia Cuartil de Monto de venta	Recencia Frecuencia Monto de venta	$f_i = \frac{i-1}{n-1}$	
	Indicadores de predicción	Porcentaje de predicción Número de clústers Número de clientes potenciales	Porcentaje Número Número	$s(i) = \frac{b-a}{\max(a,b)}$	

2.2 Tipo de investigación

2.2.1 Diseño de investigación

En base a lo que indica Supo (2020) en su libro sobre la metodología de investigación científica, esta investigación según su profundidad es aplicativa, ya que su objetivo general es implementar un modelo de clusterización en la segmentación de clientes en el área comercial de supermercados.

Además, según su naturaleza de datos es cuantitativa (Supo, 2020), ya que se usarán datos y variables cuantificables, para luego, realizar experimentos a través de un algoritmo e ir midiendo la eficiencia del modelo empleado y así obtener el número óptimo de clústers y clientes potenciales para el área comercial en supermercados.

Por último, según Sampieri (2018), el tipo de investigación es Preexperimental, debido a que se hace una intervención de los datos en base a la implementación del modelo de clusterización con el que se obtendrá el número óptimo de clústers y clientes potenciales para el área comercial en supermercados.

2.2.2 Diseño del experimento

En base a lo indicado anteriormente, en el presente trabajo se empleará una investigación experimental, por lo que se está considerando como pre experimental solo con un postest. Este diseño consistirá en dos etapas, tal como se visualiza en la figura 1.

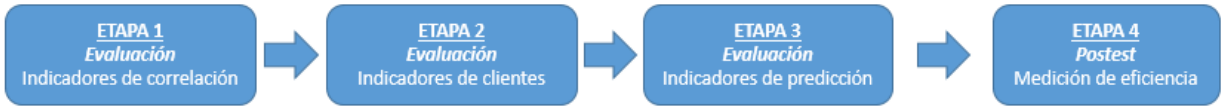


Figura 1: Etapas del diseño de experimento

En la primera etapa, se evalúa la relación entre las variables de la base de datos para que según su nivel de correlación definir cuáles son las más apropiadas para un posterior análisis e intervención en el modelo de clusterización.

Luego, en la segunda etapa de esta investigación, se obtienen los indicadores de Recencia, Frecuencia y Monto de venta como indicadores de clientes para definir el perfilamiento de clientes.

Después, con respecto a los indicadores de predicción, a través de las pruebas con el modelo de clusterización se registran en la guía de observación los datos recolectados de porcentaje de predicción, número de clústers y número de clientes potenciales.

Por último, en la etapa de post test se mide la correlación entre las variables seleccionadas anteriormente y que se utilizaron en el modelo, con respecto al porcentaje de predicción obtenido; además, se valida la normalidad de los datos recolectados en la tercera etapa mediante el test de Kolmogorov-Smirnov.

Por ello, se considera esta investigación como aplicativa, ya que lo que se busca es medir la relación de un modelo de clusterización con la segmentación de perfil de clientes para el área comercial en supermercados.

2.3 Materiales, instrumentos y métodos

2.3.1 Materiales

Los materiales que se emplean en ambas etapas, previamente detalladas en el diseño del experimento, son los que se muestran en la tabla 2:

Tabla 2:

Materiales – Software

Tipo	Nombre	Descripción	Cantidad
Software	Windows 10	Sistema operativo	1
	Excel	Herramienta de Base de datos	1
	Jupyter Notebook	Herramienta de desarrollo	1

En la primera etapa, los materiales descritos se usaron para realizar un análisis de la base de datos que se tiene y así definir los perfiles en los que se segmentarán a los clientes potenciales. Mientras que, en la segunda etapa apoyaron en la medición de qué tan óptimos son los clústers generados por el modelo de clusterización.

2.3.2 Instrumentos

Según Hernández, Fernández & Baptista (2014) un instrumento es todo aquello que se emplea para registrar la información que se medirá con respecto a las variables en una investigación. Teniendo en cuenta a los autores mencionados, este trabajo cuenta con dos variables de investigación, tal como fue operacionalizada en el 2.1, y para cada uno de ellos se usarán herramientas válidas, confiables y objetivas. Por ende, se usarán los siguientes instrumentos para determinar el nivel de relación entre el modelo de clusterización y los indicadores de correlación, cliente y predicción, respectivamente:

- **Guía de observación - Indicadores de correlación**

  Este instrumento se empleará para evaluar y registrar que variables de la base de datos utilizada tienen un mayor nivel de correlación entre ellos, para que, en base a los seleccionados, se realicé un análisis de datos. Esta evaluación se medirá en base al coeficiente de correlación a través del método de cuadrado de Pearson, explicado en la base teórica (Véase el punto 1.1.3.4.1).

  El instrumento fue diseñado tomando en cuenta las variables a correlacionar, con el fin de evaluar la relación entre el modelo de clusterización y los indicadores de correlación. Este instrumento fue validado por tres profesionales especialistas en la investigación científica de ingeniería. (Véase el anexo 2)

- **Guía de observación - Indicadores de cliente**

Este instrumento se empleará para evaluar y registrar los indicadores de Recencia, Frecuencia y Monto de venta, los cuales son, que tan reciente ha comprado un cliente, con qué frecuencia lo hace y el consumo total monetario de venta respectivamente. Esta evaluación se mediará en base a los cuartiles del Diagrama de cajas (Véase el punto 1.1.3.2.1)

El instrumento fue diseñado tomando en cuenta las variables a medir, con el fin de evaluar la relación entre el modelo de clusterización y los indicadores de cliente. Este instrumento fue validado por tres profesionales especialistas en la investigación científica de ingeniería. (Véase el anexo 3)

- **Guía de observación - Indicadores de predicción**

Este instrumento se empleará para evaluar y registrar el porcentaje de predicción, el número de clústers y los clientes potenciales obtenidos en las pruebas del modelo. Esta evaluación se mediará en base al Coeficiente de Silueta (Véase el punto 1.1.3.1.1) empleado en el algoritmo k-means (Véase el punto 1.1.3.1.2).

El instrumento fue diseñado tomando en cuenta las variables a medir, con el fin de evaluar la relación entre el modelo de clusterización y los indicadores de predicción. Este instrumento fue validado por tres profesionales especialistas en la investigación científica de ingeniería. (Véase el anexo 4)

2.3.3 Población y Muestra

En este trabajo de investigación se está considerando como población a todas las iteraciones de clientes realizadas en un supermercado en el periodo entre los años 2019 y 2020 en nuestro país, encontradas en una base de datos abierta, la cual, según indican Aleixandre, Ferrer & Peset (2019) es libre de usar, reutilizar y distribuir con el fin de compartir datos en investigaciones y así maximizar el potencial de los mismos; por lo cual no se está tomando alguna muestra sobre ella y se considera como unidad de estudio a cada iteración por cliente, las cuales se describen a continuación.

Muestra

Debido a que la población incluye a diferentes regiones del Perú, se está tomando como muestra 7,726 iteraciones de clientes en la región de Lima, ya que las compras comerciales suelen variar de acuerdo a las costumbres de sus ciudadanos, es así que el tipo de muestreo realizado fue por conveniencia, para un mejor análisis de correlación de variables de las iteraciones de los clientes.

Unidad de Estudio

Un supermercado tiene una cadena de locales distribuidas en todo el país, donde se generan millones de iteraciones de personas de manera diaria. Estos datos contienen diversas variables que permiten tener una trazabilidad correcta de una compra por parte de un cliente en cualquier local; esto como, el identificador del cliente (anonimizado), el número de ticket, la fecha de compra, el nombre de producto, categoría del producto,

cantidad de ítems por producto, ubigeo de la sucursal, monto total del ticket y la forma como realizar el pago del mismo.

De esta base, se consideran a los clientes de ambos sexos sin importar la edad que tengan siempre y cuando cuenten con la facultad de pagar con una tarjeta de crédito o débito, esto en todas las sucursales a nivel nacional de un supermercado en el periodo entre los años 2019 y 2020, mediante cualquier tipo de tarjeta MasterCard, Visa, MasterCard Internacional o Clásica, además que estén basadas en las compras de productos de las categorías Abarrotes, Frescos, Misceláneos y Non Food.

2.3.4 Estructura de trabajo

En la Figura 2 se muestra la estructura del trabajo enfocado en los objetivos planteados, donde se detalla sus fases y las relaciones entre sus tareas.

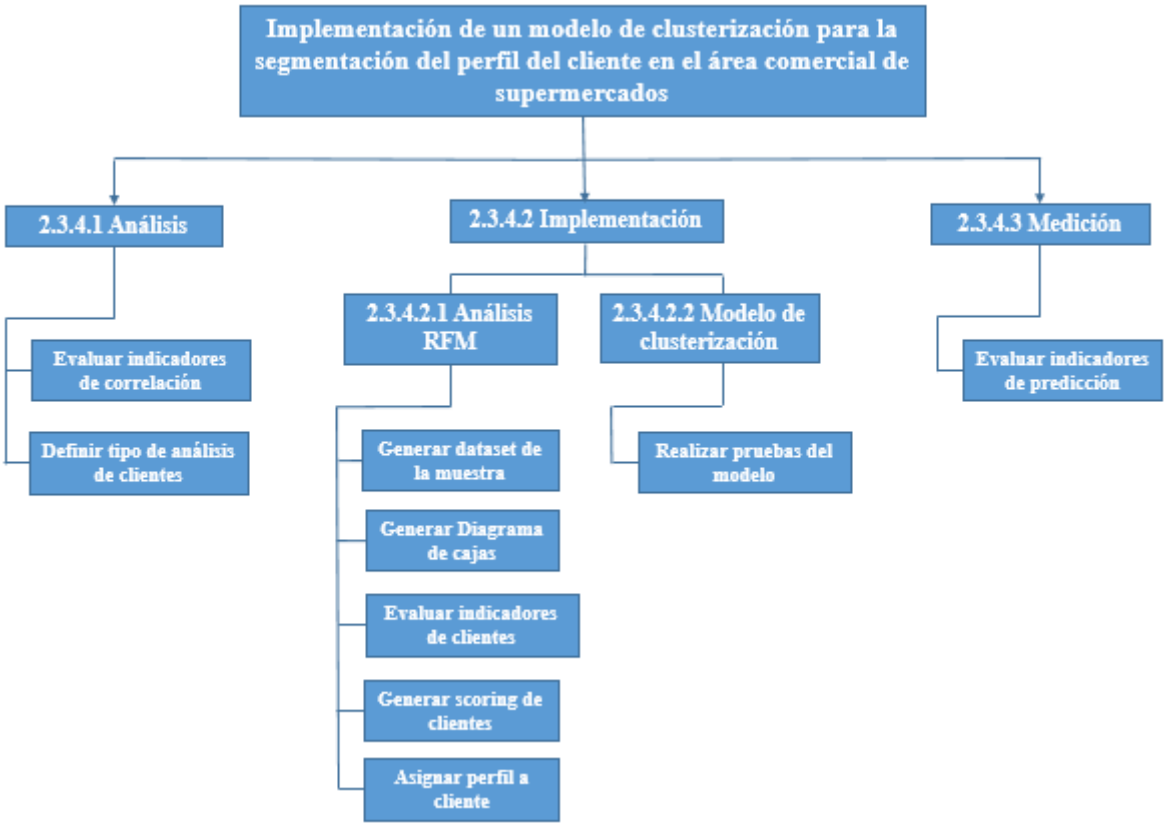


Figura 2: Estructura de trabajo

A continuación, se detalla en qué consistió cada fase y sus tareas respectivas.

2.3.4.1 Análisis

En esta etapa inicial se definió cuáles eran las variables de la base de clientes que tenían un mayor nivel de correlación para así posteriormente realizar el análisis.

Evaluar indicadores de correlación

Se comenzó por un análisis descriptivo de algunas variables; tal y como muestra la Tabla 3, se validó el número de operaciones o tickets, el número de productos únicos vendidos, el número de clientes únicos que realizaron una compra y el número de distritos donde se realizó al menos una venta.

Tabla 3:

Análisis descriptivo de variables

Indicador	Cantidad
Número de operaciones	14,268
Número de productos únicos	674
Número de clientes	11,430
Número de distritos	398

Además, se validó que no exista ningún registro de venta con un valor negativo, dando como resultado 0 registros con esta especificación; logrando así no tener problemas a la hora de generar los datasets.

También, como se ve en la Figura 3, se obtuvo un gráfico para verificar el número de clientes por región.

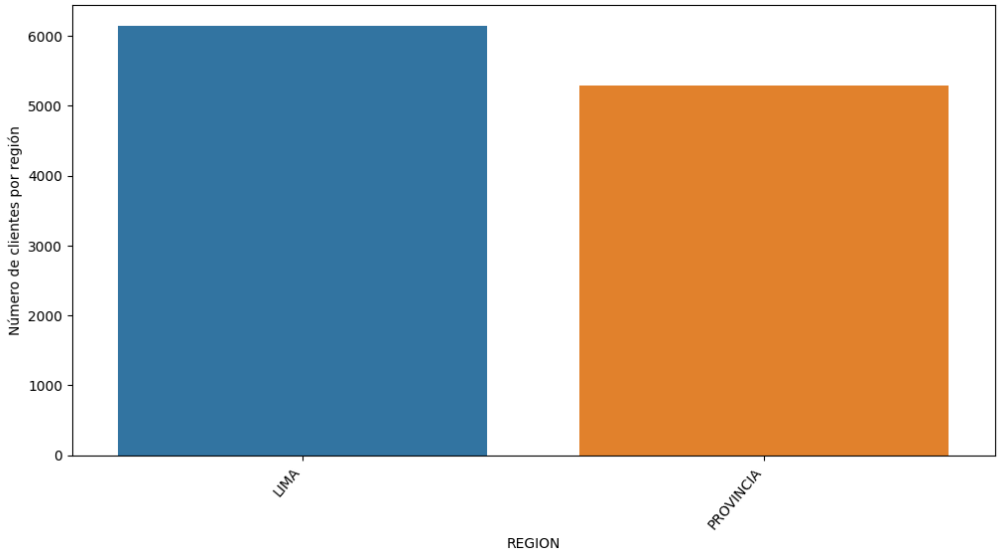


Figura 3: Número de clientes por región

Mientras que, en la Figura 4, se muestra el número de clientes que realizó una compra agrupado por categoría del producto adquirido.

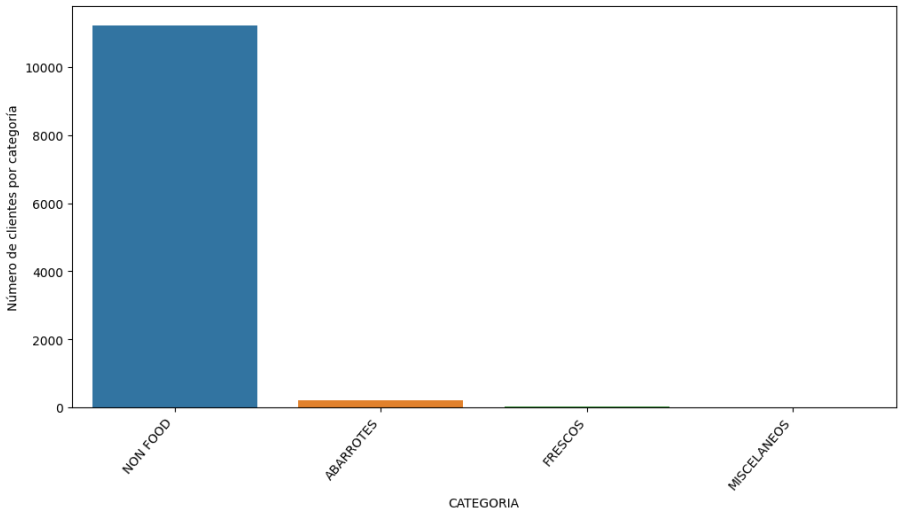


Figura 4: Número de clientes por categoría

Luego, se generó un gráfico en base a la cantidad de clientes por el uso del tipo de tarjeta a la hora de realizar su compra (Véase Figura 5).

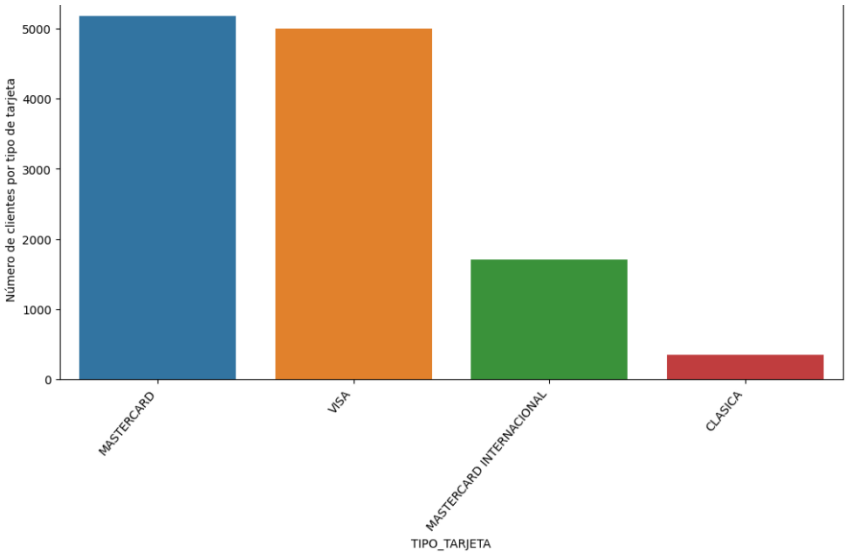


Figura 5: Número de clientes por tipo de tarjeta

Por último, se midió el nivel de correlación que había entre todas las variables a través del método de cuadrado de Pearson, el cual nos brinda un coeficiente de correlación entre variables (Véase el punto 1.1.3.4.1), tal como muestra la Figura 6.

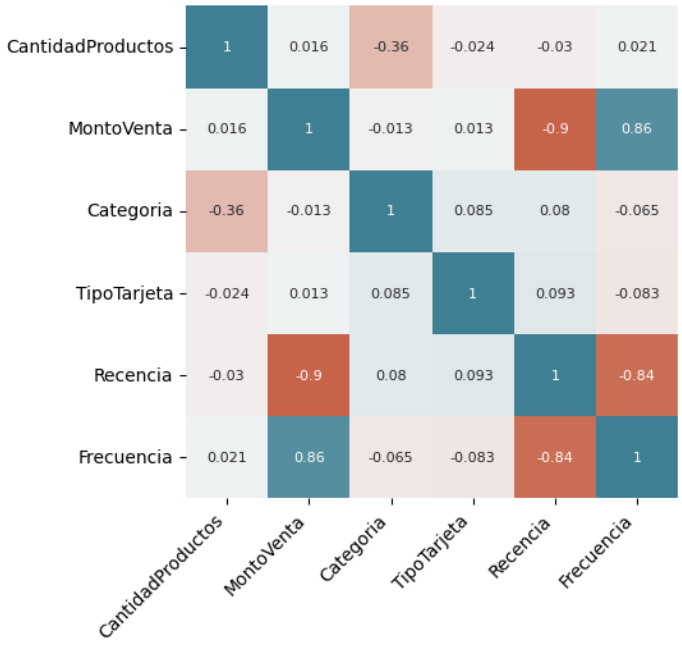


Figura 6: Evaluación de correlación entre variables

Definir tipo de análisis de clientes

Tras ver que variables tenían un mayor nivel de correlación, como son la Recencia, Frecuencia y Monto de venta, se creyó pertinente realizar el análisis RFM ya que este está basado en qué tan reciente ha comprado un cliente, que tan frecuente lo hace y cuanto es su consumo total a nivel monetario, permitiendo clasificarlo de manera adecuada.

2.3.4.2 Implementación

Esta etapa se comienza por el análisis RFM para llegar a segmentar a los clientes según el scoring generado de sus iteraciones.

2.3.4.2.1 Análisis RFM

Se definieron los perfiles de clientes por su score según tres variables, que son la Recencia, que indica los días transcurridos desde su última compra, la Frecuencia, que nos brinda el número de compras y el Monto de venta, que señala el total del consumo. Este análisis cuenta con cuatro tareas principales, las cuales se detallan a continuación.

Generar dataset de la muestra

Se generó un dataset en base la región de Lima debido a que, filtrando por este indicador, se tiene un mayor nivel de correlación entre variables y como se puede observar en la Tabla 4, se obtuvo el siguiente análisis.

Tabla 4:

Análisis del clúster de iteraciones en Lima

Indicador	Cantidad
Número de transacciones	7,726
Número de productos únicos	568
Número de clientes únicos	6,140

Además, como se observa en la Tabla 5, se obtuvo el Top 10 de los productos más vendidos en esta región.

Tabla 5:

Top de productos del dataset de iteraciones en Lima

Código	Nombre Producto	Cantidad
90803877	Claro pp chip portabilidad	24,502
107285	Soya aceite de soya refinada bt900ml	15,072
21592	Bell-s aceite de soya bt900ml	13,100
959308	Bell-s aceite vegetal bt900ml	11,420
979253	Pilsen cerveza pk 12 lt 355 ml	10,893
20080359	Bonle mezcla láctea familiar cj500gr6pk	6,020
199315	Gloria leche evap entera lt400gr 6pk	5,942
20110463	Gloria leche evap ninos lt400gr 6pk	5,036
90806745	Entel pc celular	4,874
90806743	Entel pp celular	4,815

Generar el Diagrama de cajas

Como se puede observar en la Tabla 6, en esta tarea se generaron tres cuartiles en base a las variables antes señaladas.

Tabla 6:

Cuartiles del Diagrama de cajas

	Días recientes	Frecuencia	Costo total
0.25	370	1	2,199.00
0.50	461	1	2,499.00
0.75	553	1	3,698.25

Evaluar indicadores de clientes

Aquí se le asigna un puntaje a cada indicador de cliente, los cuales son la Recencia, Frecuencia y Monto de venta y se los agrega en una columna cada uno; esto se explica en el punto 1.1.3.2.1 basado en el análisis de datos realizado.

Generar scoring de clientes

En esta tarea, como se puede ver en la Tabla 7, se genera un scoring RFM en base a los cuartiles ya obtenidos en la tarea anterior, el cual representa la unión de los 3 indicadores que son, Recencia, Frecuencia y Monto de venta.

Tabla 7:

Scoring de clientes

Cliente	Recencia	Frecuencia	Monto de venta	Cuartil R	Cuartil F	Cuartil M	Scoring RFM
1	633	1	2,799.00	1	1	3	113
2	534	1	2,299.00	2	1	2	212
3	559	1	2,299.00	1	1	2	112
4	572	1	3,499.00	1	1	3	113
5	643	1	2,399.00	1	1	2	112
6	291	38	859,898.00	4	4	4	444

Asignar perfil a cliente

Finalmente, en base a los scoring obtenidos por cada cliente, y como se observa en la Tabla 8, se agruparon a estos para así identificarlos y clasificarlos en base a qué tan leales son (Véase el punto 1.1.3.6.2).

Tabla 8:

Perfiles de clientes

Perfil	Q de Clientes
Mejores Clientes	541
Clientes leales	648
Clientes más gastadores	2,303
Clientes casi muertos	216
Clientes perdidos	115
Clientes perdidos que son baratos	1220
Otros	6,387

2.3.4.2.2 Modelo de clusterización

Se realizan las pruebas del modelo para posteriormente, evaluar el nivel de relación de los indicadores de predicción.

Realizar pruebas del modelo

Se realizó las pruebas con el algoritmo k-means (Véase el punto 1.1.3.1.2) según el número de clústers deseado para así, gracias al Coeficiente de Silueta (Véase el punto 1.1.3.1.1) ver cuál era la homogeneidad entre clústers a través del porcentaje de predicción de cada resultado obtenido, por lo cual, se optó por evaluar desde 2 a 9 clústers.

2.3.4.3 Medición

En esta última etapa se evaluaron los indicadores de predicción para así obtener el porcentaje de predicción, número de clústers y número de clientes potenciales.

Evaluar indicadores de predicción

Se realizó la evaluación en base a la combinación de las 3 variables utilizadas en el análisis RFM, los cuales son Recencia, Frecuencia y Monto de venta, de los cuales se obtuvo que solo considerando Frecuencia y Monto de venta se tenía un mayor porcentaje de predicción.

2.4 Procedimiento

2.4.1 Proceso de recolección de datos

Recolección de la base de datos a utilizar

Para el proceso de recolección de la base de datos a utilizar, se comenzó con la búsqueda de datos abiertos en la nube que se orienten a las iteraciones de clientes en un supermercado; ya que, según Aleixandre, Ferrer & Peset (2019) en su artículo señalan que un dato es abierto si alguien es libre de usarlo, reutilizarlo y distribuirlo, esto basándose en lo que dice la organización Open Knowledge International sobre la importancia de compartir datos en investigaciones para así maximizar el potencial de los mismos.

Ante esto, el portal donde se encontró una estructura de datos oportuna es "*data.world*", filtrando como búsqueda específica "*transacciones con tarjeta de crédito*". Los responsables de este portal se encargan, mediante diversas tecnologías de recolectar datos abiertos de varios países, la cual mayormente es de la parte pública de los mismos; además de tener acuerdos con empresas privadas para que estas le proporcionen información que pueda ser compartida con el público en general.

Posteriormente, se delimitó la estrategia de búsqueda basándose que los datasets encontrados contengan datos a nivel nacional, detallado por distritos, que los datos no superen una antigüedad de dos años, que los tickets de venta estén detallados a nivel producto y estos cuenten con una categoría a la cual se pueden asociar. También, se verificó que cada registro de venta debía contar con un código de cliente anonimizado, para así identificarlos, ya que el medio de pago debía ser mediante algún tipo de tarjeta.

A continuación, en la Figura 7 se resume el flujo del proceso de recolección de la base de datos a utilizar.

Figura 7: Proceso de recolección de la base de datos

Recolección de datos de los instrumentos empleados

Para poder validar las hipótesis planteadas inicialmente en el estudio, se crearon y validaron tres instrumentos que permitieron registrar y validar datos en las tres primeras etapas de la investigación.

En la etapa 1, se recolectaron los datos sobre los indicadores de correlación, a través de una guía de observación (Véase anexo 2), el cual permitió identificar las variables con mayor nivel de correlación entre sí, mediante el coeficiente de correlación de Pearson.

Luego, en la etapa 2, se recolectaron datos sobre los indicadores de cliente a través de una guía de observación (Véase anexo 3), la cual permitió registrar el puntaje del cuartil asignado a cada uno de los tres indicadores que se evaluaron mediante el diagrama de cajas.

Por último, en la etapa 3, para la recolección de datos de los indicadores de la predicción también se empleó una guía de observación (Véase anexo 4) donde se registraron las pruebas del modelo de clusterización que permitió identificar el porcentaje de predicción, el número de clústers y el número de clientes potenciales.

A continuación, en la Figura 8 se resume el flujo del proceso de recolección de datos de los instrumentos empleados.

Figura 8: Proceso de recolección de datos de los instrumentos

2.4.2 Proceso de análisis de datos

Para poder validar las hipótesis planteadas inicialmente en el estudio, se analizaron los datos recolectados en los tres instrumentos antes mencionados.

En la primera etapa, se evaluó el nivel de relación entre todos los indicadores que tenía el conjunto de datos, con la finalidad de identificar cuáles eran las variables más óptimas para realizar un análisis posterior e incluirlas en el modelo de clusterización; en la cual se identificó que las tres variables con mayor nivel de correlación eran la Recencia, Frecuencia y Monto de venta.

Luego, en la segunda etapa, se evaluaron los cuartiles de las tres variables mencionadas en el párrafo anterior, para lo cual se establecieron que los valores representados; para el primer cuartil (0.25), era la media aritmética del valor mayor del 25% de la distribución de todos los datos y su valor siguiente; para el segundo cuartil (0.5), era la media aritmética de la mediana de todos los datos y su valor siguiente; por último, para el tercer cuartil (0.75), era la media aritmética del valor que supera al 75% de todos los datos y el valor siguiente. Entonces, se evaluó los indicadores de cliente en base a las tres variables antes mencionados y a los cuartiles generados, se pudo obtener un score o puntuación para cada cliente, donde las puntuaciones eran 1, 2, 3 y 4 en base a que cuartil este clasificado; por ejemplo, si un indicador estaba en el primer cuartil, tenía una puntuación de 1; para el segundo cuartil, la puntuación era 2; en el tercer cuartil, la puntuación era 3; por último, si el indicador superaba el monto del tercer cuartil, tendría una puntuación de 4.

Además, en la tercera etapa, para la evaluación de los indicadores de predicción se hicieron las pruebas con el modelo considerando la combinación de las tres variables indicadas anteriormente, donde en base a la Frecuencia y Monto de venta de obtuvo el mayor porcentaje de predicción, con lo cual se considera que son las dos variables óptimas para el modelo.

Por último, mediante el test de Kolmogorov-Smirnov (Véase el punto 1.1.3.4.2) se pudo medir la distribución normal de los datos registrados en el instrumento de indicadores de predicción, obteniendo un valor de significancia de 0.74 el cual es muy superior al nivel de confianza de 0.05 que señala que se acepta la hipótesis nula, por lo cual se concluye que las variables tienen una distribución normal.

2.5. Aspectos éticos

En la presente investigación se consideran los aspectos éticos oportunos en cuanto a la correcta citación de fuentes empleando las normas del manual de redacción de la UPN. Ante ello, también se presentan datos confiables, fidedignos y ajustados a la investigación realizada.

Además, se asegura el buen uso de los datos de transacciones de la base de datos abierta empleada esto basado en lo que señala la Organización Open Knowledge International sobre la importancia de compartir datos en investigaciones (Aleixandre, Ferrer & Peset, 2019), así como ningún acto que vaya en contra de lo legal por parte del autor del presente trabajo.

CAPÍTULO III. RESULTADOS

En la Figura 9 se puede apreciar los registros de las iteraciones de clientes realizadas en el periodo entre 2019 y 2020, en todos los locales del supermercado agrupado por la región del local de venta.

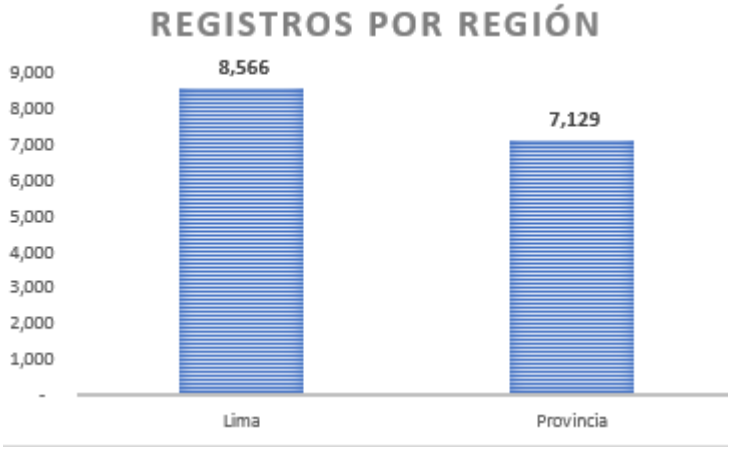


Figura 9: Registros por Región

Mientras que en la Tabla 9, se confirma que la moda los registros de iteraciones de clientes, es en la región Lima.

Tabla 9:

Análisis de la variable Región

Código	Región	Frecuencia	Moda
1	Lima	8,566	1
2	Provincia	7,129	

En la Figura 10 se puede apreciar los registros de las iteraciones de clientes realizadas en el periodo entre 2019 y 2020, en todos los locales del supermercado agrupado por la tarjeta que uso el cliente a la hora de la compra.

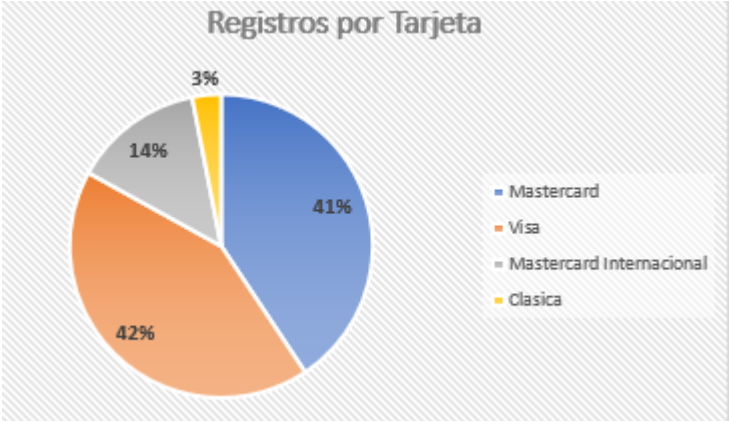


Figura 10: Registros por Tarjeta

Mientras que en la Tabla 10, se confirma que la moda del uso de tarjeta a la hora de hacer una compra por parte del cliente es Visa.

Tabla 10:

Análisis de la variable Tarjeta

Código	Tarjeta	Frecuencia	Moda
1	MasterCard	6,380	2
2	Visa	6,649	
3	MasterCard Internacional	2,183	
4	Clásica	483	

En la Figura 11 se puede apreciar los registros de las iteraciones de clientes realizadas en el periodo entre 2019 y 2020, en todos los locales del supermercado agrupado por la categoría del producto de la venta.

Figura 11: Registros por Categoría

Mientras que en la Tabla 11, se confirma que la moda de los registros de iteraciones de clientes, en base a productos por categoría es Non Food.

Tabla 11:

Análisis de la variable Categoría

Código	Categorías	Frecuencia	Moda
1	Abarrotes	545	4
2	Frescos	68	
3	Misceláneos	1	
4	Non Food	15,081	

En la Figura 12 se puede apreciar los registros de las iteraciones de clientes realizadas en el periodo entre 2019 y 2020, en todos los locales del supermercado sumarizados agrupado por el sexo del cliente que realizó la compra.

Figura 12: Registros por Sexo

Mientras que en la Tabla 12, se confirma que la moda de los registros de iteraciones de clientes, en base al sexo del cliente que realiza la compra es Masculino.

Tabla 12:

Análisis de la variable Sexo

Código	Sexo	Frecuencia	Moda
1	M	8,664	1
2	F	7,031	

En la Figura 13 se puede apreciar los registros de las iteraciones de clientes realizadas en el periodo entre 2019 y 2020, en todos los locales del supermercado agrupado por rangos de venta de producto por pedido, donde en las barras, el color azul indica el límite inferior y el color anaranjado el límite superior del rango, mientras que, la línea gris, indica la cantidad de registros por rango.

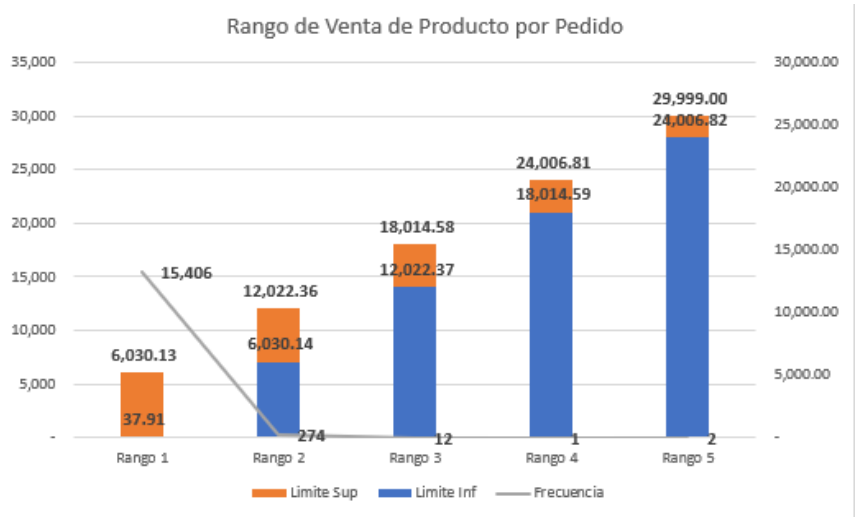


Figura 13: Registros por rangos de venta de Producto

Mientras que en la Tabla 13, se puede observar que la moda en los montos de venta de producto es de S/. 2,499. Además, que el rango de ventas que cuenta con mayor número de registros es el rango más pequeño, que entre S/.37.91 y S/. 6,030.13. También, se indica que entre todos los registros existe un promedio de S/. 2,799.29 y la mitad de estos tiene un monto de venta menor o igual a S/. 2,499. Por último, se tiene una desviación estándar pequeña con referencia a la media, por lo que se deduce que los valores están correctamente distribuidos.

Tabla 13:

Análisis de la variable Rango de venta de Producto

Código	Rango	Limite Inf	Limite Sup	Frecuencia	Media	Mediana	Moda	Desv. Estándar
1	Rango 1	37.91	6,030.13	15,406				
2	Rango 2	6,030.14	12,022.36	274				
3	Rango 3	12,022.37	18,014.58	12	2,799.29	2,499.00	2,499.00	1,143.26
4	Rango 4	18,014.59	24,006.81	1				
5	Rango 5	24,006.82	29,999.00	2				

En la Figura 14 se puede apreciar los registros de las iteraciones de clientes realizadas en el periodo entre 2019 y 2020, en todos los locales del supermercado agrupado por rangos de cantidad de ítems por pedido, donde en las barras, el color azul indica el límite inferior y el color anaranjado el límite superior del rango, mientras que, la línea gris, indica la cantidad de registros por rango.

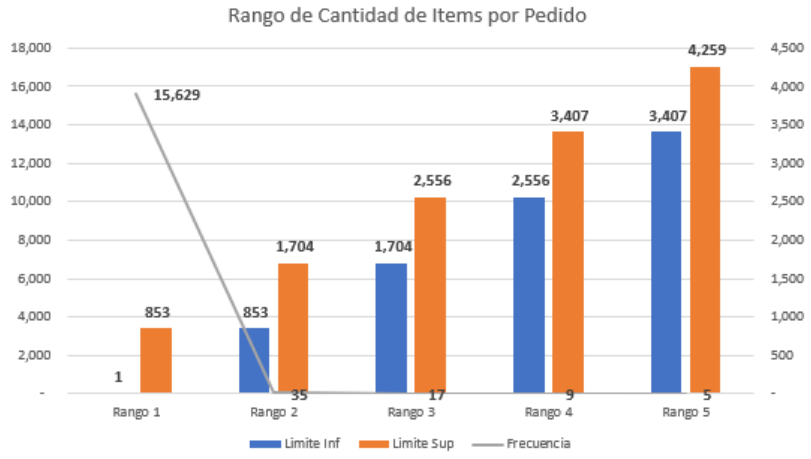


Figura 14: Registros por rangos ítems de Producto

Mientras que en la Tabla 14, se puede observar que la moda en la cantidad de ítems de producto es de 1. Además, que el rango de ítems que cuenta con mayor número de registros es el rango más pequeño, que entre 1 y 853. También, se indica que entre todos los registros existe un promedio de 16.68 y la mitad de estos tiene una cantidad de ítems menor o igual a 1. Por último, se tiene una desviación estándar pequeña con referencia a la media, por lo que se deduce que los valores están correctamente distribuidos.

Tabla 14:

Análisis de la variable Rango de ítems de Producto

Código	Rango	Limite Inf	Limite Sup	Frecuencia	Media	Mediana	Moda	Desv. Estándar
1	Rango 1	1	853	15,629				
2	Rango 2	853	1,704	35				
3	Rango 3	1,704	2,556	17	16.68	1	1	139.52
4	Rango 4	2,556	3,407	9				
5	Rango 5	3,407	4,259	5				

Contrastación de hipótesis específica 1

En la pregunta de investigación 1 se cuestiona lo siguiente: ¿Cómo se puede identificar los indicadores de correlación para el área comercial en supermercados?

Para responder esta pregunta, se recolectaron datos de cada una de las variables contenidas en la región de Lima, que es la muestra, para posteriormente hacer un análisis descriptivo, con el objetivo de evaluar la correlación entre todas las variables de la base de datos, con lo cual se pudo identificar que las que tenían un mayor nivel de correlación eran Recencia, Frecuencia y Monto de venta, por lo que posteriormente se realizó el análisis RFM gracias a las variables indicadas.

En la Figura 15, se puede apreciar lo indicado anteriormente, donde resaltan con un color más pronunciado los valores altos de correlación.

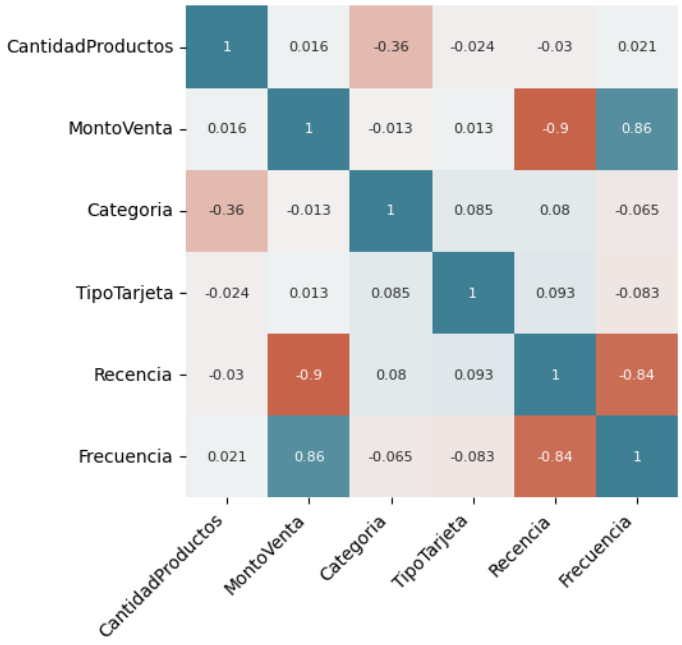


Figura 15: Coeficientes de correlación entre variables

Contrastación de hipótesis específica 2

En la pregunta de investigación 2 se cuestiona lo siguiente: ¿Cómo se puede identificar el perfil de lealtad en base a los indicadores de cliente para el área comercial en supermercados?

Para responder esta pregunta, se recolectaron los datos de cuartiles empleando el diagrama de cajas, por lo que se la Tabla 15 se detalla los tres cuartiles generados para cada indicador del análisis, que son la Recencia, la Frecuencia y el Monto de venta; de donde se observa la definición de los cuartiles con lo que se asignarán los puntajes respectivos a cada cliente.

Tabla 15:

Diagrama de Cajas

	Recencia	**Frecuencia**	**Monto de venta**
0.25	370	1	2,199.00
0.5	461	4	2,499.00
0.75	553	9	3,698.25

Luego, mediante la fórmula mostrada en el punto 1.1.3.2.1 se pudo determinar el puntaje de cada cuartil por indicador. En la Tabla 16, como ejemplo, se puede observar lo recolectado mediante el instrumento elaborado; en la Guía de observación se registró los puntajes obtenidos para cada cuartil.

Tabla 16:

Resultados de indicadores de cliente

Código Cliente	Recencia	Frecuencia	Monto de venta	Cuartil R	Cuartil F	Cuartil M
1	633	1	2,799.00	1	1	3
2	534	1	2,299.00	2	1	2
3	559	1	2,299.00	1	1	2
4	572	1	3,499.00	1	1	3
5	643	1	2,399.00	1	1	2
6	303	2	5,698.00	4	4	4
7	337	5	58,296.00	4	4	4
8	311	2	7,998.00	4	4	4
9	339	2	4,998.00	4	4	4
10	342	2	4,998.00	4	4	4

Mientras que, en la Tabla 17 se puede observar, como ejemplo, el score generado en base a sus cuartiles para diez clientes.

Tabla 17:

Scoring de clientes según RFM

Código Cliente	Recencia	Frecuencia	Monto de venta	Cuartil R	Cuartil F	Cuartil M	Score RFM
1	633	1	2,799.00	1	1	3	113
2	534	1	2,299.00	2	1	2	212
3	559	1	2,299.00	1	1	2	112
4	572	1	3,499.00	1	1	3	113
5	643	1	2,399.00	1	1	2	112
6	303	2	5,698.00	4	4	4	444
7	337	5	58,296.00	4	4	4	444
8	311	2	7,998.00	4	4	4	444
9	339	2	4,998.00	4	4	4	444
10	342	2	4,998.00	4	4	4	444

En base al análisis de las iteraciones de todos los clientes que se tienen como unidad de estudio, en la Tabla 18 se describe el resumen de seis perfiles según el score de RFM o de los cuartiles generados.

Los tres primeros perfiles son de un análisis positivo que se obtuvo; a los clientes que tuvieron un score RFM de 444, se los denominó Mejores Clientes, ya que tienen un alto consumo y frecuencia de compras, además de que su última fecha de compra es reciente; a los que tuvieron una puntuación en el cuartil de Frecuencia de 4, se los denominó como Clientes leales; los clientes que obtuvieron un puntaje de 4 en el cuartil de Monto Total, se los llamó como Clientes más gastadores.

Mientras que, los tres últimos perfiles refleja a los clientes con un bajo scoring; a los que tuvieron un score RFM de 244 se les denominó como Clientes casi muertos, ya que si bien tienen un alto consumo y una buena frecuencia de compras, estas no han sido muy recientes; luego se tienen a los Clientes perdidos que están basados en un score RFM de 144, que al igual que los Clientes casi muertos, tuvieron un buen consumo y frecuencia de compras, pero su última fecha de compra es demasiado lejana; por último, se tienen a los Clientes perdidos que son baratos con un score RFM de 111, lo que indica su bajo consumo y frecuencia de compras y su última fecha de compra es muy lejana.

Tabla 18:

Perfiles de clientes según modelo RFM

Perfil	Q de Clientes
Mejores Clientes	541
Clientes leales	648
Clientes más gastadores	2,303
Clientes casi muertos	216
Clientes perdidos	115
Clientes perdidos que son baratos	1220
Otros	6,387

Contrastación de hipótesis específica 3

En la pregunta de investigación 3 se cuestiona lo siguiente: ¿Cómo se puede medir la relación entre el modelo de clusterización y los indicadores de predicción para el área comercial en supermercados?

Para responder esta pregunta, se recolectaron los datos a través de las pruebas realizadas en el modelo de clusterización, por lo que, en la Tabla 19 se muestra los coeficientes de correlación de los tres indicadores del análisis RFM con respecto al porcentaje de predicción obtenido con el Coeficiente de Silueta, lo que permite medir el nivel de relación hay entre ellos.

Tabla 19:

Correlación con respecto a la predicción

	Recencia	**Frecuencia**	**Monto de venta**	**Silhouette**
Recencia	1	-0.84	-0.9	0.88
Frecuencia	-0.84	1	0.86	0.67
Monto de venta	-0.9	0.86	1	0.9
Silhouette	0.88	0.67	0.9	1

Mientras que, la Figura 16 es la representación gráfica de la tabla anterior, donde, se puede decir que existe una relación de nivel alto entre las variables empeladas en el modelo con respecto al porcentaje de predicción del mismo.

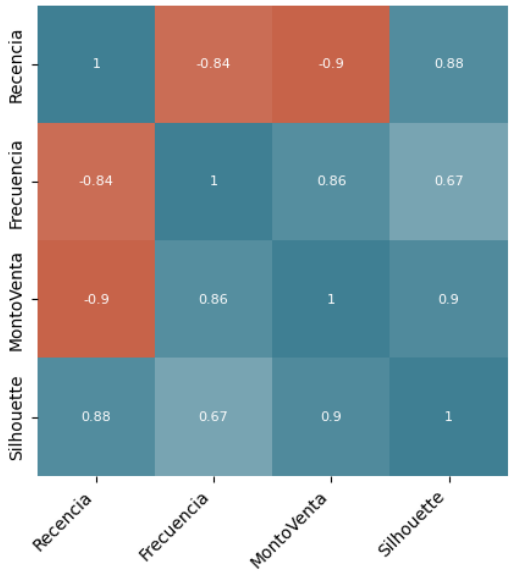


Figura 16: Correlación con respecto a la predicción

Para definir cuál era el número óptimo de clústers, se utilizó el coeficiente de Silueta, calculando la distancia media dentro de un conglomerado de datos y la distancia media con respecto al conglomerado de datos más cercado; esto ayudó a obtener un puntaje de predicción según el número de segmentaciones posibles basados en la combinación de indicadores con mayor porcentaje de predicción que se recolectó en la guía de observación, los cuales se muestran en la Tabla 20.

Tabla 20:

Resultados de indicadores de predicción

VARIABLES			Número de clústers															
Variable 1	Variable 2	Variable 3	2		3		4		5		6		7		8		9	
			%PP	#CP	%PP	#CP	%PP	#CP	%PP	#CP	%PP	#CP	%PP	#CP	%PP	#CP	%PP	#CP
Recencia	-	-	0.63	2,632	0.603	2,206	0.587	1,780	0.579	1,354	0.562	928	0.558	745	0.556	679	0.553	547
Frecuencia	-	-	0.928	2,956	0.882	2,530	0.886	2,104	0.863	1,678	0.878	1,252	0.905	826	0.908	400	0.945	533
MontoVenta	-	-	0.73	3,280	0.624	2,854	0.579	2,428	0.552	2,002	0.56	1,576	0.566	1,150	0.564	724	0.562	519
Recencia	Frecuencia	-	0.628	3,604	0.598	3,178	0.580	2,752	0.568	2,326	0.551	1,900	0.54	1,474	0.537	1,048	0.529	505
Recencia	MontoVenta	-	0.716	3,928	0.592	3,502	0.533	3,076	0.482	2,650	0.468	2,224	0.471	1,798	0.472	1,372	0.43	491
Frecuencia	MontoVenta	-	0.73	4,252	0.624	3,826	0.579	3,400	0.551	2,974	0.56	2,548	0.565	2,122	0.565	1,696	0.562	477
Recencia	Frecuencia	MontoVenta	0.716	4,052	0.592	3,626	0.532	3,200	0.482	2,774	0.468	2,348	0.471	1,922	0.471	1,496	0.43	463

De tabla anterior se puede observar que la combinación de variables con mayor porcentaje de predicción (%PP) son la de Frecuencia y Monto de venta con dos clústers; por lo que, en la Tabla 21, según esta combinación, se muestra los resultados obtenidos en base hasta 9 clústers posibles, donde el mayor valor está asociado a solo 2 clústers con un 73% de predicción y un número de clientes potenciales (#CP) de 4,252; el cual, será el número final de agrupaciones.

Tabla 21:

Puntajes de predicción de clústers

Número de clústers	Puntaje de predicción
2	0.730
3	0.624
4	0.579
5	0.551
6	0.560
7	0.565
8	0.565
9	0.562

Mientras que, en la Figura 17 están representados los dos clústers finales según el puntaje obtenido anteriormente, donde se puede observar que en el clúster de color amarillo, los clientes no están tan dispersos, por lo que tienen una frecuencia de compra entre 7.5 a 10 días y un consumo total entre S/.5,600 y S/.6,400; mientras que en el clúster de color azul, el consumo total entre S/.5,600 y S/.5,800 se da en la frecuencia de compra de los clientes entre 10 y 15 días, posterior a ello el consumo disminuye.

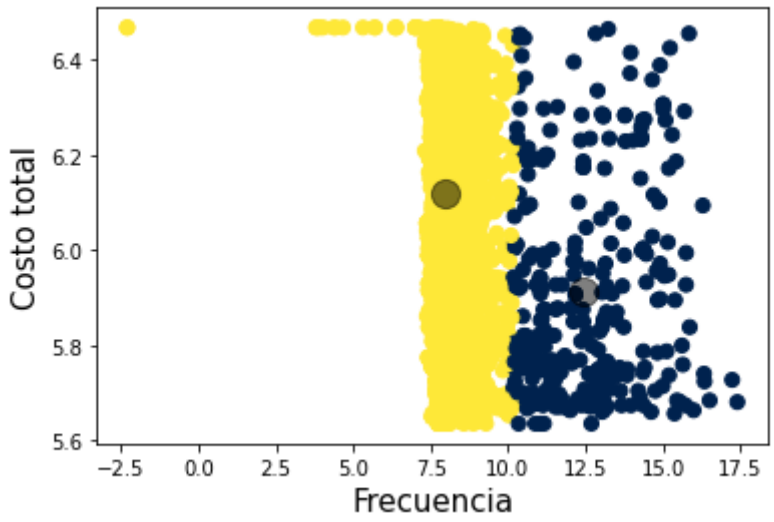


Figura 17: Distribución de clústers finales

Por lo que en el clúster amarrillo se tienen a los clientes categorizados en el análisis RFM como mejores clientes, clientes leales y clientes más gastadores, mientras que en el clúster azul los restantes. Además, en la Tabla 22 se muestra el número de clientes potenciales para cada clúster final.

Tabla 22:

Número de clientes potenciales

	Amarillo	**Azul**
	Clúster 1	**Clúster 0**
# de clientes potenciales	3,740	512

Por último, gracias al test de normalidad de Kolmogorov-Smirnov se logró obtener una significancia de 0.74, superando el valor mínimo esperado de 0.05 que señala que se acepta la hipótesis nula; por lo que las variables tienen una distribución normal del conjunto de datos que se registraron en el instrumento de observación que se empleó para los indicadores de predicción.

En base a lo mencionado, en la Figura 18 se puede observar la distribución del porcentaje de predicción en la guía de observación de los indicadores de predicción, donde resalta que, la mayor cantidad de datos registrados figuran entre 0.5 y 0.6 de porcentaje de predicción, mientras que, los que tienen una menor distribución de datos se aproximan a 0.9 considerando el análisis solo con una variable.

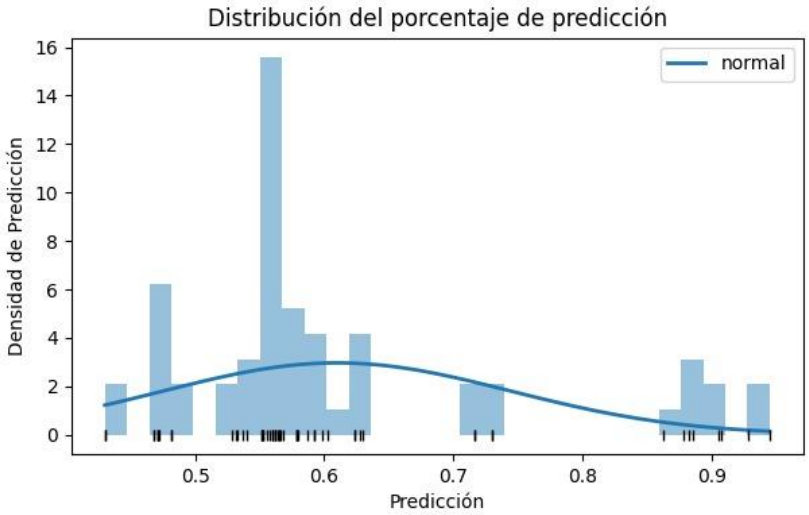


Figura 18: Distribución del porcentaje de predicción

Mientras que, en la Figura 19 se puede apreciar el gráfico de cuantiles teóricos (Q-Q), los cuales comparan los cuantiles del conjunto de datos en evaluación con los de una distribución normal y como se puede ver en la figura indicada, los datos entre el cuantil -1 y 1 están más alineados a la línea roja que muestra la distribución normal de las variables.

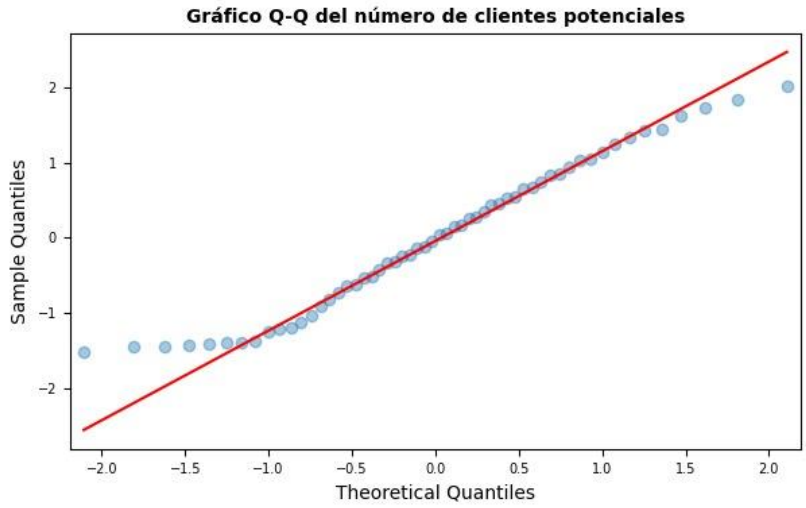


Figura 19: Gráfico Q-Q del número de clientes potenciales

CAPÍTULO IV. DISCUSIÓN Y CONCLUSIONES

4.1 Discusión de resultados

De los resultados obtenidos, se identificó que el modelo de clusterización tiene un alto nivel de relación con respecto a los indicadores de correlación, ya que la correlación entre las variables de Recencia y Frecuencia es de -0.84; entre Recencia y Monto de venta es de -0.9 y Frecuencia con Monto de venta es de 0.86. Además, los indicadores de clientes permitieron identificar el perfil de lealtad en base al puntaje por cada cuartil de las variables ya mencionadas; y con respecto a la relación del modelo de clusterización con los indicadores de predicción se puede afirmar que es alto ya que se obtuvo un 73% de porcentaje de predicción y 4,252 clientes potenciales evaluando con dos clústers.

Luego, cabe señalar que se obtuvo como resultado que con dos clústers se representó el mayor porcentaje de predicción con un 73% a la hora de generar el modelo de clusterización, pero se tuvo una diferencia considerable con el 99.5% que llegó a tener en la investigación "*Segmentación de clientes de un casino utilizando el algoritmo partición alrededor de medoides (PAM) con datos mixtos*" con tres clústers. Esto no es algo con lo que se pueda inferir que hay algo malo en alguno de los trabajos de investigación, sino que, esta diferencia depende de muchos factores, tales como, la generación de los datasets con que se han trabajado los modelos, la antigüedad de los datos con los que se trabajó y otros más (Elguera, 2018).

También, el uso de guías de observación para el registro de datos de las distintas evaluaciones, tal y como se hace en esta investigación, no es muy común en las investigaciones analizadas. Así lo muestra la tesis "*Aplicación de minería de datos para pronosticar el riesgo de morosidad de los estudiantes de la Universidad Autónoma del Perú*", donde se emplea como instrumento para la recolección de datos una encuesta a los estudiantes para saber datos que estén relacionados a su nivel académico y situación financiera; para luego de ello, analizar la distribución de los datos registrados a través del método de Shapiro Wilk ya que a diferencia de esta investigación, el número de datos era menor a cincuenta, por lo que no usaron el mismo método de análisis de la normalidad de los datos (Córdova & Torres, 2018).

Luego, en comparación con Milla (2017), quién empleó arboles de decisión para llegar a evaluar la relación de su modelo con la generación de compañas comerciales más eficientes, en esta investigación se optó por utilizar el algoritmo k-means, el cual me permitió evaluar los indicadores de porcentaje de predicción, número de clústers y número de clientes potenciales. Esto permitió, a diferencia de Milla (2017), identificar que variables eran las más adecuadas a emplear en el modelo, ya que previamente se midió la relación con respecto a los indicadores de correlación.

Por último, Del Pino & Cortez (2017) utilizaron técnicas de minería de datos en otro rubro a nivel organizacional, como son los procesos petitorios, pero a similitud de esta investigación, donde se segmenta la información para encontrar clientes potenciales; la aplicación de ellos estuvo enfocada en encontrar patrones de afinidad, mediante la evaluación de indicadores de clientes, como son la Recencia, Frecuencia y Monto de venta, para que luego, empleando el algoritmo k-means, pueda ir registrando los datos arrojados en sus pruebas en una ficha de datos y medir cuál es el testeo más óptimo para predecir el volumen de procesos petitorios y así mejorar la calidad en los servicios brindados.

En base al registro de datos empleado en la investigación de Del Pino & Cortez (2017), tuvo una significancia de 0.62, valor que le permitió señalar que la hipótesis nula era aceptada, por lo cual, al igual que en esta investigación, la distribución de sus datos recolectados era normal.

4.2 Conclusiones y limitaciones

4.2.1 Conclusiones

En este trabajo se planteó un objetivo de investigación en base a la implementación de un modelo de clusterización en la segmentación de perfiles de clientes para el área comercial en supermercados; por lo que se propuso la hipótesis de que este modelo se relaciona correctamente con el perfil de clientes en base a los indicadores de correlación, cliente y predicción. Después de los resultados obtenidos, queda claro que esta hipótesis es correcta, ya que gracias al modelo desarrollado se puede decir lo siguiente:

Las variables de Recencia, Frecuencia y Monto de venta, obtenidas en la evaluación de los indicadores de correlación permiten al área comercial identificar de manera más óptima comportamientos de clientes.

Además, con respecto a los indicadores de clientes, en base a la evaluación de los puntajes de cuartiles, se puede identificar que los “Mejores clientes”, “Clientes leales” y “Clientes más gastadores” son los que mayor tendencia tienen a ser re compradores.

Por último, en base a la evaluación de los indicadores de predicción, se puede deducir que el modelo empleado se puede mejorar a través de mayor cantidad de variables numéricas se tengan disponibles.

4.2.2 Limitaciones

Las principales tres limitaciones de la presente investigación están basadas en el aspecto técnico, al no poder contar con una institución específica para implementar el desarrollo del modelo de clusterización diseñado, ni utilizar como fuente de datos la información de la misma, se tuvo que utilizar datos abiertos encontrados en la nube, pero asegurando que esta esté orientada al tema del presente trabajo.

En el desarrollo de la primera etapa, se contó solo con seis variables numéricas lo cual limitó el enriquecimiento de la data para la construcción de un modelo predictivo más eficiente.

En la segunda etapa, al querer evaluar el cuartil de Frecuencia se encontró la limitación de solo tener dos años de historia de iteraciones, ocasionando que el indicador no tenga una medición más óptima.

En la última etapa, respecto a los indicadores de predicción, se tuvo como limitante los dos puntos antes mencionados, ya que estos influyen directamente en el porcentaje de predicción del cliente potencial.

4.3 Implicancias y futuras investigaciones

4.3.1 Implicancias

En lo que respecta a las implicancias prácticas, se sabe que, el objetivo de evaluar la relación de la implementación de un modelo de clusterización para la segmentación del perfil del cliente en el área comercial de supermercados se basa en entender cómo se puede involucrar técnicas de minería de datos en el proceso de encontrar clientes potenciales y así mejorarlo, ya que, según Flores (2016), la generación de clústers permite segmentar a los clientes que se tienen en base a las similitudes y diferencias que puedan tener entre ellos.

Además, en base a los indicadores de correlación, se puede asegurar para futuras investigaciones que también tengan el mismo objetivo de esta, que deben medir las tres variables identificadas como las más óptimas, que son la Recencia, Frecuencia y Monto de venta.

También, con referencia a los indicadores de cliente, para futuras investigaciones sobre el mismo tema de investigación, deben tener en cuenta que necesitan tener un historial de iteraciones de clientes mayor a dos años para evitar la limitación explicada en el punto 4.2.2.

Por último, en lo que respecta a los indicadores de predicción, las futuras investigaciones deben tener claro que el modelo se puede mejorar en base a que se tenga mayor cantidad de data y variables numéricas a evaluar.

4.3.2 Recomendaciones para futuras investigaciones

Con relación a futuras investigaciones, aún quedan algunos temas pendientes de resolver, por ejemplo, a nivel de organización, no está definido si el involucrar el modelo de clusterización en la segmentación de clientes incluye un aumento o disminución de presupuesto, ya que tal vez pueda ahorrarse en recursos humanos, pero existirá un costo a nivel de recursos tecnológicos. Además, a un nivel técnico, se puede buscar el utilizar un algoritmo que no te exija colocar manualmente el número clústers a probar para determinar su porcentaje de predicción, sino más bien, que lo calcule de manera automática y te devuelva como resultado el número de clústers adecuado.

En futuros estudios, se podría mejorar los perfiles de los clientes definidos en el análisis RFM para así ir evaluando en base a los clústers que se tendrán como resultado, así como también, el contar con un mayor número de variables en la base de datos, como, la hora de compra, la edad del cliente, estado civil del cliente, si el cliente compra solo o acompañado, etc.

Por ello, el desarrollo realizado en esta investigación demuestra que es posible involucrar las iteraciones de los clientes en el proceso de segmentación de clientes.

REFERENCIAS

Roberto Hernández Sampieri. Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta (2018). *Universidad de Celaya*.

José Supo. Metodología de la investigación científica: Para las Ciencias de Salud y las Ciencias Sociales – 3era Edición (2020). *Seminarios de Investigación Científica*.

Heber Héctor Milla Caballero. Propuesta de metodología para mejorar la efectividad de las campañas comerciales de un ISP utilizando Data Mining (2017). *Universidad Nacional de Ingeniería*.

Natali Flores Lafosse. Extracción de patrones semánticamente distintos a partir de los datos almacenados en la plataforma Paideia (2016). *Pontificia Universidad Católica del Perú*.

Karina Atalaya Tello, Nancy Flores Aedo, Ángela Flores Alvarado. Propuesta de analytics a los patrones de comportamiento en el proceso de clasificación socioeconómica en el MIDIS (2019). *Universidad Peruana de Ciencias Aplicadas*.

Rubén Darío Gómez Ríos. Modelo predictivo de gestión administrativa y deserción estudiantil en programa pregrado adulto trabajador de universidad privada de Lima Metropolitana del año 2017 (2018). *Universidad Privada del Norte*.

Jean Pierre Prati Oblitas, Jushua Rai Ganhi Baldoceda Puentes. Modelo de clasificación de clientes con telefonía móvil y uso de canal de autogestión USSD en una empresa de telecomunicaciones (2017). *Universidad Nacional de Ingeniería*.

Aldo Richard Meza Rodríguez. Predicción de fuga de clientes en una empresa de telefonía utilizando el algoritmo Adaboost desbalanceado y la regresión logística asimétrica (2018). *Universidad Nacional Agraria La Molina*.

Víctor Andrés Edgard Cáceres Chian. Predicción de precios de acciones de bolsa de valores utilizando Support Vector Regression (2018). *Universidad de Lima*.

Eiriku Yamao. Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo de la escuela profesional de Ingeniería de Computación y Sistemas, Universidad de San Martín de Porres (2018). *Universidad de San Martín de Porres*.

Zoraida Mamani Rodríguez, Luz del Pino Rodríguez, Augusto Cortez Vásquez. Minería de datos distribuida usando clustering k-means en la predictibilidad del proceso petitorio en una organización pública (2017). *Industrial Data – Universidad Mayor de San Marcos*.

Carlos Eduardo Marulanda Echeverry, Marcelo López Trujillo, María Helena Mejía Salazar. Minería de datos en gestión del conocimiento de pymes de Colombia (2017). *Revista Virtual Universidad Católica del Norte*.

Delimiro Visbal Cadavid, Adel Mendoza Mendoza, Sonia Jacqueline Orejuela Pedraza. Predicción de la eficiencia de las instituciones de educación superior colombianas con análisis envolvente de datos y minería de datos (2017). *Pensamiento & Gestión*.

Rogelio Morejón Rivera, Félix A. Cámara, Dany E. Jiménez, Sandra H. Díaz. SISDAM: Aplicación web para el procesamiento de datos según un diseño aumentado modificado (2016). *Cultivos Tropicales*.

Agustí Cerrillo-Martínez. Datos masivos y datos abiertos para una gobernanza inteligente (2018). *El profesional de la información*. ISSN 1699-2407, Vol. 27, N. 5 (Ejemplar dedicado a: Información política y redes sociales (I)), 1128-1135.

Diana Arce, Fernando Lima, Marcos Orellana, John Ortega, Chester Sellers, Patricia Ortega. Descubriendo patrones de comportamiento entre contaminantes del aire: Un enfoque de minería de datos (2018). *Enfoque UTE*.

Marcos Orellana, Priscilla Cedillo. Detección de valores atípicos con técnicas de minería de datos y métodos estadísticos (2020). *Enfoque UTE*.

doi: https://doi.org/10.29019/enfoque.v11n1.584

E.M. Ruiz Lobaina, C.P. Romero Suárez. Resultados obtenidos en un proceso de minería de datos aplicado a una base de datos que contiene información bibliográfica referida a cuatro segmentos de la ciencia (2018). *JISTEM: Journal of Information System and Technology Management*.

Fernando Rogelio Simonato. La innovación en el área comercial a través de la gestión de las experiencias (2018). *Ciencias Administrativas - Argentina*.

Merlinda Clarke-Bloomfield, Yoanis Cisneros-Arias, Yurianela Paneca-González. Gestión Comercial: diagnóstico del atractivo y rentabilidad del punto de ventas (2018). *Ciencias Holguín*.

Fernando Martínez-Abad, Juan Pablo Hernández-Ramos. Técnicas de minería de datos con software libre para la detección de factores asociados al rendimiento (2018). *REXE. Revista de Estudios y Experiencias en Educación*.

Carlos Arcila Calderón, Luz Marina Alonso, Antonio García Jiménez. Enfoques big data para la comunicación en salud: análisis de redes y análisis de sentimientos a gran escala (2018). *Salud Uninorte*.

Elisa García Morales. Seis contribuciones de la gobernanza de la información a la transparencia y la lucha contra la corrupción (2016). *Anuario ThinkEPI*.

Rienties Bart, Cross Simon, Marsh, Vicky and Ullmann, Thomas. Making sense of learner and learning Big Data: reviewing 5 years of Data Wrangling at the Open University UK (2017). *Open Learning: The Journal of Open, Distance and e-learning*, 32(3) pp. 279–293.

Fernando Díaz. Importancia de los aspectos no tecnológicos de las iniciativas de Big Data (2017). *Economía industrial*, ISSN 0422-2784, N.º 405.

Alenster Córdova & Karen Torres. Aplicación de minería de datos para pronosticar el riesgo de morosidad de los estudiantes de la Universidad Autónoma del Perú (2018). *Repositorio de la Universidad Autónoma del Perú*.

Zoraida Mamani. Aplicación de la minería de datos distribuida usando algoritmo de clustering k-means para mejorar la calidad de servicios de las organizaciones modernas (2015). *Repositorio de la Universidad Mayor de San Maros*.

Rodolfo Pacco. Análisis predictivo basado en redes neuronales no supervisadas aplicando algoritmo de k-means y CRISP-DM para pronóstico de riesgo de morosidad de los alumnos en la Universidad Peruana Unión (2015). *Repositorio de la Universidad Peruana Unión*.

Rhony Elguera. Segmentación de clientes de un casino utilizando el algoritmo Partición alrededor de Medoides (PAM) con datos mixtos (2018). *Repositorio de la Universidad Nacional Agraria*.

Juan Carrión & Marvin Espinoza & Milagros Lártiga & Lisha Yangali. Planeamiento estratégico de la empresa Supermercados Peruanos SPSA (2018). *Repositorio de la Pontifica Universidad Católica del Perú*.

Ximena Videla Cabello. Medidas de posición y gráfico de caja y bigote: una propuesta didáctica (2017). *Repositorio de la Pontifica Universidad Católica de Valparaíso*.

Elizabeth León Guzmán. Métricas para la validación de Clustering (2019). *Repositorio de la Universidad Nacional de Colombia*.

Rafael Aleixandre-Benavent & Antonia Ferrer Sapena & Fernanda Peset. Compartir los recursos útiles para la investigación: datos abiertos – open data (2019). *Repositorio de la Universidad Politécnica de Valencia, España*.

ANEXOS

ANEXO 1

En la siguiente tabla se muestra la matriz de consistencia, donde se resume todo lo desarrollado en el presente trabajo.

Tabla 23:

Matriz de consistencia

<table>
<tr><td colspan="6">Tema de investigación: Implementación de un modelo de clusterización para la segmentación del perfil del cliente en el área comercial de supermercados</td></tr>
<tr><td>Problema General</td><td>Objetivo General</td><td>Hipótesis General</td><td colspan="3">Variable Independiente</td></tr>
<tr><td rowspan="3">¿De qué manera un modelo de clusterización segmenta el perfil del cliente para el área comercial en supermercados?</td><td rowspan="3">Implementar un modelo de clusterización en la segmentación del perfil del cliente para el área comercial en supermercados.</td><td rowspan="3">El modelo de clusterización segmenta correctamente el perfil del cliente en base a los indicadores de correlación, cliente y predicción.</td><td colspan="3">Modelo de clusterización (Flores, 2016)</td></tr>
<tr><td colspan="3">Variable Dependiente</td></tr>
<tr><td colspan="3">Perfil de clientes (Simonato, 2018)</td></tr>
<tr><td>Problemas Específicos</td><td>Objetivos Específicos</td><td>Hipótesis Específicas</td><td>Dimensiones</td><td>Indicadores</td><td>Unidad de medida</td></tr>
<tr><td rowspan="2">¿Cómo se puede identificar los indicadores de correlación para el área comercial en supermercados?</td><td rowspan="2">Identificar los indicadores de correlación para el área comercial en supermercados.</td><td rowspan="2">Los indicadores de correlación están entre -1 y 1.</td><td rowspan="2">Indicadores de correlación</td><td>Correlación entre variables</td><td rowspan="2">Nivel de correlación</td></tr>
<tr><td>Correlación con respecto a la predicción</td></tr>
<tr><td rowspan="3">¿Cómo se puede identificar el perfil de lealtad mediante los indicadores de cliente para el área comercial en supemercados?</td><td rowspan="3">Identificar el perfil de lealtad en base a los indicadores de cliente para el área comercial en supermercados.</td><td rowspan="3">El perfil de lealtad se basa en los puntajes entre 1 y 4 de los indicadores de cliente.</td><td rowspan="3">Indicadores de cliente</td><td>Cuartil de Recencia</td><td>Recencia</td></tr>
<tr><td>Cuartil de Frecuencia</td><td>Frecuencia</td></tr>
<tr><td>Cuartil de Monto de venta</td><td>Monto de venta</td></tr>
<tr><td rowspan="3">¿Cómo se puede medir la relación entre el modelo de clusterización y los indicadores de predicción para el área comercial en supemercados?</td><td rowspan="3">Evaluar la relación entre el modelo de clusterización y los indicadores de predicción para el área comercial en supermercados.</td><td rowspan="3">El nivel de relación entre los indicadores de predicción y el modelo de clusterización supera el 70%.</td><td rowspan="3">Indicadores de predicción</td><td>Porcentaje de predicción</td><td>Porcentaje</td></tr>
<tr><td>Número de clústers</td><td>Número</td></tr>
<tr><td>Número de clientes potenciales</td><td>Número</td></tr>
<tr><td>Tipo y diseño</td><td>Población y muestra</td><td>Técnicas e instrumentos</td><td colspan="3">Estadística a utilizar</td></tr>
<tr><td rowspan="2">Tipo: Aplicada

Nivel: Aplicativo

Enfoque: Cuantitativo

Diseño: Pre-experimental</td><td rowspan="2">Población: 14,268 iteraciones de clientes en un supermercado de Perú durante el periodo 2019 – 2020

Muestra: 7,726 iteraciones de clientes en la región Lima</td><td rowspan="2">Técnica: Observación

Instrumento:
Guía de observación</td><td colspan="3">Descriptiva
Se usan tablas y figuras, datos emitidos por el instrumento, lo cual ayudará a fijar de manera visual y estructurada la comprensión sencilla de todos los datos numéricos.
Inferencial</td></tr>
<tr><td colspan="3">Para el análisis inferencial, se comprobó la normalidad de los datos obtenidos mediante la prueba Test de Kolmogorov-Smirnov.</td></tr>
</table>

ANEXO 2

En la siguiente figura se muestra el instrumento de recolección de datos para la dimensión de Indicadores de correlación.

INSTRUMENTO DE RECOLECCIÓN DE DATOS: Nivel de relación entre variables

Guía de Observación			
Investigador	Gustavo Alfonso Arrelucea Zapata	**Tipo de prueba**	Post
Descripción	Obtener el nivel de relación entre las variables de la base de datos.		
Variable		**Fórmula**	
PERFIL DE CLIENTE		$r = \frac{\alpha xy}{\alpha_x * \alpha_y}$	
Dimensión			
INDICADORES DE CORRELACIÓN			
Indicador	**Medida**		
Correlación entre variables y con respecto a la predicción	Nivel de correlación	Donde: r = Es un coeficiente Si r ≈ 1, existe una correlación directa fuerte Si r ≈ -1, existe una correlación indirecta fuerte Si r = 1 o r = -1, hay una correlación funcional Si r ≈ 0, no existe una correlación lineal	

VARIABLE	VALORES								
	Correlación entre variables								Correlación con respecto a la predicción
	Cantidad de productos	Categoría de producto	Tipo de tarjeta	Sexo de cliente	Región	Recencia	Frecuencia	Monto de venta	Silhouette
Cantidad de productos									
Categoría de producto									
Tipo de tarjeta									
Sexo de cliente									
Región									
Recencia									
Frecuencia									
Monto de venta									
Silhouette									

ANEXO 3

En la siguiente figura se muestra el instrumento de recolección de datos para la dimensión de Indicadores de cliente.

INSTRUMENTO DE RECOLECCIÓN DE DATOS: Nivel de relación de indicadores de cliente

<table>
<tr><th colspan="4">Guía de Observación</th></tr>
<tr><td>Investigador</td><td>Gustavo Alfonso Arrelucea Zapata</td><td>Tipo de prueba</td><td>Post</td></tr>
<tr><td>Descripción</td><td colspan="3">Obtener el nivel de relación de los indicadores de cliente</td></tr>
<tr><td colspan="2">Variable</td><td colspan="2">Fórmula</td></tr>
<tr><td colspan="2">PERFIL DE CLIENTE</td><td colspan="2" rowspan="4">$f_i = \frac{i-1}{n-1}$
Donde:
i = Es el índice del valor
n = Es el número de valores</td></tr>
<tr><td colspan="2">Dimensión</td></tr>
<tr><td colspan="2">INDICADORES DE CLIENTE</td></tr>
<tr><td>Indicador</td><td>Medida</td></tr>
<tr><td>• Cuartil de Recencia
• Cuartil de Frecuencia
• Cuartil de Monto de Venta</td><td>• Recencia
• Frecuencia
• Monto de venta</td><td colspan="2"></td></tr>
</table>

Código de cliente	Cuartil de Recencia	Cuartil de Frecuencia	Cuartil de Monto de venta

ANEXO 4

En la siguiente figura se muestra el instrumento de recolección de datos para la dimensión de Indicadores de predicción.

INSTRUMENTO DE RECOLECCIÓN DE DATOS: Indicadores de predicción

Guía de Observación			
Investigador	Gustavo Alfonso Arrelucea Zapata	**Tipo de prueba**	Post
Descripción	Obtener los indicadores de la predicción		
Variable		**Fórmula**	
PERFIL DE CLIENTE		$s(i) = \frac{b - a}{\max(a, b)}$ Donde: **a** = Es el promedio de las disimilitudes (o distancias) de la observación i con las demás observaciones del clúster al que pertenece i. **b** = Es la distancia mínima a otro clúster que no es el mismo en el que está la observación i. Ese clúster es la segunda mejor opción para i y se lo denomina vecindad de i. El coeficiente de Silueta [s(i)] es un valor comprendido entre -1 y 1.	
Dimensión			
INDICADORES DE PREDICCIÒN			
Indicador	**Medida**		
• Número de clústers • **%PP**: Porcentaje de predicción • **#CP**: Número de clientes potenciales	• Número • Porcentaje • Número		

VARIABLES			Número de clústers															
Variable 1	Variable 2	Variable 3	2		3		4		5		6		7		8		9	
			%PP	#CP	%PP	#CP	%PP	#CP	%PP	#CP	%PP	#CP	%PP	#CP	%PP	#CP	%PP	#CP
Recencia	-	-																
Frecuencia	-	-																
MontoVenta	-	-																
Recencia	Frecuencia	-																
Recencia	MontoVenta	-																
Frecuencia	MontoVenta	-																
Recencia	Frecuencia	MontoVenta																

ANEXO 5

A continuación, se adjuntan los certificados de validez de instrumentos realizada por expertos.

CERTIFICADO DE VALIDEZ DEL INSTRUMENTO

Título de la investigación: Implementación de un modelo de clusterización en la segmentación de perfil de clientes para el área comercial en supermercados
Autor: Arrelucea Zapata, Gustavo Alfonso **Instrumento**: Guía de observación

DATOS DEL EXPERTO
Apellidos y Nombres : _OVALLE PAULINO CHRISTIAN_____ **DNI:** 40234321 **CIP**: 213553 **Especialidad del validador**: ING SISTEMAS
Grado Académico : Magister (_) Doctor (X)

ASPECTOS DE VALIDACIÓN

Variable	Perfil de cliente							
Dimensión	Indicador	Unidad Medida	CLARIDAD		PERTINENCIA		RELEVANCIAS		SUGERENCIAS
			SI	NO	SI	NO	SI	NO	
Indicadores de Correlación	Correlación entre variables y con respecto a la predicción	Nivel de correlación	X		X		X		
Indicadores de Cliente	• Cuartil de Recencia • Cuartil de Frecuencia • Cuartil de Monto de venta	• Recencia • Frecuencia • Monto de venta	X		X		X		
Indicadores de Predicción	• # de clústers • % de predicción • # de clientes potenciales	• Número • Porcentaje • Número	X		X		X		

Opinión de aplicabilidad	Aplicable (X)	Aplicable después de corregir (_)	No aplicable ( )

Observaciones:

Firma del Validador

Fecha: 27-09 - 2022

CERTIFICADO DE VALIDEZ DEL INSTRUMENTO

Título de la investigación: Implementación de un modelo de clusterización en la segmentación de perfil de clientes para el área comercial en supermercados
Autor: Arrelucea Zapata, Gustavo Alfonso **Instrumento**: Guía de observación

DATOS DEL EXPERTO
Apellidos y Nombres : Ramos Gonzales Carlos **DNI:** 25771858 **CIP**: 228967 **Especialidad del validador**: Ing. Electrónico
Grado Académico : Magister (X) Doctor ( )

ASPECTOS DE VALIDACIÓN

Variable	Perfil de cliente								
Dimensión	**Indicador**	**Unidad Medida**	**CLARIDAD**		**PERTINENCIA**		**RELEVANCIAS**		**SUGERENCIAS**
			SI	**NO**	**SI**	**NO**	**SI**	**NO**	
Indicadores de Correlación	Correlación entre variables y con respecto a la predicción	Nivel de correlación	X		X		X		
Indicadores de Cliente	• Cuartil de Recencia • Cuartil de Frecuencia • Cuartil de Monto de venta	• Recencia • Frecuencia • Monto de venta	X		X		X		
Indicadores de Predicción	• # de clústers • % de predicción • # de clientes potenciales	• Número • Porcentaje • Número	X		X		X		

Opinión de aplicabilidad	Aplicable ( )	Aplicable después de corregir ( )	No aplicable ( )

Observaciones:

Firma del Validador

Fecha: 27 – 09 - 2022

CERTIFICADO DE VALIDEZ DEL INSTRUMENTO

Título de la investigación: Implementación de un modelo de clusterización en la segmentación de perfil de clientes para el área comercial en supermercados
Autor: Arrelucea Zapata, Gustavo Alfonso **Instrumento**: Guía de observación

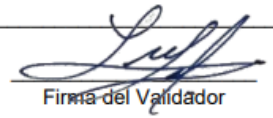
DATOS DEL EXPERTO
Apellidos y Nombres : Boza Ccoyllar Brando **DNI:** 71468243 **CIP**: 243479 **Especialidad del validador**: Ingeniero Mecatrónico
Grado Académico : Magister (X) Doctor ( )

ASPECTOS DE VALIDACIÓN

Variable	Perfil de cliente							
Dimensión	**Indicador**	**Unidad Medida**	**CLARIDAD**		**PERTINENCIA**		**RELEVANCIAS**		**SUGERENCIAS**
			SI	**NO**	**SI**	**NO**	**SI**	**NO**	
Indicadores de Correlación	Correlación entre variables y con respecto a la predicción	Nivel de correlación	X		X		X		
Indicadores de Cliente	• Cuartil de Recencia • Cuartil de Frecuencia • Cuartil de Monto de venta	• Recencia • Frecuencia • Monto de venta	X		X		X		
Indicadores de Predicción	• # de clústers • % de predicción • # de clientes potenciales	• Número • Porcentaje • Número	X		X		X		

Opinión de aplicabilidad	Aplicable (X)	Aplicable después de corregir ( )	No aplicable ( )

Observaciones:



Firma del Validador

Fecha:27/09/2022