SEGMENTACION DE CLIENTES DE UNA EMPRESA COMERCIALIZADORA DE PRODUCTOS DE CONSUMO MASIVO EN LA CIUDAD DE POPAYÁN SOPORTADO EN MACHINE LEARNING Y ANALISIS RFM (Recency, Frecuency y Money)

FABIAN ANTONIO PALACIOS ABADÍA
NELSON ANDRES PASTOR PATIÑO

FUNDACIÓN UNIVERSITARIA DE POPAYÁN
Programa de ingeniería de sistemas
Línea(s) de investigación: sistemas telemáticos inteligentes
Popayán, abril de 2020

SEGMENTACION DE CLIENTES DE UNA EMPRESA COMERCIALIZADORA DE PRODUCTOS DE CONSUMO MASIVO EN LA CIUDAD DE POPAYÁN SOPORTADO EN MACHINE LEARNING Y ANALISIS RFM (Recency, Frecuency y Money)

FABIAN ANTONIO PALACIOS ABADÍA
NELSON ANDRES PASTOR PATIÑO

Monografía de trabajo de grado para optar al título de:

Ingeniero de sistemas

Director:
PhD. Armando Ordoñez

FUNDACIÓN UNIVERSITARIA DE POPAYÁN
Programa de ingeniería de sistemas
Línea(s) de investigación: sistemas telemáticos inteligentes
Popayán, abril de 20

Contenido

Índice de tablas

Índice de ecuaciones

Índice de ilustraciones

CERTIFICACION DE AUTORIA

Certificamos que conocemos el concepto de plagiar según la Real Académica de la lengua ("Copiar en lo sustancial obras ajenas, dándolas como propias.")

Y certificamos que el contenido de este documento es de nuestra autoría, no hay contenido que haya sido copiado directamente y al pie de la letra de ninguna fuente. En el caso de ideas, teorías, conceptos, resultados y otros contenidos tomados de otros autores se menciona explícitamente la fuente original, y sólo en unos pocos casos se han mantenido el mismo texto, colocándolo entre comillas.

Reconocemos las consecuencias académicas, jurídicas y económicas que conlleva el plagio.

Firma

Fabian A. Palacios Abadia

Fabian Antonio palacio Abadía

CC. 1.123.305.128

Firma

Nelson Pastor

Nelson Andrés Pastor Patiño

CC.1.061.743.290

Resumen

Esta investigación se plantea desde la necesidad de una empresa de la ciudad de Popayán que genera sus ingresos con la venta de productos de consumo masivo y que a su vez quiere también conocer la distribución de sus clientes para lograr la fidelización de su marca. Para darle solución a dicha problemática hemos tomado como muestra los datos de las transacciones del año 2019 de 2837 clientes y le implementamos la técnica de clustering por medio del modelo RFM en Excel y la implementación del algoritmo de K-Means realizado en Python y la herramienta Weka, en este desarrolló utilizamos las fases de la metodología CRISP-DM dándonos como resultado 5 segmentos de los clientes de la empresa comercializadora de productos lácteos en el modelo RFM y 7 en K-Means, permitiéndole a la empresa el uso de esta información para generar estrategias de marketing.

Palabras claves: Clientes, segmentación, clustering, RFM, K-Means y CRISP-DM.

Introducción

Las grandes empresas del mundo desde hace ya muchos años vienen implementando tecnologías y aplicaciones que segmentan sus clientes. Permitiendo a las empresas formular estrategias de marketing, tener una mejor atención al usuario teniendo en cuenta sus necesidades y una mayor venta de sus productos que tienen una baja rotación, gracias a la fidelización de dichos clientes aportando mayores ingresos para las empresas.

El entorno empresarial cada día es más competitivo y es necesario que los clientes que se tienen en las empresas no se pierdan sino al contrario que día a día sean aún más fieles a las marcas de una empresa, es por esto que nacen nuevas técnicas que nos permiten segmentar a nuestros clientes en diferentes grupos ya sea por sus gustos, su número de compras, su edad, su género, su estado civil, su nivel social o su ubicación demográfica.

Dentro de las técnicas minería de datos podemos encontrar el clustering, las redes neuronales y los árboles de decisión entre otros. Para lo cual es importante la preparación de los datos y las metodologías que nos permite evaluar los modelos y generar excelentes resultados que mejoran aún más la información recolectada y convertir esa información en el mayor activo para la empresa sacándole el máximo beneficio.

CAPÍTULO I ASPECTOS GENERALES DE LA INVESTIGACIÓN

1.1 Planteamiento del problema

Bajo la dinámica actual de los negocios, sobresalir ante la competencia hace imperativo identificar grupos de clientes con necesidades, características o comportamientos en común, permitiendo optimizar la atención y el servicio, logrando clientes satisfechos y leales a la empresa y generando una relación de largo plazo con ellos. Para lograr dicho propósito se implementa un proceso denominado Segmentación de clientes, el cual es definido por BBVA como: "una tarea que consiste en dividir en pequeños grupos homogéneos de clientes en un mercado concreto". Su objetivo fundamental es el de poder determinar con precisión las necesidades de cada grupo, de tal manera que la empresa las pueda atender mejor, ofreciéndole a cada uno de ellos un producto o servicio adecuado. (BBVA, 2017)

Para realizar segmentación de clientes existen diversas técnicas que varían en función del mercado objetivo, el tipo y dimensionalidad de los datos. Dentro de las más populares se encuentra el análisis RFM, él cual es un método cuantitativo de segmentación de clientes. En su versión estándar está diseñado para trabajar únicamente con variables de tipo transaccional (Frecuencia, Recencia y Monto), convirtiéndolo en un método practico, fácil de implementar y que ofrece resultados a corto plazo, aunque en su versión extendida es posible considerar más variables. Está fundamentada en el principio de Pareto, también denominado como la regla del 80:20, la cual pondera que, en proporción; el 80% de las consecuencias se derivan del 20% de las causas. En el sector comercial este principio infiere que cerca del 80% de las ganancias de una empresa la genera el 20% de los clientes, o que aproximadamente el 80% de las ganancias provienen del 20% de los productos. (Córdoba, 2011)

Otra tecnología comúnmente utilizada para segmentar clientes es Machine learning, la cual puede definirse como un método analítico que permite que un sistema, por sí mismo sin intervención humana y en forma automatizada, aprenda a descubrir patrones, tendencias y relaciones en los datos. (Alpaydin, 2020)

En el mundo empresarial actual, el acceso a la información de manera clara, precisa y oportuna se ha convertido en uno de los principales activos de las empresas, por tal motivo es imperativo incurrir en estudios que permitan a las empresas identificar diferentes grupos de clientes con características de consumo afines, lo que posibilitará conocer el valor que representa cado uno

de los clientes respecto a la empresa, permitiendo realizar campañas de fidelización y retención, de manera eficiente y efectiva, al enfocar los recursos exactamente a aquellos grupos de clientes a los que se quiere llegar, y entregándoles soluciones o productos que requieran.

En la ciudad de Popayán las empresas guardan toda clase de información relativa a las operaciones diarias que se desarrollan en sus establecimientos; sin embargo, muchas no explotan su valor pues la información que se encuentra implícita en estas bases de datos no es fácil de discernir, debido a su elevada dimensionalidad.

La empresa comercializadora que se analiza en este proyecto no es ajena a esta realidad. La actividad comercial de la empresa comercializadora de lácteos, al ser de consumo masivo tiene como mercado objetivo la población en general. La información de los clientes y las ventas realizadas de la empresa, son registradas en hojas de Excel para la gestión comercial de la misma; los datos almacenados son utilizados para analizar el comportamiento de inventario y proyectar las compras a los proveedores en los diferentes períodos comerciales del año, pero no se hace uso de la información del cliente para ningún proceso de marketing, lo cual hace que pierdan competitividad.

Existen muchos proyectos que han desarrollado estrategias similares en la segmentación de mercado y/o clientes en empresas pymes en Colombia, utilizando herramientas CRM (**Gestión de Relaciones con Clientes)** disponibles en el mercado, las cuales permiten:

- Agilizar la atención con los clientes: poder fidelizarlos reduciendo el tiempo de espera en las consultas.
- Aumentar la productividad: controlando la información de los clientes se obtiene un mayor volumen en la productividad.
- Realizar campañas de marketing específicas: a través de las redes sociales y para cada tipo de cliente, ya que se tiene toda la información acerca de ellos.
- Automatización marketing: las herramientas de CRM pueden automatizar las tareas repetitivas para mejorar los esfuerzos realizados en marketing.

Sin embargo, estas opciones no han sido aplicadas en la empresa que hace parte de esta investigación por falta de conocimiento en dichas tecnologías, personal poco capacitado en el tema tecnológico y la falta del conocimiento que tienen los directivos hacia la proyección que puede tener el negocio a la hora de aplicar estas herramientas.

1.1.1 Pregunta de Investigación

¿Qué beneficios puede obtener para la empresa comercializadora de productos de consumo masivo en Popayán la segmentación de los clientes utilizando aprendizaje de máquina?

1.2 Objetivos

1.2.1 Objetivo general

Caracterizar los clientes de la empresa comercializadora de lácteos en la ciudad de Popayán, implementado unsupervised machine learning y análisis RFM.

1.2.2 Objetivos específicos

- Realzar una revisión del estado del arte sobre machine learning para segmentación de mercados.
- Definir un modelo de segmentación de los clientes de una empresa comercializadora de lácteos en Popayán soportado en machine learning.
- Implementar Análisis RFM para la segmentación de clientes de la empresa comercializadora de lácteos en Popayán.
- Evaluar el modelo RFM mediante análisis de resultados y el modelo de clustering mediante métricas de validación internas.

1.3 Justificación

En la obra de Roberto Hernández Sampieri (Sampieri, Collado, & Lucio, 1996) se exponen algunos criterios para evaluar la importancia potencial de una investigación, los cuales fueron adoptados para justificar este estudio:

- **Valor metodológico de la investigación**:

Este proyecto busca, con base a los patrones de compra y otros factores identificados en información recolectada a través de la empresa distribuidora de productos de consumo masivo durante el año 2019, definir un modelo de clustering evaluado mediante las métricas de validación interna y externa más comunes identificadas en el análisis de la revisión bibliográfica: suma de error cuadrático (SSE), el índice Dunn y el índice Davies-Boulding, distancia Euclidiana, distancia de Manhattan y coeficiente correlación de Pearson (Grabusts, 2011), (Maimon & Rokach, 2010).

Implementar y evaluar el Análisis RFM para segmentación de clientes, y contrastar los resultados obtenidos con el modelo de machine learning.

- **Valor práctico de la investigación**:

El proceso de segmentación de los clientes de la empresa comercializadora de productos lácteos en la ciudad de Popayán, le permitirá a la empresa identificar grupos de clientes con diferentes necesidades, características y comportamientos que requieren estrategias de marketing diferenciadas.

- **Valor tecnológico**:

Desarrollo de una herramienta software integral de Minería de datos y Machine Learning, la cual permitirá al personal asignado del área, realizar actualizaciones de los clústers de clientes a lo largo del tiempo (Agregar nuevos clientes o actualizar datos).

- **Valor de Emprendimiento e Innovación**:

A mediano plazo (fase siguiente de este proyecto) se desarrollará una herramienta software usable con un algoritmo de segmentación de clientes para la empresa comercializadora de productos lácteos en Popayán. La herramienta tiene como objetivo realizar actualizaciones al modelo desarrollado en la fase anterior, lo que permitirá la adaptabilidad del modelo a las actualizaciones futuras en los datos de los clientes.

CAPÍTULO II MARCO REFERENCIAL

La presente investigación se orienta a la implementación de Machine learning en la segmentación de mercado para clientes de productos de consumo masivo, utilizando las metodologías de segmentación, las bases de datos y los diferentes framework.

2.1 Marco conceptual

2.1.1 Segmentación de mercado: (Sanchéz Galán, 2019) afirma que "es un proceso de marketing mediante el que una empresa divide un amplio mercado en grupos más pequeños para integrantes con semejanzas o ciertas características en común".

2.1.2 Cliente: Desde el punto de vista de la economía, hace referencia es una persona natural o jurídica la cual tiene de manera frecuente u ocasional, una relación comercial que involucra bienes, productos o servicios; los que pone a su disposición un profesional, un comercio o una empresa. (significados, 2015)

2.1.3 Mercadeo: (Vergara, 2019) **"**Consiste en un proceso administrativo y social gracias al cual determinados grupos o individuos obtienen lo que necesitan o desean a través del intercambio de productos o servicios".

2.1.4 Marketing personalizado: "Es la implementación de una estrategia mediante la cual las empresas entregan contenido individualizado a los destinatarios mediante la recopilación de datos, el análisis y el uso de la tecnología de automatización" (manzana, 2019).

2.1.5 Machine Learning: Conocido en español como Aprendizaje Automático, según (El naga & Murpy, 2015, pág. 6) es una disciplina científica en el campo de la Inteligencia Artificial (IA). Básicamente, es una rama en desarrollo de los algoritmos computacionales diseñados para simular la inteligencia humana al aprender del entorno circundante. Las técnicas basadas en Machine Learning se han aplicada en diferentes ámbitos que van desde la ingeniería de naves espaciales, las finanzas hasta las aplicaciones médicas.

2.1.6 Aprendizaje supervisado: Según (Gago Utreta, 2017) "los datos en estos casos disponen de atributos adicionales que son los que se intentan predecir. Dentro de esta categoría destacan los algoritmos de clasificación, en los que las muestras están

etiquetadas como como pertenecientes a dos o más clases y se requiere aprender a predecir la clase de datos sin etiquetar".

2.1.7 **Aprendizaje no supervisado**: Según "los datos de entrenamiento consisten en un conjunto de vectores sin ningún valor o etiqueta correspondiente. El objetivo en estos casos puede ser descubrir grupos de ejemplos similares dentro de los datos"

2.1.8 **El análisis RFM**: (por Recency, Frequency, Monetary) es una técnica de marketing usada para determinar cuantitativamente el valor que representa cada uno de los clientes para la empresa. Esta técnica permite identificar por medio de segmentación, los clientes fieles, así como también aquellos a los que se necesita enfocar esfuerzos de fidelización y retención. Esto se consigue examinando tres factores sobre la información de las tracciones comerciales realizadas por el cliente, los cuales son: (R) Recencia de compra, (F) Frecuencia de compra y (M) Monto de la compra en términos monetarios. (Morelo Tapias K. A., 2014)

2.1.9 **Clustering:** También conocido como agrupamiento, es una de las técnicas de minería de datos, el proceso consiste en la división de los datos en grupos de objetos similares. Cuando se representan la información obtenida a través de clústers se pierden algunos detalles de los datos, pero a la vez se simplifica dicha información. (Ecured, 2017)

2.2 Estado del arte

En esta sección se utilizó un mapeo sistemático (Carrizo & Ortiz, 2016). Este comienza con la especificación de las preguntas de investigación que se desean responder en el estudio. En este caso, el objetivo de la investigación se declara con una pregunta principal y tres secundarias" (Carrizo & Ortiz, 2016).

La pregunta principal es:

RQ1: ¿Cómo realizar segmentación de clientes de una empresa utilizando técnicas de clustering?

para el presente estudio, las 3 preguntas de investigación secundarias planteadas son:

2. ¿Qué estudios se han desarrollado para la segmentación de mercado?

3. ¿Existen procesos para la aplicación segmentación de mercado utilizando machine learning?

4. ¿Hay alguna metodología específica para segmentación de mercado?

2.1.10 Búsqueda de la literatura

En la elección de la literatura adecuada para la investigación se implementarán los 3 siguientes pasos: 1) establecer las palabras claves para la búsqueda, 2) establecer las bases de datos bibliográficas en las se establecerá la búsqueda, 3) definir las cadenas de búsqueda.

1) establecer las palabras claves para la búsqueda: se tiene en cuenta las palabras más relevantes para iniciar la búsqueda, esto nos puede conceder una información adecuada en los resultados, las palabras que se definen son:

a. "Marketing"
b. "Market segmentation"
c. " Data mining"
d. " Personalized marketing "
e. "Clustering"
f. "CRISP-DM"
g. "Machine Learning"

2) Bases de datos bibliográficas: se implementa la búsqueda en las siguientes bases de datos:

A. Google Scholar
B. RedIB
C. Scopus
D. Ebsco

3) Definición de cadenas de búsqueda: en cada una de las bases de datos se observan resultados diferentes, al igual que plantea diferentes formas de realizar las búsquedas, por lo tanto, se implementan cadenas simples de máximo dos palabras claves. Cada cadena tendrá en una de sus palabras claves alguna variación.

De esta forma se plantean las siguientes cadenas de búsqueda:

A. “Marketing " AND " Market segmentation " OR "Data mining"
B. “Market segmentation” AND ("Data mining" OR “Clustering” OR “Machine Learning” OR “CRISP-DM”)
C. “Personalized marketing” AND ("Data mining" OR “Clustering” OR “Machine Learning” OR “CRISP-DM”)

A partir de estas cadenas se realizaron 3 búsquedas en cada una de las bases de datos, en las que se observan 4 aspectos relevantes: título, resumen, introducción y conclusiones. Esto permitirá decidir cuáles son los artículos importantes para responder a las preguntas de investigación planteadas. En total se implementarán 12 búsquedas independientes de las cuales se muestran los resultados en la tabla 1 con el total de artículos encontrados en cada base de datos según las cadenas de búsqueda. En la primera etapa se encuentran una gran cantidad de artículos que no eran correspondientes para la investigación realizada, por esta razón, no se analizan casos muy específicos.

Tabla 1 Resultados bibliográficos

Búsqueda	Google Scholar	RedIB	Scopus	Ebsco
“Marketing" AND "Market segmentation" OR "Data mining"	2500	1591	10	20
“Market segmentation” AND ("Data mining" OR “Clustering” OR “Machine Learning” OR “CRISP-DM”)	58	11	6	4
“Personalized marketing” AND ("Data mining" OR “Clustering” OR “Machine Learning” OR “CRISP-DM”)	4	2	2	1

2.2.2 Selección de los artículos relevantes

Dentro del proceso del mapeo sistemático se debe elegir de manera cuidadosa los artículos que nos pueden contribuir información o evidencia directa para dar le respuestas a las preguntas planteadas. Para esto se definen criterios de inclusión que nos muestra las características que deben cumplir los sujetos o unidades de observación para participar y los criterios de exclusión que nos definen las características que impiden participar en la investigación, con la intensión de enfatizar los artículos con información específica que aportan a la respuesta de las preguntas planteadas anteriormente, por esto, los juicios definidos para la presente investigación son:

Exclusión: Artículos, revistas de investigación o tesis que tengan similitud con la investigación que estamos realizando, donde se muestre la aplicación de minería de datos, la

implementación Clustering en la segmentación de mercado, la fecha de la implementación de la investigación, los algoritmos implementados y las variables utilizadas para la investigación.

Inclusión: artículos, revistas trabajos de grado que se encuentren publicadas en librerías top, dónde se muestre el modelo de estudio, un análisis de minería de datos en la segmentación de mercados, fuentes de datos que utilizaron, implementaciones generales, organizaciones y análisis sociales.

Se realiza la lectura de: título, resumen, introducción y conclusiones de los artículos. A partir de los principios mencionados anteriormente se descartaron los artículos que a pesar de tener una o todas las palabras claves de la búsqueda su tema no está directamente relacionado con la investigación, o que cumplieran los criterios de exclusión. De esta manera la lista inicial de documentos se reduce hasta llegar a 10 artículos.

2.2.3 Clasificación de los artículos

En este paso de la metodología se debe rememorar el objeto central de la investigación, solucionar la(s) pregunta(s) de investigación y fijar los parámetros o criterios necesarios a observar para lograr al final del estudio responder a estas. Para la clasificación, en este caso se implementan los siguientes criterios de análisis:

1. Tecnología
2. Metodología
3. Fuente de los datos
4. Implementación automática
5. Aplicación de Minería de datos
6. Año de implementación

2.2.4 Extracción y agregación de datos

Determinamos los criterios para la clasificación de los artículos, se procede a leer de manera independiente la totalidad de cada uno de los 10 artículos elegidos, al mismo tiempo se

establecen los datos en una ficha bibliográfica que contiene una descripción del contenido del artículo al igual que el análisis de las características definidas en la fase anterior (extracción de los datos). A partir de este proceso se obtiene un análisis con las inclinaciones de estudio en la

implementación de minería de datos en la segmentación de mercado y las tecnologías implementadas. Los datos alcanzados se presentan en la tabla 2.

Dentro de la tabla se utilizan acrónimos que se encuentran o se referencian en cada uno de los artículos, tanto para la fuente de datos como para la tecnología. Específicamente la sigla. La tabla 2 permite observar de una manera ordenada la información para responder directamente las preguntas de investigación formuladas en secciones pasadas, a continuación, se presenta la forma en que se interpretó la información para dar respuesta a cada una. La tabla está compuesta por cinco columnas; cada columna tiene una descripción, en la primera se tiene el título que indica el nombre del documento, en segundo lugar "imple" que nos indica cual fue la tecnología implementada y el algoritmo que se desarrolló en cada investigación, en tercer lugar "var" nos indica cuales fueron las variables de estudio en cada uno de las investigaciones, en cuarto lugar las fuentes de datos esto nos indica si fueron externas o internas, el quinto que nos indica el año de publicación de cada artículo y por último la referencia que es la referencia bibliográfica de cada una de las investigaciones.

Tabla 2 Selección de artículos relevantes

TITULO	**IMPLE**	**VAR**	**FUENTE DE DATOS**	**AÑO**	**REFERENIA**
Aplicación de minería de datos para la segmentación de clientes y desarrollo de estrategias de comunicación para la empresa DPC Studio S.A.S.	K-Means	Id del cliente, nombre del cliente, los portafolios de los productos y el tipo de producto que compra.	Externos	2013	(Cubides Proaños, 2013)

Sistemas para caracterización de perfiles de clientes de la empresa Zona T.	k-Means, K-SSE	Datos sociodemográficos y datos de comportamientos de clientes.	Externos	2014	(Morelo Tapias K. a., 2014)
Programa para la identificación del comportamiento del cliente de MIPYES con base en RFM.	Análisis de RFM	Recency, frequency, monetary y una variable independiente de comportamiento de los clientes.	Externos	2014	(Jacome Ortega & Mariella, 2014)
Segmentación de mercados sobre datos de alta dimensión.	Post-hoc, K-Means	Visitas de usuarios a la página Web, la fecha, el medio empleado para conectarse, el documento visitado y la ubicación del usuario.	Externos	2015	(Betanzos, Berdinas, Betanzos, & Antonio, 2015)
Propuesta de plataforma de procesamiento de datos para marketing directo.	Weka, Moa, Samoa y Apache Spark	Datos demográficos e históricos de compras.	Externos	2016	(Rivera, Peña, & Martinez, 2016)
Big Data para la segmentación de mercado en	Java	Resultados de la toma de encuestas.	Externos	2016	(Chirinos, 2016)

redes sociales en accesorios de moda emergente.					
Estudio del tanger objetivo de la empresa Madecentro Colombia S.A.S sucursal Santander.	Clúster Ward, software SPSS.	Resultados de encuestas personales a cada uno de los clientes (variables de los clientes, comportamiento de la compra y el uso de productos).	Externos	2016	(Rincon Boneth, 2016)
Análisis multivariado para la segmentación de clientes basado en RFM.	Análisis multivariable y modelo RFM	Recency, frequency, monetary, ganancia, porcentaje de la ganancia y días de vencimiento de las facturas.	Externos	2017	(Cuadros López & Gonzales Caicedo, 2017)
Metodología de análisis y segmentación de clientes usando secuencias de comportamiento.	MSC2	Conjunto de datos dinámicos a partir de datos reales.	Externos	2019	(casariego, 2019)
Solución basada en inteligencia de negocio para apoyar a la toma	Data Marts, K-Means	Resultados de encuestas aplicadas a personal de la	Externos	2019	(Cornejo Arce, 2019)

de decisiones en el área de ventas de una empresa en la ciudad de Chiclayo.		empresa y datos recolectados a los clientes (género, estado civil y productos que más compran).			
Análisis del modelo RFM según el método convencional y el método de las 2-tuplas	Análisis RFM	Frecuencia, Recencia, Monto	Externos	2016	(María Mas Diaz, 2016)

2.3 Antecedentes

El análisis de los antecedentes que se realiza se agrupa en tres tipos de investigación: investigaciones internacionales para efecto del trabajo de investigación, aporta a las diferentes investigaciones ya realizadas a nivel internacional sobre el tema de segmentación de mercado utilizando Minería de Datos, a su vez se observan investigaciones nacionales que muestran los aportes que se han realizado en el país y por ultimo las investigaciones locales dando un punto de inicio a la hora de mejorar esta problemática en nuestra región.

Algunos autores, han realizado una ardua investigación dando su punto de vista sobre la implementación de herramientas para la segmentación de mercado en productos del consumo masivo, permitiendo desarrollo de dichas capacidades gracias a la implementación de las metodologías.

2.3.1 Investigaciones internacionales

(Jacome Ortega & Mariella, 2014) desarrollaron una aplicación para la identificación del comportamiento con base en el análisis de RFM y la clasificación de los clientes de acuerdo con la

fase de vida. Esta aplicación se enfocó a las empresas comerciales micro, pequeñas y medianas empresas (MIPYMES). Se definieron tres variables independientes que son: Recency, Frequency, Monetary y una variable dependiente será la de segmentos, que es la encargada de recoger el comportamiento de los clientes. El programa desarrollado funciona como aplicativo a una hoja de cálculo de electrónica de Microsoft Excel que copila la operación de las variables definitivas.

Un grupo de investigadores de la ciudad de Coruña- España, implementaron un algoritmo de agrupamiento de datos diádicos (**post-hoc**) en una plataforma llamada **Apache Spark** que se utilizó en la segmentación de mercado para la empresa **outbrain** que aporto los datos de sus clientes. La información que se tuvo en cuenta para este estudio fue la siguiente: Las visitas de usuarios a la página web, la fecha, el medio empleado para conectarse, el documento visitado y la ubicación. Con estos datos se realizaron 2 experimentos. En primer lugar, se calculó un agrupamiento con una entropía ponderada comparable a la obtenida para una agrupación de 100 grupos obtenidos mediante **K-Means**. El segundo consistió en comparar el efecto sobre el tiempo de ejecución de más nodos de cómputo para el cálculo distribuido. (Betanzos, Berdinas, Betanzos, & Antonio, 2015)

.

(Rivera, Peña, & Martinez, 2016) propusieron la implementación de una plataforma de procesamiento de datos basada en tecnologías de software libre y con un alto nivel de escalabilidad en un proyecto de marketing personalizado. Dado que la naturaleza de las fuentes de datos generadas por los hábitos de consumo de los clientes resulta propagada, ellos propusieron un procesamiento de flujo de datos distribuido, capaz de resolver las tareas de cada una de las etapas de la metodología **CRIPS-DM.** Para esto utilizaron las siguientes herramientas: **WEKA, MOA, SAMOA Y APACHE SPARK**.

En el procesamiento distribuido de flujos de datos demográficos e históricos de compras procedentes de **Bank Marketing Data Set** se ejecutó **SAMOA** que fue configurado con varios motores de procesamiento **(SPE)** entre ellos **STORM** y **S4.** Estos se encargaron de tareas como serialización de datos, los cuales son evaluados en **SAMOA** y **WEKA** para las etapas de procesamiento y modelado de datos (Rivera, Peña, & Martinez, 2016) .

En Venezuela se desarrolló una investigación sobre **Big Data** para segmentación de mercados en redes sociales en accesorios de moda emergente. Esta investigación se dividió en dos

etapas: inicialmente se realizó un análisis de la problemática y luego se establecieron las bases teóricas que daban sustentación a los planteamientos de esta investigación; de manera que posibilito la estructura basada en su sistema de variables con sus respectivas dimensiones, subdimensiones e indicadores. (Chirinos, 2016)

Esta investigación se realizó con un enfoque cuantitativo, el cual se fundamenta en la recolección de datos para la medición de fenómenos sociales. Como población muestra se tomaron 3 profesores y 384 encuestados a los que se les realizo 2 encuestas: la primera de escala tipo **Likert** de 9 ítems y la segunda de selección y consta de 23 ítems. Esta población muestra se dividió en 3 grupos: **grupo "A"** (usuarios reales y clientes potenciales de diseñadores de accesorios de moda emergente), **grupo "B"** (profesores de las escuelas de computación de la universidad Rafael Urdaneta y la universidad privada Dr. Rafael Belloso con experiencia de 1 a 3 años) y **grupo "C"** (communy managers de habla hispana que son miembros comunidades activas orientadas a la gestión de redes sociales dentro de Facebook). Para la recolección de los datos extraídos de las redes sociales se utilizó un programa desarrollado en java. Los datos recolectados de las encuestas realizadas al **grupo "A"** y los extraídos de sus respectivos perfiles de Instagram se codificaron agrupando las respuestas de las preguntas con sus números correspondientes. En el análisis y modelado de algunos datos se utilizó la metodología **CRISP-DM,** con las siguientes fases: Análisis de datos a menor escala y extracción análisis de las redes sociales. (Chirinos, 2016)

En Madrid-España (casariego, 2019) defino una metodología de segmentación basada en secuencias de comportamiento (**MSC2**) enfocada en el comportamiento dinámico de los clientes. El objetivo de esta metodología **MSC2** consiste en proporcionar a los decisores de marketing una herramienta que complementa las actuales prácticas de clientes, ciudadanos y pacientes. Donde se pueden ver patrones de comportamiento, identificar su evolución y anticiparse al mismo. La investigación se implementó en dos escenarios distintos (**tienda de modas y servicios sanitarios**) y se utilizaron dos conjuntos de datos sintéticos a partir de datos reales.

En la ciudad Chiclayo realizo una investigación basada en una solución inteligente de negocio para apoyar a la toma de decisiones en el área de ventas de una empresa comercial de la ciudad Chiclayo. Como población, muestra y muestreo se tomaron 3 empleados de la empresa que son el gerente, jefe marketing y el supervisor de ventas, que son las personas que más conocen de las ventas de la empresa y a los que se les efectúo una encuesta de satisfacción de la información solicitada, otra encuesta de utilidad y facilidad de uso percibido de la solución implementada y una

entrevista. Los aspectos que se tuvieron en cuenta de los clientes para ejecutar la segmentación fueron los siguientes: género, estado civil y productos que más compra y con estos datos ofrecer las promociones adecuadas y los descuentos más justos. En la implementación de la solución de inteligencia de negocio se utilizó la metodología **Kimball** permitiendo la implementación de **Data Marts** para el área de ventas y luego formar el **Data Warehouse,** teniendo así el enfoque de menor a mayor en los datos de los clientes, también se utilizó la metodología **CRISP-DM** para la parte de modelado de minería de datos para la implementación del algoritmo de clustering **K-Means** y su correspondiente interpretación. Como resultado de la implementación del algoritmo de clustering se definieron 14 segmentos clientes permitiendo a la empresa una mejora en la demanda de sus productos y un mayor ingreso en ventas. (Cornejo Arce, 2019)

2.3.2 Investigaciones nacionales

En la ciudad de Bogotá se realizó la implementación de minería de datos para la segmentación de clientes para la empresa **DPC Studio S.A.S,** para esta investigación se manipularon los datos que fueron brindados por la empresa que tenían como atributos el id del cliente, el nombre del cliente, el portafolios del producto y el tipo de producto que compra. Para el análisis de los datos y poder lograr los objetivos de la investigación se utilizó un modelo no supervisado con la técnica de segmentación y la implementación del algoritmo **K-Means** que se encargó de clasificar los clientes a partir de un conglomerado de datos. Esto dio como resultado 3 tipos de clúster que son los más relevantes para la investigación teniendo en cuenta los productos del portafolio y los sectores más relevantes (Cubides Proaños, 2013). Se observaron también los siguientes beneficios para la empresa:

- Aumento en el consumo de los portafolios.
- Ingresos mensuales constantes.
- Incremento en clientes fidelizados con la marca.

A su vez también se observan los siguientes beneficios para los clientes:

- Un contante acompañamiento de expertos.
- El manejo de su presupuesto según la necesidad del consumidor.
- Tarifas preferenciales.

En la ciudad de Bucaramanga (Rincon Boneth, 2016) realizo el estudio del tanger objetivo de la empresa Madecentro Colombia S.A.S sucursal Santander, por medio de la segmentación se conoció el perfil real de los clientes y los potenciales de la zona. Para esto se manipularon las siguientes fuentes de información: Las encuestas personales que se realizaron a los clientes de los 3 puntos de venta de la zona Santander, estas encuestas se agruparon en dos ejes temáticos que agrupan las diferentes variables de los clientes y el comportamiento de la compra y uso del producto, también se utilizaron las bases de datos suministradas por la empresa, específicamente en el área de mercadeo y Retail dando como número de clientes analizados a 1000 a los que luego se aplicó la depuración quedando como datos reales para la compañía 242 clientes. Se utilizó el software SPSS para el análisis del clustering jerárquico que permitió la segmentación de clientes por tipología, utilizando el método de agrupación de clúster Ward y una media de distancia euclídea al cuadrado. Con esta investigación se puede concluir que los clientes tienen una satisfacción favorable en el servicio, pero hay que mejorar en la calidad del producto y los precios.

(Cuadros López & Gonzales Caicedo, 2017) en la ciudad de Cali implementaron un análisis multivariado para la segmentación de clientes basado en el modelo **RFM** para una microempresa dedicada a la manufactura y comercialización de productos desechables plásticos que costa de un portafolios con más de 23 líneas que se comercializan a mayoristas y minoristas en varias ciudades del suroccidente del país. Se tomaron los datos de los clientes a los que se les realizo una depuración quedando como muestra de estudio 304 clientes que, durante 8 meses han realizado 5962 transacciones. Luego de tenerlos datos listados se calcularon las variables clásicas del modelo **RFM:**

- Reciente (R): Compras recientes**.**
- Frecuencia (F): Número de veces que cada cliente realizo una compra.
- Monetario (M): Sumatoria de todas las transacciones del periodo.

Adicionalmente se agregaron otras tres variables para la segmentación:

- Ganancia.
- Porcentaje de la ganancia.
- Días vencidos.

Para la selección de las variables se utilizó una técnica de análisis multivariable permitiendo validar que las variables que se introdujeron en el modelo, realmente ofrezcan información

adicional sobre los clientes. Como resultado de la implementación del modelo **RFM** se observa la segmentación de cinco grupos que están organizados de forma descendente en función de cada variable de análisis.

Se efectuó un sistema para la caracterización de perfiles de clientes de la empresa **Zona T**. Se tomaron los datos sociodemográficos y de comportamiento de los clientes de la juguetería Zona T del centro comercial la plazuela de la ciudad de Cartagena que dio como resultado 5 atributos cuantitativos y 6 cualitativos para un total de 11 variables de 180 clientes durante los años 2012 y 2013. En la segmentación de los clientes se utilizó el algoritmo **K-Means** al que se le realizó una adaptación creando un nuevo algoritmo llamado **K-SSE** y teniendo en cuenta esto se desarrolló una herramienta de software orientada a la web cuyo diseño fue implementado de forma adecuada a la técnica de minería de datos. Como resultado la empresa **Zona T** puede identificar los 3 grupos de clientes que posee y los productos más compran permitiendo una mayor rentabilidad. (Morelo Tapias K. a., 2014)

Todo este contexto nos lleva a pensar que cada día se hace necesaria la intervención de la tecnología en la vida cotidiana, de esta premisa nace la necesidad de incentivar la creatividad dirigida a la consecución de herramientas que permitan agilizar de manera adecuada procesos como el que nos ocupa en este documento, “las técnicas de segmentación de clientes para empresas de consumo masivo”, así pues es necesario ahondar en temas relacionados no solo al diseño de herramientas, sino más bien a temas que giren en torno al desarrollo de proyectos dirigidos a la implementación de minería de datos para la segmentación de mercados, el cual se crea de la necesidad de las empresas por conocer el comportamiento de sus clientes y las estrategias que se deben utilizar para tener una mayor rentabilidad. De ahí nace el análisis comercial de la venta productos de consumo masivo de una empresa comercializadora de lácteos en Popayán implementado técnicas de Data Analytics y Machine Learning.

CAPÍTULO III IMPLEMENTACION DEL MODELO RFM

En este capítulo se presenta la implementación del modelo RFM, el cual da cumplimiento al objetivo específico número 3, en el que nos comprometemos a: "**Implementar Análisis RFM para la segmentación de clientes de la empresa comercializadora de lácteos en Popayán**."

El modelo RFM se enfoca en el análisis de tres variables ligadas directamente con la interacción comercial de los clientes con la empresa. Las tres variables de la metodología RFM, que son sus siglas en inglés y que describen el modelo, significan:

- Recency (Recencia): es el tiempo transcurrido entre la fecha actual y la fecha de la transacción más reciente del socio.
- Frecuency (Frecuencia): es el número total de transacciones que un socio ha realizado dentro de un período determinado de tiempo.
- Monetary (Monetario): es el valor total en dinero de las transacciones realizadas por un socio dentro de un período determinado de tiempo. (Jacome Ortega & Mariella, 2014)

Por medio de la implementación de este modelo se pretende responder a la siguiente pregunta: ¿Qué valor tienen nuestros clientes?

Los datos que vamos a emplear ya se encuentran disponibles y forman parte de la información de ventas de una empresa comercializadora de lácteos en la ciudad de Popayán recolectados en el año 2019. A continuación, definimos la metodología de trabajo:

La implementación del modelo RFM se llevó a cabo en Excel y algunos procesos se implementaron en Python.

Dado que la naturaleza del modelo RFM dista mucho de un proceso de minería de datos, hemos decidido que para su desarrollo nos basaremos parcialmente en la metodología Krisp-DM.

3.1 Descripción del dataset

A continuación, se presenta las principales características del dataset sobre el cual realizaremos el análisis RFM

Tabla 3 Descripción del dataset

Características del dataset	
Fuente de datos	49 archivos de Excel
Número de registros	85538
Número de variables	112
Año de las muestras	2019
Número de clientes	2837
Variables	ejecutivo, nomejecu, proveedor, nomproveedor, cantidad, devolución, cantneta, valortotal, valordevo, valorneto, iva, impdevo, ivaneto, costo, impuesto, consumo, deporte, neto, producto, alterno, nomproducto, cliente nomcliente, ciudad, nomciud, margen, direcc, identi, comercial, dv, tipoid, telefo, poriva, línea, nomlinea, marca, nommarca, familia, nomfamilia, categoría, nomcate, grupo, nomgrupo, almacen, nomalmacen, departam, nomdepar, porcendev, margen, pormargen, venta, margenfin, ventafin, pormargenfin, posición, saldo, factor, diaruta, diaventa, fecha, factura, pedido, planilla, auxiliar, nombreaux, transpor, nombretrans, cx, cy, barriogeo, ciudadgeo, clirazons, clinombre1, clinombre2, cliapelli1, cliapelli2, mes embase, cartones, cartonven, costouni, tieneaereo, estra, nitprovee, pago, invinicial,

	iicosto, iiventaiicostoiva, iiventaiva, iacosto, iaventa, iacostoiva, iaventaiva, cubeta, área, areanombre, rango, nomrango, descto, adicion, invunidad, invventa, invcosto, notas, canal, nomcanal, negocio, nomnego,final

3.2 Cargar el dataset

Los datos de las ventas se obtuvieron divididos en 49 archivos de Excel, cada uno correspondiente a una semana del año 2019. Por motivos prácticos, fue necesario implementar una macro de Excel para unificar los archivos en un solo documento.

A continuación, en la ilustración 1 se presenta la macro implementada para unificar los archivos de Excel.

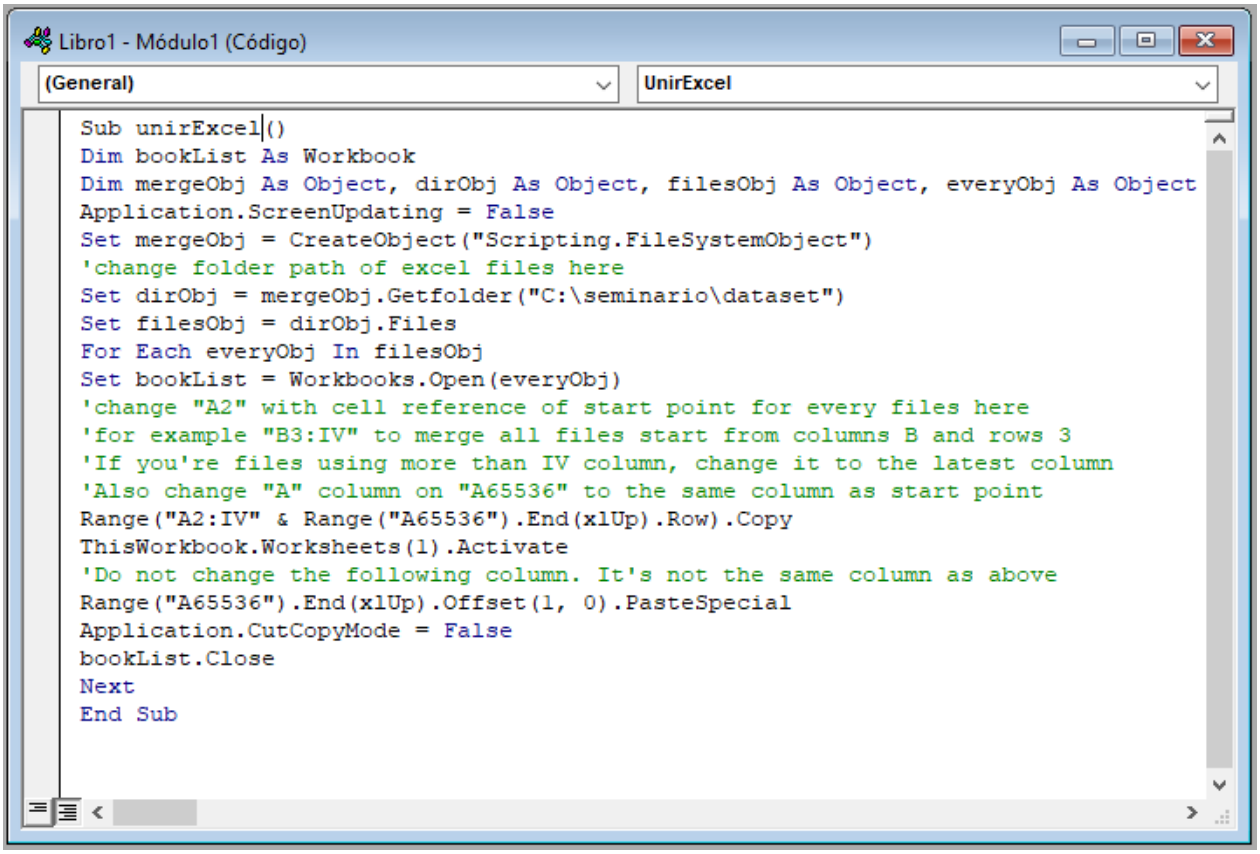


Ilustración 1 Implementación de una macro para unificar archivos de Excel

3.3 Selección de las variables de interés

Como ya se ha mencionado anteriormente, el modelo RFM se basa en tres variables (Recencia, Frecuencia y Monto), las cuales son indispensables para su aplicación, en consecuencia; se seleccionaron las siguientes variables:

a) **Fecha**: indispensable para obtener los valores de Frecuencia y Recencia.
b) **Monto_neto**: variable indispensable para la implementación del modelo, la cual hace referencia a el valor de cada una de las compras realizadas por el cliente.
c) **codigo_cliente**: variable que nos permitirá identificar la clasificación individual de cada cliente.

3.4 Identificación de datos vacíos

Una vez realizada la selección de las variables de interés, el primer paso para la comprensión de los datos es realizar una búsqueda de posibles registros vacíos o nulos, este procedimiento es indispensable para garantizar que no existan un sesgo de información en los datos y por consiguiente afectar el los resultados de nuestro modelo.

Este proceso se aplicó a las variables seleccionadas en el apartado 3.3 haciendo uso de unas cuantas líneas de código en Python, podemos observar que nuestras variables de interés no presentan campos vacíos, por lo que podemos continuar con el análisis de datos.

A continuación, se muestran los la implementación y los resultados de la búsqueda de campos vacíos.

```
#verificacion de campos vacios en las variables de interes
nombre_columnas=rfm.columns.tolist()
for column in nombre_columnas:
    print ("valores nulos en <{0}>: {1}".format(column, datos[column].isnull().sum()))
```

```
valores nulos en <valorneto>: 0
valores nulos en <cliente>: 0
valores nulos en <fecha>: 0
```

Ilustración 2 Valores nulos

3.5 Adecuación de los datos

Hasta este punto hemos reducido nuestro conjunto de datos de 85.538 registros y 112 columnas a 85.538 registros y 3 colunas. Debido a que 85.538 registros forman parte de las compras realizas por los clientes en el trascurso de un año, ahora debemos realizar un filtro avanzado para obtener el monto total y la fecha de la última compra de cada cliente en ese año.

A continuación, se indica el proceso que se implementó para obtener el monto total por cada cliente y la fecha de su última compra.

A. Se seleccionaron todos los datos y se les aplico un orden avanzado tal y como se muestra a continuación.

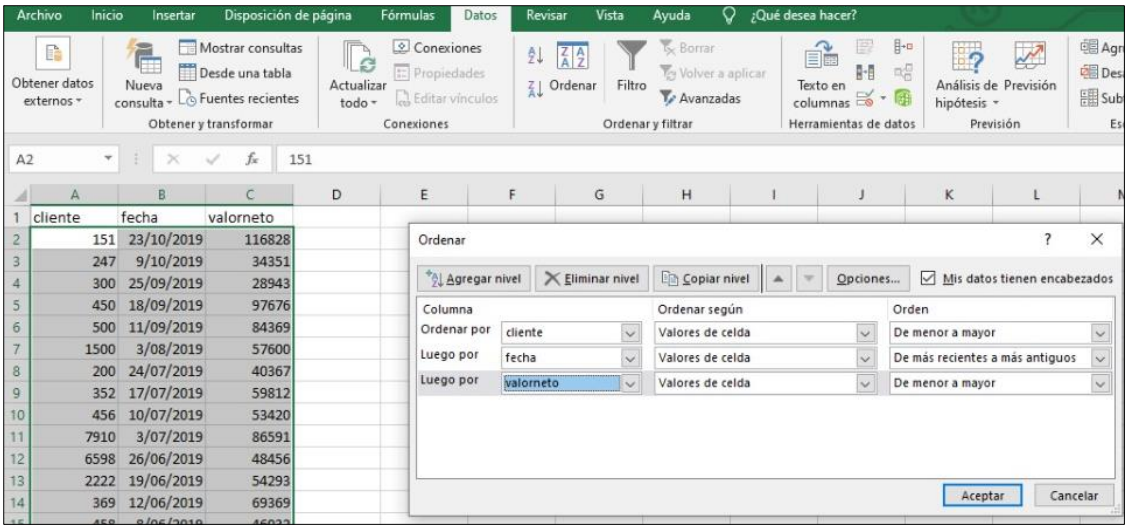


Ilustración 3 Implementación de orden avanzado

Al aplicar cambios obtenemos los datos ordenados por cliente, con sus respectivos montos de compra y la fecha ordenada de la más reciente a la más antigua, tal y como se indica en la siguiente imagen.

A	B	C
cliente	fecha	valorneto
151	23/10/2019	116828
151	9/10/2019	34351
151	25/09/2019	28943
151	18/09/2019	97676
151	11/09/2019	84369
151	3/08/2019	57600
151	24/07/2019	40367
151	17/07/2019	59812
151	10/07/2019	53420

Ilustración 4 Variables para calcular RFM

Ordenar los datos nos permite obtener la fecha más reciente de compra de cada cliente ordenándole a Excel que tome la primera fecha que encuentre para cada código de cliente.

B. Obtener valores totales por cliente

Una vez ordenados los registros, procedemos a realizar un filtro avanzado para obtener un solo registro por cliente. A continuación, se muestra la aplicación y los resultados de este procedimiento.

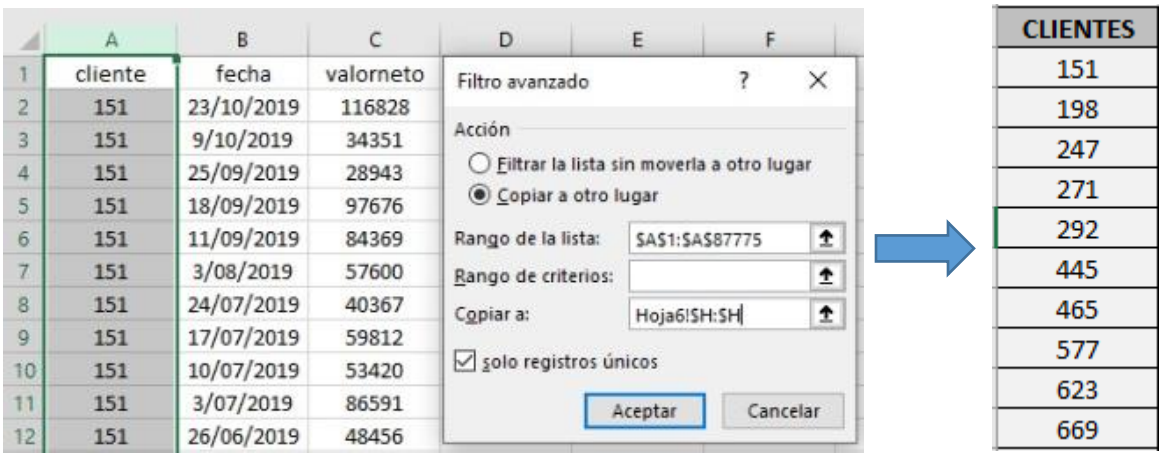


Ilustración 5 Implementación de filtro avanzado

C. **Obtener última fecha de compra**

Para obtener el registro de la última compra realizada por cada cliente se aplicó una formula a la columna de fechas, tal y como se muestra a continuación.

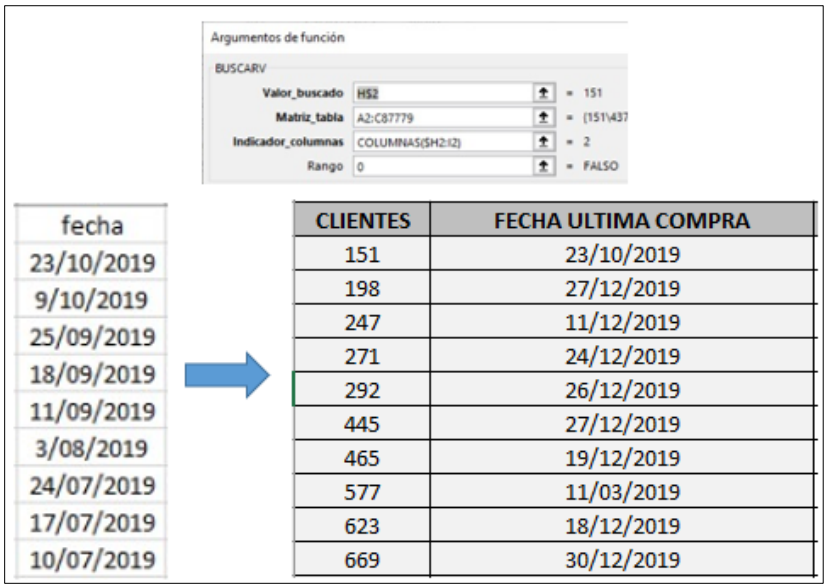


Ilustración 6 Obtener fecha más reciente por cliente

3.6 Calculo de la Recencia

Como se mencionó anteriormente, la recencia hace mención a los días transcurridos desde la última compra de un cliente. Para efectos de este proyecto se tomó como referencia la fecha 01/03/2020 la cual es la fecha hábil posterior a la última fecha de estudio de nuestros dataset (31/12/2019).

$$R = fecha\_de\_referecia - fecha\_ultima\_compra$$

Ecuación 1 Calculo de Recencia

A continuación, se presenta un fragmento de la tabla obtenida con los valores de Recencia.

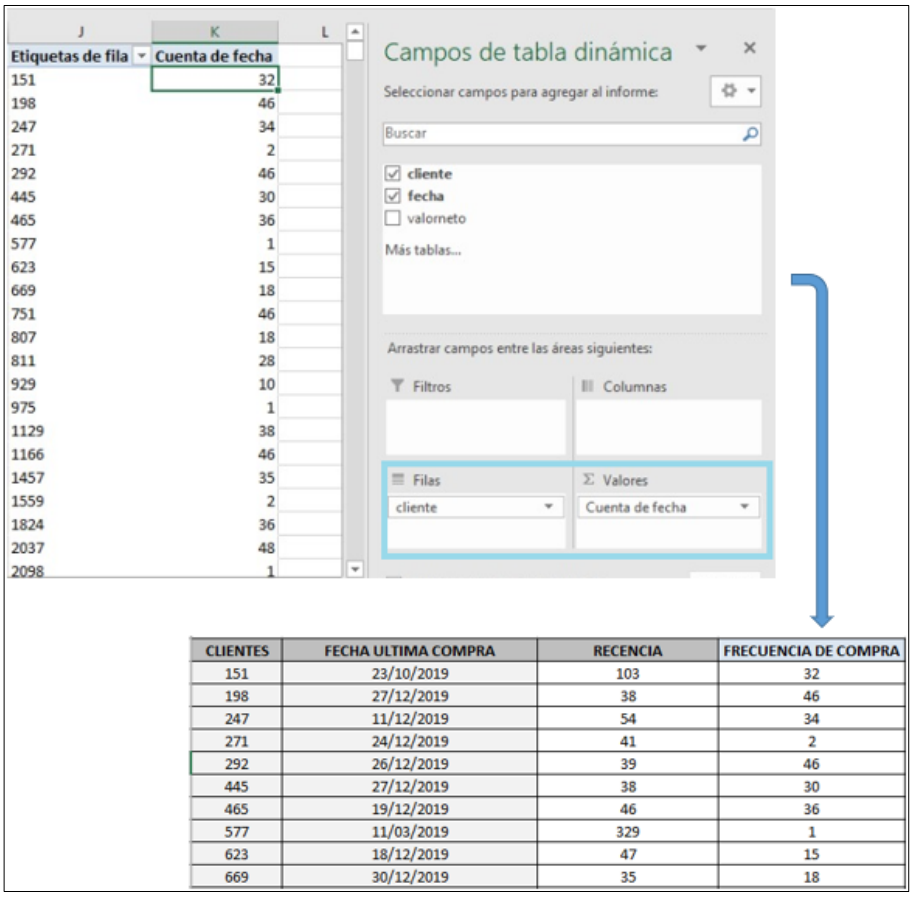
CLIENTES	FECHA ULTIMA COMPRA	RECENCIA
151	23/10/2019	103
198	27/12/2019	38
247	11/12/2019	54
271	24/12/2019	41
292	26/12/2019	39
445	27/12/2019	38
465	19/12/2019	46
577	11/03/2019	329
623	18/12/2019	47
669	30/12/2019	35

Ilustración 7 Calculo de recencia

3.7 Cálculo de la Frecuencia

Como se mencionó al inicio del capítulo, la frecuencia hace referencia a la cantidad de compras realizada por cada cliente en un rango de tiempo establecido, el cual para este caso es de 1 año. Para obtener estos valores se implementó una tabla dinámica para realizar una sumatoria de los registros de fecha filtrados por cliente, dando como resultado el número de transacciones total para cada cliente.

A continuación, se presenta la tabla dinámica implementada y un fragmento de la tabla con los valores de frecuencia obtenidos.

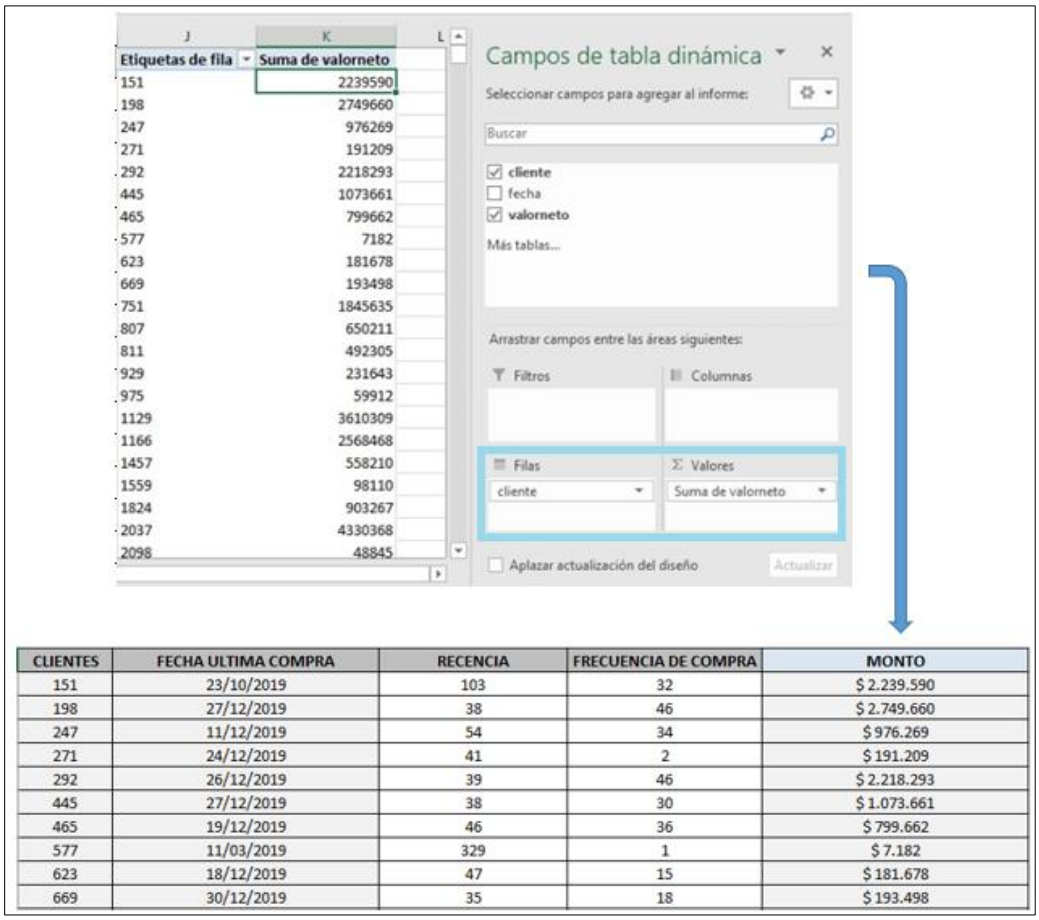


CLIENTES	FECHA ULTIMA COMPRA	RECENCIA	FRECUENCIA DE COMPRA
151	23/10/2019	103	32
198	27/12/2019	38	46
247	11/12/2019	54	34
271	24/12/2019	41	2
292	26/12/2019	39	46
445	27/12/2019	38	30
465	19/12/2019	46	36
577	11/03/2019	329	1
623	18/12/2019	47	15
669	30/12/2019	35	18

Ilustración 8 Calculo de la Frecuencia

3.8 Obtener Monto total

Para obtener el valor de la compra total realizada por cada cliente en todo el periodo del año 2019, se realizó una sumatoria de las compras totales en el año filtradas por cliente, tal y como se muestra a continuación.



CLIENTES	FECHA ULTIMA COMPRA	RECENCIA	FRECUENCIA DE COMPRA	MONTO
151	23/10/2019	103	32	$ 2.239.590
198	27/12/2019	38	46	$ 2.749.660
247	11/12/2019	54	34	$ 976.269
271	24/12/2019	41	2	$ 191.209
292	26/12/2019	39	46	$ 2.218.293
445	27/12/2019	38	30	$ 1.073.661
465	19/12/2019	46	36	$ 799.662
577	11/03/2019	329	1	$ 7.182
623	18/12/2019	47	15	$ 181.678
669	30/12/2019	35	18	$ 193.498

Ilustración 9 Calculo del monto total

En este punto ya hemos obtenidos las tres variables de RFM, también hemos transformado y reducido la dimensionalidad del dataset de 85.538 registros a 2.837. A continuación, se muestra un fragmento de la tabla final que se obtuvo.

Tabla 4 Valores de RFM

CLIENTES	FECHA ULTIMA COMPRA	RECENCIA	FRECUENCIA DE COMPRA	MONTO
151	23/10/2019	103	32	$ 2.239.590
198	27/12/2019	38	46	$ 2.749.660
247	11/12/2019	54	34	$ 976.269
271	24/12/2019	41	2	$ 191.209
292	26/12/2019	39	46	$ 2.218.293
445	27/12/2019	38	30	$ 1.073.661
465	19/12/2019	46	36	$ 799.662
577	11/03/2019	329	1	$ 7.182
623	18/12/2019	47	15	$ 181.678
669	30/12/2019	35	18	$ 193.498
751	26/12/2019	39	46	$ 1.845.635

3.9 Crear la matriz RFM

Teniendo los valores de las variables del modelo RFM nos dispondremos a realizar la matriz de RFM obteniendo los rangos de las variables.

3.9.1 Obtener puntaje de rangos para la Recencia, Frecuencia y Monto

Para obtener los rangos de la Recencia, Frecuencia y Monto primero se deben calcular los siguientes valores:

a) MIN: hace referencia a los valores mínimos de las columnas de Recencia, Frecuencia y Monto.
b) Max: hace referencia al valor máximo de la Recencia, Frecuencia y Monto.
c) RANGO: corresponde a la diferencia entre el valor máximo y el mínimo
d) N_INTERVALOR: corresponde a un valor definido a criterio del desarrollador, el cual hace referencia al número de rangos o segmentos de clientes para la clasificación RFM.
e) AMPLITUD: Es la división del RANGO entre el número de INTERVALOS.

A continuación, se presenta la tabla con los resultados de los valores anteriormente mencionados, aplicados a la recencia, frecuencia y monto.

Tabla 5 Datos de recencia, frecuencia y monto

RECENCIA		FRECUENCIA		MONTO	
MIN	34	MIN	1	MIN	934
MAX	394	MAX	49	MAX	$ 8.000.000
RANGO	360	RANGO	48	RANGO	7999066
N_INTERVALOS	5	N_INTERVALOS	5	N_INTERVAL	5
AMPLITUD	72	AMPLITUD	9,6	AMPLITUD	1599813

Una vez calculadas las variables a, b, c, d y e procedemos a calcular con ellas los rangos de puntaje correspondientes a la Recencia, Frecuencia y Monto de la siguiente manera.

a) LIMITE SUPERIO: corresponde al valor del límite inferior más la amplitud para cada intervalo.
b) LIMITE INFERIOR: inicia con el valor MIN y se le adiciona la AMPLITUD para cada intervalo.
c) RANGO DE PUNTAJE: Los rangos de puntaje están dados por el LIMITE INFERIOR y SUPERIOR para cada intervalo, así mismo se le asigna el PUNTAJE de los rangos de recencia, como se implementaron 5 intervalos la puntuación es de 1 a 5.

Tabla 6 Rangos de recencia

RECENCIA			
N_INTERVALO	**LIMITE INFERIOR**	**LIMITE SUPERIOR**	**RANGOS DE PUNTAJE**
1	34	106	5. [34- 106)
2	106	178	4. [106- 178)
3	178	250	3. [178- 250)
4	250	322	2. [250- 322)
5	322	394	1. [322- 394)

Tabla 7 Rangos de frecuencia

FRECUENCIA			
N_RANGO	**LIMITE INFERIOR**	**LIMITE SUPERIOR**	**RANGOS DE PUNTAJE**
1	1	11	1. [1- 10,6)
2	11	20	2. [10,6- 20,2)
3	20	30	3. [20,2- 29,8)
4	30	39	4. [29,8- 39,4)
5	39	49	5. [39,4- 49)

Tabla 8 Rangos de monto

MONTO			
N_RANGO	**LIMITE INFERIOR**	**LIMITE SUPERIOR**	**RANGOS DE PUNTAJE**
1	934	1600747	1. [934- 1600747,2)
2	1600747	3200560	2. [1600747,2- 3200560,4)
3	3200560	4800374	3. [3200560,4- 4800373,6)
4	4800374	6400187	4. [4800373,6- 6400186,8)
5	6400187	8000000	5. [6400186,8- 8000000)

Posteriormente, con los rangos calculados para las tres variables se construyó la matriz RFM, a continuación, se muestra un fragmento de la tabla.

MATRIZ RFM							
CLIENTES	FRECUENCIA DE COMPRA	FECHA ULTIMA COMPRA	RECENCIA	MONTO	PUNTAJE RECENCIA	PUNTAJE FRECUENCIA	PUNTAJE MONTO
151	32	23/10/2019	103	$ 2.239.590	5	4	2
198	46	27/12/2019	38	$ 2.749.660	5	5	2
247	34	11/12/2019	54	$ 976.269	5	4	1
271	2	24/12/2019	41	$ 191.209	5	1	1
292	46	26/12/2019	39	$ 2.218.293	5	5	2
445	30	27/12/2019	38	$ 1.073.661	5	4	1
465	36	19/12/2019	46	$ 799.662	5	4	1
577	1	11/03/2019	329	$ 7.182	1	1	1
623	15	18/12/2019	47	$ 181.678	5	2	1
669	18	30/12/2019	35	$ 193.498	5	2	1
751	46	26/12/2019	39	$ 1.845.635	5	5	2
807	18	29/11/2019	66	$ 650.211	5	2	1
811	28	20/12/2019	45	$ 492.305	5	3	1
929	10	08/08/2019	179	$ 231.643	3	1	1
975	1	02/04/2019	307	$ 59.912	1	1	1
1129	38	16/12/2019	49	$ 3.610.309	5	4	3
1166	46	26/12/2019	39	$ 2.568.468	5	5	2

Ilustración 10 Matriz RFM

3.9.2 Puntaje RFM

Finalmente, una vez obtenida la matriz RFM nos disponemos a calcular la columna de puntajes del modelo RFM. Se obtuvo el puntaje RFM a través del ponderado de los resultados de los rangos para Recencia, Frecuencia y Monto.

Con base en la tipología del negocio y en la investigación (Yánez Peter, 2012) en donde se dice que los valores de los rangos de RFM se pueden multiplicar por el valor correspondiente al peso asignado en wR, wF y wM, de acuerdo a la importancia que se le da a cada una de las variables del modelo RFM dentro del negocio. Se decidió dar más peso a la variable de PUNTAJE DE MONTO asignándole un peso del 40%, seguida de la variable PUNTAJE DE FRECUENCIA con

un valor del 35%, por último; la variable de PUNTAJE RECENCIA se le asigno un 25% de peso. En la siguiente tabla se presentan los pesos asignados a cada variable.

Tabla 9 Valores ponderados por variable

VALOR PONDERADO POR VARIABLE			
R	F	M	TOTAL
0,25	0,35	0,4	100%

Así mismo en relación a estos valores y al resultado de calcular w(RFM) se asignaron las siguientes etiquetas:

- CLIENTES VIP: de acuerdo a peso que se les dio a las variables, lo clientes VIP son aquellos con w(RFM) igual 5.
- CLIENTES EXCELENTES: este segmento de clientes corresponde a aquellos que tienen resultados w(RFM) mayores o iguales 4 pero menores a 5.
- CLIENTES BUENOS: corresponde al segmento con w(RFM) mayor o igual a 3.5 y menores a 4.
- CLIENTES REGULARES: el segmento de clientes regulares abarca los resultados de w(RFM) mayores o iguales a 2.5 pero menores a 3.5.
- CLIENTES POCO APORTE. Son todos aquellos clientes con w(RFM) menor a 2.5.

Cabe enfatizar que se puso como variable principal el resultado del monto, luego la frecuencia y por último la recencia ya que se dedujo con base en la tipología del negocio que el monto es la variable más importante porque hace referencia al principal valor de sus clientes y la frecuencia y recencia como el segundo y tercer ítem de más relevancia respectivamente.

A continuación, se presenta la tabla final obtenida con los valores ponderados de W(RFM) tras aplicar la multiplicación de cada variable por su respectivo peso.

PUNTAJE RFM	W(RFM)	CLASIFICACION
542	3,45	CLIENTES REGULARES
552	3,8	CLIENTES BUENOS
541	3,05	CLIENTES REGULARES
511	2	CLIENTES DE POCO APORTE
552	3,8	CLIENTES BUENOS
541	3,05	CLIENTES REGULARES
541	3,05	CLIENTES REGULARES
111	1	CLIENTES DE POCO APORTE
521	2,35	CLIENTES DE POCO APORTE
521	2,35	CLIENTES DE POCO APORTE
552	3,8	CLIENTES BUENOS
521	2,35	CLIENTES DE POCO APORTE
531	2,7	CLIENTES REGULARES
311	1,5	CLIENTES DE POCO APORTE
111	1	CLIENTES DE POCO APORTE

Ilustración 11 Puntaje RFM, W(RFM) y Clasificación

CAPÍTULO IV DESARROLLO DEL MODELO DE CLUSTERING

En este capítulo se presenta la implementación del modelo de clustering, el cual da cumplimiento al objetivo específico número 2, en el que nos comprometemos a "**Definir un modelo de segmentación de los clientes de una empresa comercializadora de lácteos en Popayán soportado en machine learning**"

El proceso de minería de datos fue ejecutado con el lenguaje de programación Python 3.7 a través de la herramienta iPython Notebook de Anaconda. Este proceso contempla la implementación de las Fases 1, 2, 3 y 4 de la metodología de desarrollo CRISP-DM.

4.1 Selección del modelo

Como lo menciona el autor (Rogers & Schroedl, 2001) K-means es el algoritmo más implementado a la hora de identificar segmentos. (Plazas Cardenas & Plazas Cardenas, 2013) Mencionan en su tesis doctoral enfocada en la segmentación de clientes, que el algoritmo de K-means se enfoca en trabajar en datos de tipo numérico, y posee una gran capacidad para trabajar con grandes volúmenes de datos a costa de un pequeño requerimiento en términos de espacio ya que el algoritmo solo almacena los puntos y sus centroides. Los autores también afirman que "En cuanto a costo computacional, K-means también requiere poco tiempo, básicamente lineal O(I $*$ k $*$ N $*$ d), donde I es el número de iteraciones requeridas para la convergencia."

Siendo consecuentes con lo anteriormente expresado y basados en el estudio realizado en la revisión bibliográfica, determinamos que el algoritmo de K-means es uno de los más idóneos el ámbito de la segmentación de clientes y el clustering en general.

4.2 Metodología

Para (Román Villena, 2016) la metodología CRISP-DM contempla el proceso de análisis de datos como un proyecto profesional, estableciendo así un contexto mucho más rico que influye en la elaboración de los modelos. Este contexto tiene en cuenta la existencia de un cliente que no es parte del equipo de desarrollo, así como el hecho de que el proyecto no sólo no acaba una vez se halla el modelo idóneo (ya que después se requiere un despliegue y un mantenimiento), sino que

está relacionado con otros proyectos, y es preciso documentarlo de forma exhaustiva para que otros equipos de desarrollo utilicen el conocimiento adquirido y trabajen a partir de él.

4.3 Comprensión del negocio

En esta fase se determina los objetivos del negocio y las necesidades actuales. Para esto es necesario realizar una revisión de las tecnologías actuales en la segmentación de clientes, la selección de variables, características relacionadas con las principales necesidades del negocio, con esta información se plantea y se limita el desarrollo del proyecto.

4.4 Fase 2 Comprensión de los datos

En esta sección se abordarán los principales conceptos relacionados con la comprensión de las características de los datos.

4.4.1 Descripción de los datos

La fase de preprocesamiento de los datos se realizó anteriormente en la implementación del modelo RFM en el capítulo 3. Por tal motivo, para la implementación del algoritmo de clustering no es necesario volver a realizar este procedimiento, puesto que ya contamos con las variables de recencia, frecuencia y monto, cliente y fecha.

Como resultado de la implementación del modelo RFM obtuvimos un conjunto de datos con las siguientes características generales:

Tabla 10 Características del dataset

CARACTERISTICAS DEL DATASET	
Numero de registros	2837
Numero de variables	5
Año a analizar	1 (2019)
Número de clientes	2837
Variables	clientes, monto, frecuencia, recencia, fecha

4.4.2 Cargar y observar el conjunto de datos

El primer paso que se realizó para la comprensión de los datos fue cargar e imprimir el dataset haciendo uso de la librería pandas, tal y como se muestra a continuación.

```
#Cargamos el dataset
datos= pd.read_csv("DatosRFM_final3.csv",header=0, sep=';', encoding='ISO-8859-1')
datos # Imprimimos los datos
```

	CLIENTES	ESTRATO	MONTO X COMPRA ULTIMA VISITA	FRECUENCIA DE COMPRA	FECHA ULTIMA COMPRA	RECENCIA	MONTO	RANGO RECENCIA	RANGO FRECUENCIA	RANGO MONTO
0	151	2	116828	32	23/10/2019	103	2239590	5	4	1
1	198	3	22267	46	27/12/2019	38	2749660	5	5	1
2	247	3	37849	34	11/12/2019	54	976269	5	4	1
3	271	3	181513	2	24/12/2019	41	191209	5	1	1
4	292	2	57175	46	26/12/2019	39	2218293	5	5	1
...	...	...	...	...	...	...	...	...	...	...
2832	9030931	2	110031	40	28/12/2019	37	5903762	5	5	2
2833	9030942	2	26785	33	16/12/2019	49	1093542	5	4	1
2834	9030946	2	146804	31	28/12/2019	37	3505463	5	4	1
2835	9030955	2	161562	46	26/12/2019	39	8209167	5	5	3
2836	9030992	2	38218	36	11/12/2019	54	2541667	5	4	1

2837 rows × 10 columns

Ilustración 12 Cargar dataset en Python

4.4.3 Aplicación de estadística descriptiva

Se implementa la función datos.describe(), la cual devuelve una tabla con diferentes parámetros estadísticos de cada variable de nuestro conjunto de datos como se muestra a continuación.

```
rfm.describe()
```

	CLIENTES	FRECUENCIA	RECENCIA	MONTO	RANGO_RECENCIA	RANGO_FRECUENCIA	RANGO_MONTO	PUNTAJE_RFM
count	2.837000e+03	2837.000000	2837.000000	2.837000e+03	2837.000000	2837.000000	2837.000000	2837.000000
mean	2.757444e+06	30.939020	63.012337	2.894454e+06	4.698978	3.595347	1.898837	38.543532
std	4.139135e+06	15.802917	62.582652	7.036791e+06	0.943190	1.514325	1.293082	36.257074
min	1.510000e+02	1.000000	34.000000	9.340000e+02	1.000000	1.000000	1.000000	1.000000
25%	1.952700e+04	18.000000	39.000000	5.126660e+05	5.000000	2.000000	1.000000	10.000000
50%	2.759500e+04	36.000000	41.000000	1.293701e+06	5.000000	4.000000	1.000000	20.000000
75%	9.000300e+06	45.000000	49.000000	3.049511e+06	5.000000	5.000000	2.000000	50.000000
max	9.030992e+06	49.000000	394.000000	1.442114e+08	5.000000	5.000000	5.000000	125.000000

Ilustración 13 Tabla descripción de datos

Esta imagen nos indica valores importantes de nuestros datos, tales como: la media, la desviación estándar, los valores máximos y mínimos, y la distribución de los datos en cuartiles.

4.4.3.1 Índices de correlación

Haciendo huso de una gráfica de mapa de calor, obtendremos el coeficiente de correlación, el cual mide grado y el sentido de la relación lineal entre dos variables cuantitativas. Este rango de medición varia valores oscilan entre -1 y 1. La magnitud de la relacion viene dada por el valor numerico reflejando el signo la direccion de tal valor. En esse sentido tan fuerte es una relacion de 1 como de -1.

Estos valores nos permiten identificar variables que estan altamente correlacionas entre si, por lo que este tipo de variables se consideran redundantes dado que no aportan informacion relevante al modelo. (Benesty, Chen, & Huang, 2009)

Teniendo ya los datos en una escala similar procedemos a generar un mapa de calor con los niveles de correlación de las variables, tal y como se muestra a continuación:

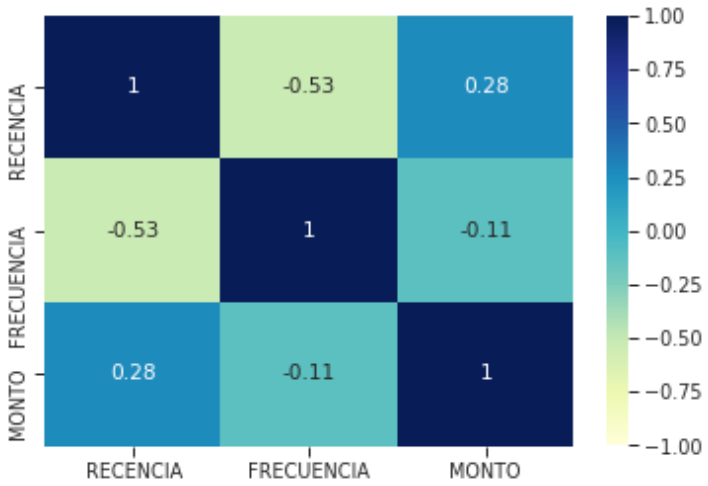


Ilustración 14 Mapa de calor de la correlación entre variables

Podemos observar que existe una mayor correlación negativa moderada-fuerte entre Frecuencia y la Recencia y una correlación positiva débil-moderada entre la variable de Recencia y Monto, y ninguna o muy débil relación entre Frecuencia y Monto.

4.5 Preparación de datos

Dado que el algoritmo K-Means, las métricas de validación internas y el índice de correlación utilizan las distancias como factor de agrupamiento, es necesario estandarizar nuestros datos con el fin de llevarlos a una misma escala y evitar que lo atributos de escala mayor dominen las distancias.

A continuación, se presentan a través de histogramas la normalización de las muestras de Recencia, Frecuencia y Monto.

a) Distribución de los datos escalados de Recencia

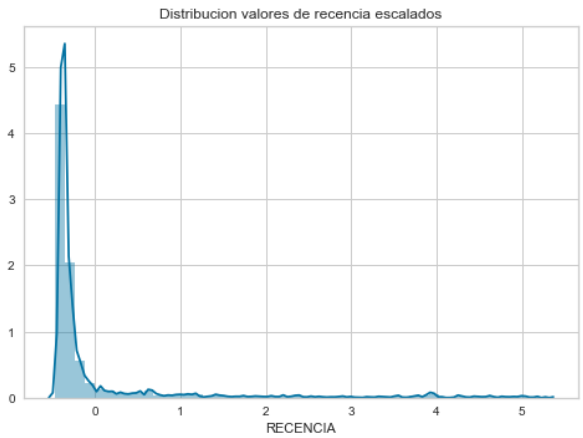


Ilustración 15 Distribución de los valores de la recencia normalizados

b) Distribución de los datos escalados de Frecuencia

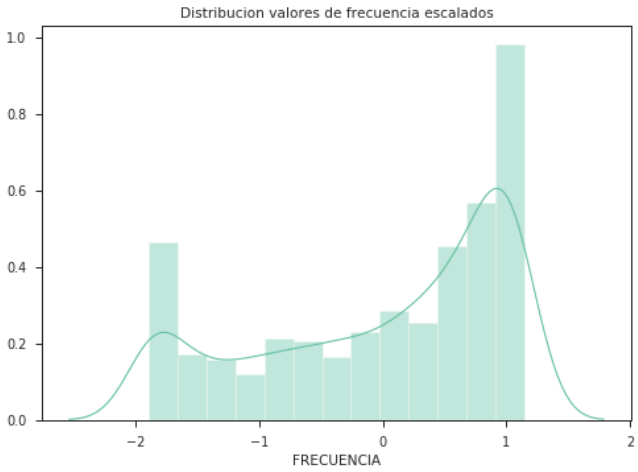


Ilustración 16 Distribución de los valores de la frecuencia normalizados

c) Distribución de los datos escalados de Monto

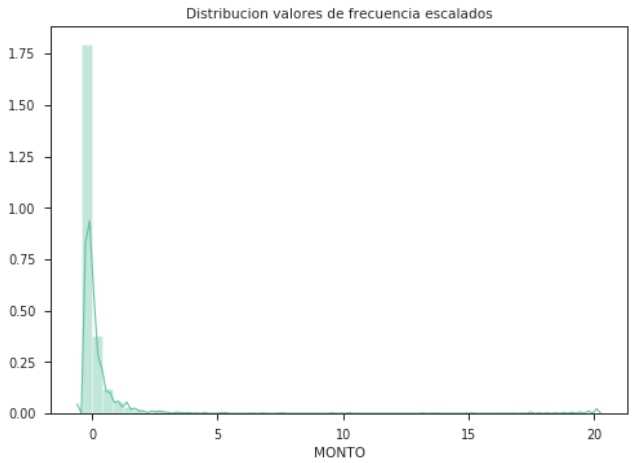


Ilustración 17 Distribución de los valores del monto normalizados

4.6 Fase de modelado

En esta fase se realiza el entrenamiento del algoritmo K-means. Aborda la selección del número de centroides y construcción del modelo.

4.6.1 Algoritmo K-Means

K-Means es un algoritmo de clustering no supervisado, ampliamente utilizado por su robustez en el tratamiento de grandes volúmenes de datos. Se ha utilizado en una variedad de dominios de aplicación, como la segmentación de imágenes (Wagstaff & Cardie, 2000) y la recuperación de información (Marroquin & Girosi, 1993). Su principal objetivo es optimizar la partición de los datos en áreas conforme a sus características implementado la minimización de las sumas de las distancias entre cada uno de los objetivos y el centroide en su clúster como se muestra en la siguiente ecuación:

$$minSE(\mu i) = minS\sum i = 1k\sum xj \in Si \parallel xj - \mu i \parallel 2$$

Ecuación 2 Reducción de distancias

4.6.1.1 Selección del número óptimo de clústers

No existe un criterio específico para la selección de numero de clústers a implementar, se pueden realizar diferentes métodos basados en medir la cohesión intra-clúster y la separación inter-clúster bajo diferentes criterios, los cuales nos ayudan a elegir un número apropiado de clústers para agrupar los datos; uno de ellos el método de error de inercia, también conocido popularmente como método del codo. (Garrido Agenjo, 2017)

La determinación de K es la siguiente:

- Si K es muy pequeño, se agruparán grupos "distintos".
- Si se elige un K muy grande, hay centros que pueden quedar huérfanos, o sin agrupación.
- El valor de K puede determinarse según alguna heurística. Por consiguiente, para lograr un K óptimo o una aproximación concluyente, se optará por realizar varias pruebas con los datos, para así al analizar los resultados, lograr estimar de mejor manera la variable K.

4.6.1.2 Método de error de inercia

El análisis de error de inercia, es una técnica ampliamente utilizada para identificar el número óptimo de clústers a implementar en un algoritmo de agrupamiento. Tras aplicar el algoritmo K-Means a un número definido de clústers, el método de error de inercia utiliza los valores de la inercia arrojados para identificar el número óptimo de clústers, siendo la inercia la suma de las distancias al cuadrado de cada objeto del clúster a su respectivo centroide.

$$Inercia = \sum_{i=0}^{N} ||x_i - \mu||^2$$

Ecuación 3 formula de la inercia

Este proceso se representar gráficamente a través de una gráfica lineal, la cual muestra una disminución de la evolución de inercia conforme aumenta el número de clústers. En la mayoría de los casos la línea representada toma una forma similar a la de un brazo y su codo, donde el codo

está representado por un cambio en la inercia de forma acentuada, indicando de esta manera el número óptimo de clúster para implementar sobre un conjunto de datos.

A continuación, se presenta la gráfica de error de inercia obtenida al aplicar K-Means sobre nuestro conjunto de datos.

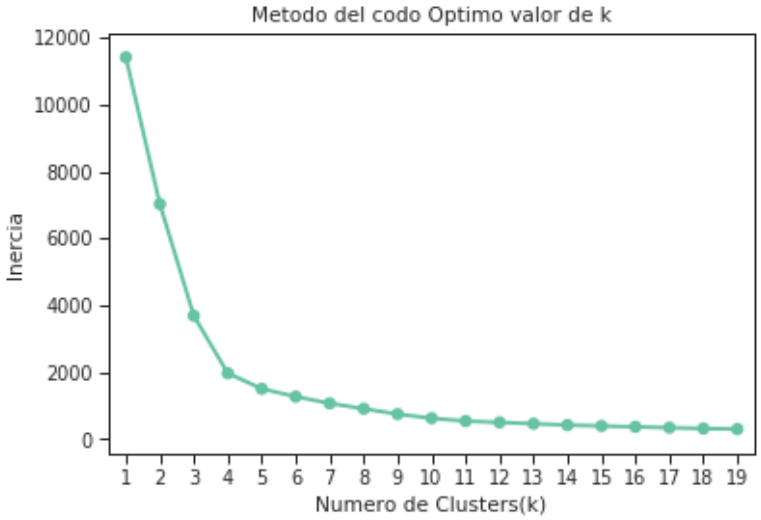


Ilustración 18 Evaluación de la inercia vs el número de clústers

Podemos observar en la ilustración 18, que los resultados de la gráfica de error de inercia son concluyentes, vemos que el codo se forma en k = 4 indicando este como el número óptimo de clústers para el entrenamiento del algoritmo. Como se mencionó anteriormente esto se deduce al observar un cambio muy mínimo en la variación de los valores inercia.

Este paso nos ha dado un punto de partida para iniciar el entrenamiento del algoritmo, posteriormente se realizará la aplicación de otras métricas de validación que aportaran mayor información acerca de la calidad del agrupamiento.

4.6.1.3 Implementación de K-Means

En este apartado se presentan los resultados obtenidos al implementar el algoritmo de K-Means haciendo uso de la librería *sklearn* en Python. El entrenamiento se realizó con implementado 4 clústers y haciendo uso de la medida la distancia Euclidiana.

A continuación, se presentan los resultados del modelo por medio de una gráfica de dispersión, donde cada clúster está representado por una forma y color en concreto.

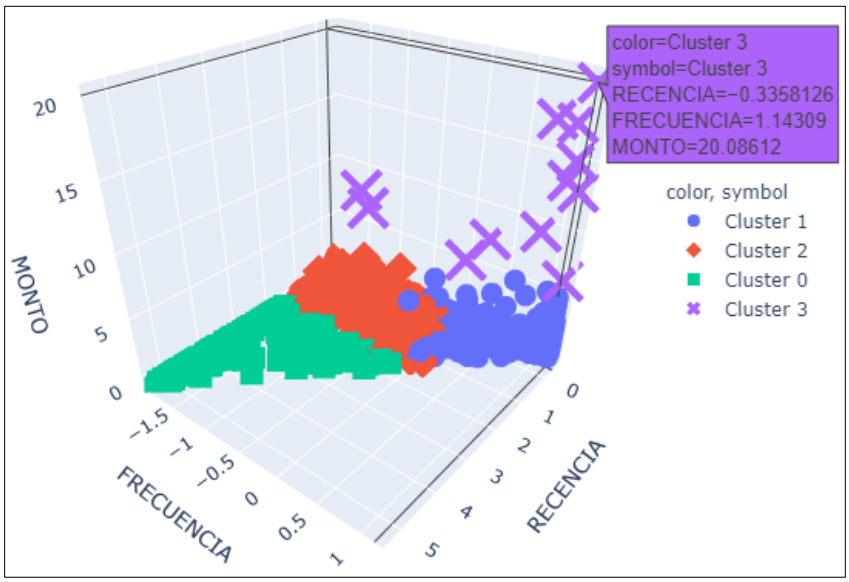


Ilustración 19 Asignación de clustering en Python

Una vez obtenidos los clústers procedemos a guardar en un archivo csv los valores de Recencia, Frecuencia, Monto y Clúster respecto a cada cliente, esto con el fin de analizar los resultados del proceso de clustering e identifican el grupo poblacional que representa cada clúster.

4.7 Evaluación del modelo

En esta sección junto con lo expuesto en el apartado 7.1 se da cumplimiento al objetivo específico número 4, en el cual nos comprometimos a: "**Evaluar el modelo RFM mediante análisis de resultados y el modelo de clustering mediante métricas de validación internas.**" Esto se realizará con la implementación de diferentes métricas de evaluación interna, las cuales permiten medir la cohesión intra-clúster y la separación inter-clústers.

4.7.1 Coeficiente de Silueta

El análisis del coeficiente de silueta se utiliza para estudiar la distancia de separación entre los grupos resultantes. El gráfico de silueta muestra una medida de qué tan cerca está cada punto de un grupo a los puntos en los grupos vecinos a través de una escala entre [-1, 1].

Los coeficientes de silueta (como se hace referencia a estos valores) cerca de +1 indican que la muestra está muy lejos de los clústers vecinos. Un valor de 0 indica que la muestra está muy cerca del límite de decisión entre dos clústers vecinos y los valores negativos indican que esas muestras podrían haberse asignado al clúster incorrecto. (Blanco & Hermida, 2016)

El coeficiente de Silhouette viene dado por la siguiente formula:

$$Silhouette_p = \frac{A-B}{max(A,B)}$$

Ecuación 4 Coeficiente de silueta

A continuación, se presenta los resultados de la puntuación del análisis de silueta aplicado a nuestro algoritmo de K-Means, obteniendo diferentes resultados según el número de clústers implementado.

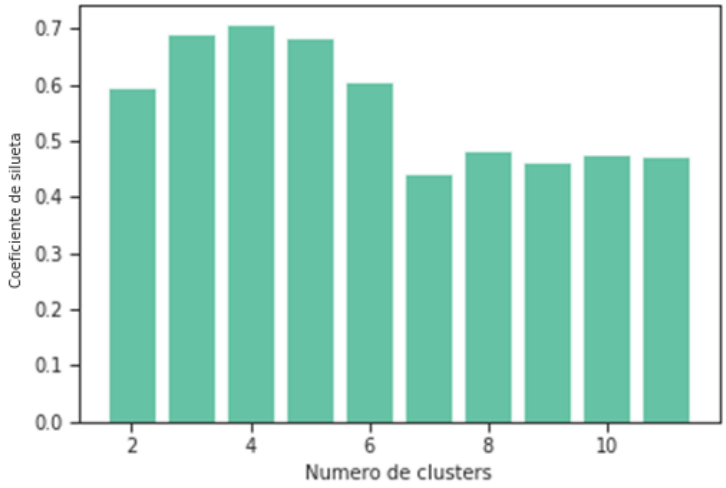


Ilustración 20 Resultado de coeficiente de silueta

En la ilustración 20 podemos constatar que el valor de k=4 presenta el índice de silueta con la puntuación más alta.

A continuación, se presenta los gráficos de siluetas correspondientes a los puntajes obtenidos en la ilustración 20, el análisis de estos gráficos nos proporciona información más completa que nos facilita la correcta interpretación de los resultados.

Nos basaremos en el puntaje de silueta, así como en la información proporcionada por los gráficos de silueta para identificar en qué condiciones se obtuvo una mejor agrupación.

a) Grafica de silueta para k = 2

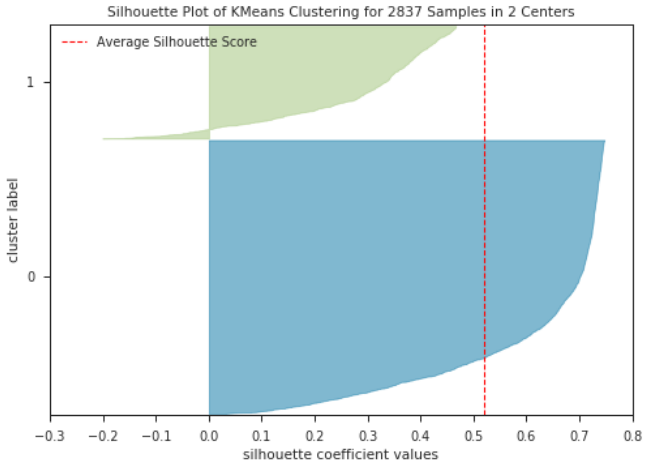


Ilustración 21 Coeficiente de silueta con 2 clústers

El agrupamiento con k = 2 obtuvo un índice de silueta de 0.57, y en la ilustración 21 podemos apreciar que una cantidad de muestras mal agrupadas en el clúster, adicionalmente también se pode conocer que el clúster 2 está por debajo de la media del índice de silueta.

b) Grafica de silueta para k =3

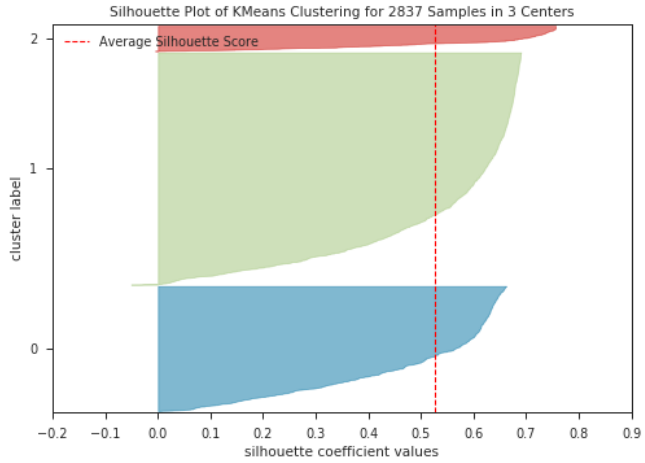


Ilustración 22 Coeficiente de silueta con 3 clústers

K=3 podría haber sido un buen resultado para entrenar K- Means si estuviese soportado por otras métricas, pero fue descartado por que tiene menor índice de silueta que el caso de k=4 y según el método del codo tampoco es la opción más eficiente.

c) Grafica de silueta para k = 4

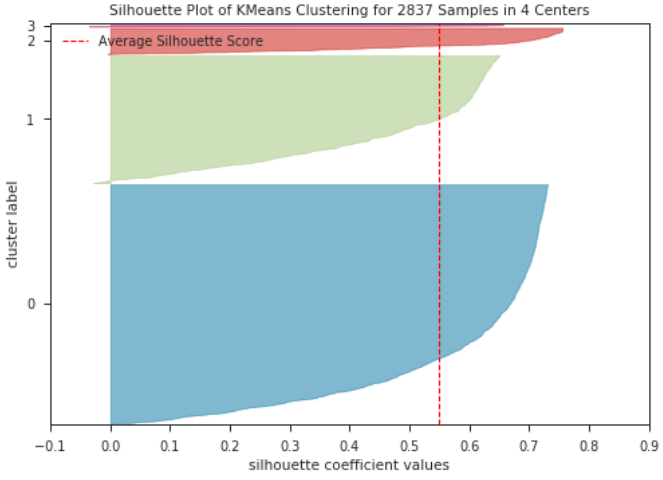


Ilustración 23 Coeficiente de silueta con 4 clústers

En la ilustración 23 se puede observar que la implementación de K-Means con k = 4 el mayor coeficiente de silueta, aunque el clúster 3 tiene considerablemente menos muestras asignadas, basados en la tipología de negocio esto no representa un factor negativo. También se observa una cantidad mínima de muestras posiblemente mal agrupadas.

d) Grafica de silueta para k = 5

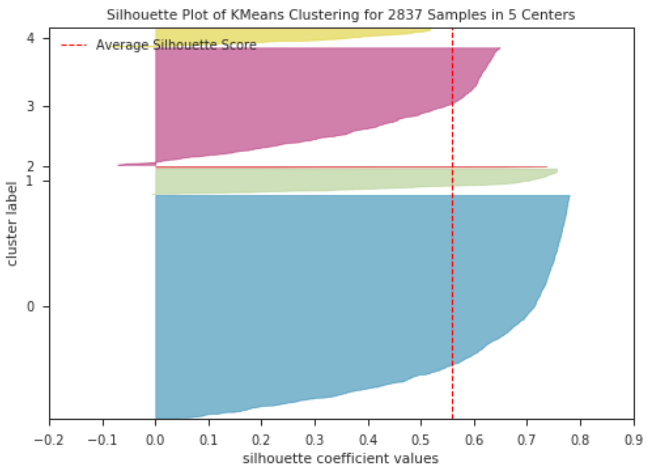


Ilustración 24 Coeficiente de silueta con 5 clústers

A partir K=5 decrece y se mantiene el coeficiente de silueta bajo, aumentan los clústers con muy pocas muestras asignadas y también aumento la cantidad de muestras asignadas al clúster incorrecto.

Podemos determinar que el entrenamiento de K-Means con k = 4, esta soportado en los resultados obtenidos con el análisis de error de inercia presentados en la ilustración 18 y con el análisis del coeficiente y grafica de silueta presentados en la ilustración 20 y 23 respectivamente. También hay que mencionar que los resultados del coeficiente de silueta son consecuentes con los resultados obtenidos en la implementación del método de error de inercia.

4.7.2 Índice Davies-Bouldin

El Índice Davies-Bouldin está definido por la siguiente formula:

$$DB = \frac{1}{k}\sum_{1} \leq i \leq k^{max_{i \neq j}(\frac{\sigma_I + \sigma_J}{||c_i - c_j||})}$$

Ecuación 5 Formula índice Davies-Bouldin

donde k es la cantidad de clústers, c_x es el centroide del clúster C_x, σ_x es la distancia promedio de todos los puntos en el clúster C_x hacia el centroide c_x, y $||c_i - c_j||$ es la distancia entre los centroides c_i y c_j . Los algoritmos que producen clústers con la mayor cohesión intra-clúster y

mayor separación entre clústers arrojan un Índice Davies-Bouldin bajo. Basados en este criterio, el modelo que arroja un índice Davies-Bouldin bajo es considerado con mejor. (Chun-Hau, 2012)

A continuación, se presentan la implementación y el índice obtenido al aplicar esta métrica de evaluación con diferente número de clústers.

```
from sklearn.cluster import KMeans
from sklearn.metrics import davies_bouldin_score

# K-Means
for i in range(1,10):
    i=i+1
    kmeans2 = KMeans(n_clusters=i, random_state=1).fit(datos_escalados)
    labels = kmeans2.labels_
    print("Cluster #"+str(i)+" puntaje indice Davies-Boulding: "+str(davies_bouldin_score(datos_escalados, labels)))
```

```
Cluster #2 puntaje indice Davies-Boulding: 0.9392976785949866
Cluster #3 puntaje indice Davies-Boulding: 0.6626189598080273
Cluster #4 puntaje indice Davies-Boulding: 0.5698636400827801
Cluster #5 puntaje indice Davies-Boulding: 0.5698692887724169
Cluster #6 puntaje indice Davies-Boulding: 0.6391608479338278
Cluster #7 puntaje indice Davies-Boulding: 0.6374786669807869
Cluster #8 puntaje indice Davies-Boulding: 0.6212450999181179
Cluster #9 puntaje indice Davies-Boulding: 0.6620653920236711
Cluster #10 puntaje indice Davies-Boulding: 0.6640588019522172
```

Ilustración 25 Resultados Índice Davies-Bouldin

Para el índice Davies-Bouldin cuanto más baja es la puntuación es mejor la agrupación. Basados en esta premisa, los resultados obtenidos en la gráfica 25 reafirman la elección de K=4 debido a que se obtiene la mejor puntuación según los parámetros establecidos por esta métrica.

4.7.3 Índice de Dunn

El índice *Dunn* es otra medida de validación interna que se obtiene de la siguiente forma:

1) Para cada *clúster* calcular la distancia entre cada una de las observaciones que lo forman y las observaciones de los otros *clústers*.
2) Seleccionar como “representante” de la distancia entre *clústers* a la menor de todas las distancias calculadas en el paso anterior (separación mínima *inter-clústers*).
3) Para cada *clúster* calcular la distancia entre las observaciones que lo forman (*intra-clúster distance*).
4) Seleccionar como “representante” de la distancia *intra-clúster* a la mayor de todas las distancias calculadas en el paso anterior (separación máxima *intra-clúster*).

Calcular el índice Dunn como:

$$Dunn\ index\ = \min_{1\le i\le c}\left\{\min_{1\le j\le c,j\ne i}\left\{\frac{\delta(X_i,X_j)}{\max_{1\le k\le c}\{\Delta(Xk)\}}\right\}\right\}$$

Ecuación 6 Índice de Dunn

Si el modelo esta conformados por clústers compactos y bien separados, el numerador es grande y el denominador pequeño, arrojaría valores altos para ***D***; por lo tanto, el objetivo es maximizar el índice **Dunn**. (Chun-Hau, 2012)

Debido a problemas de compatibilidad con la librería que implementa el índice de Dunn en Python, esta métrica fue necesario aplicarla en código R. A continuación, se presenta la implementación del índice de Dunn y el resultado obtenido para el caso de 4 clústers.

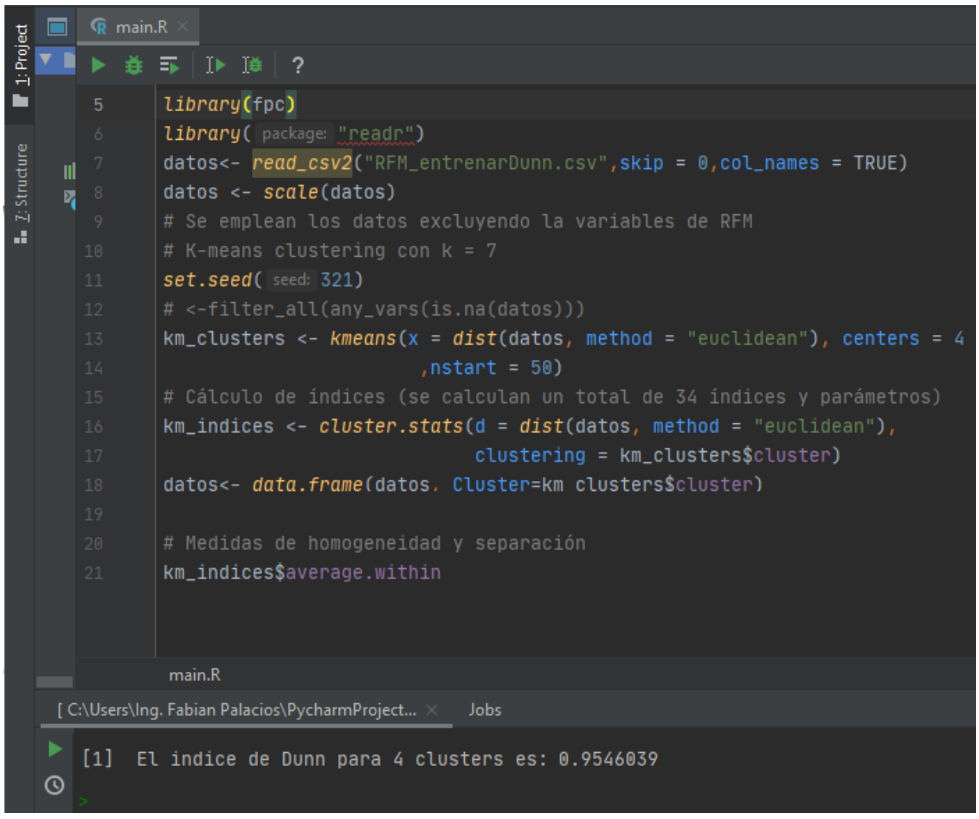


Ilustración 26 Implementación del Índice de Dunn

Si bien el índice de Dunn no tiene un umbral definido para determinar si una agrupación es buena o mala, en este caso se pretende encontrar la agrupación de muestras que maximice el valor

del índice. También se debe tener en cuenta que la interpretación de esta de este índice debe estar soportada en el análisis conjunto de otras métricas, debido a que el índice de Dunn tiende a ser mayor a medida que aumenta el número de clústers.

5. Resultados y Discusión

En este apartado se presentan los resultados de la segmentación RFM y se discutirán algunos conceptos a tener en cuenta cuando se emplea este tipo de metodologías para segmentar clientes.

5.1 Resultados del modelo RFM

En esta sección y en el apartado 4.4 se da cumplimiento al objetivo específico número 4 en el cual nos comprometemos a: "**Evaluar el modelo RFM mediante análisis de resultados y el modelo de clustering mediante métricas de validación internas.**"

En esta fase se mostrarán los resultados que se obtuvieron a la hora de realizar la implementación de análisis RFM con los datos de la empresa comercializadora de productos lácteos por medio de una tabla en la que se puede observar la influencia de cada uno de los segmentos con respecto a la compra de los productos de la empresa, a su vez también se realizó una gráfica de pastel donde se puede observar el porcentaje de los clientes que conforman cada uno de los segmentos.

Tabla 11 Resultados del análisis RFM

ANÁLISIS RFM

SEGM X VALOR	CLIENTES	% CTES	VENTAS $	% VENTAS	VTA X CLIENTE
Clientes Vip	224	8%	$ 3,416,250,880	42%	$ 15,251,120
Clientes Excelentes	383	14%	$ 1,952,670,083	24%	$ 5,098,355
Clientes Buenos	451	16%	$ 1,376,944,558	17%	$ 3,053,092
Clientes Regulares	1021	36%	$ 1,200,952,163	15%	$ 1,176,251
Clientes poco aporte	758	27%	$ 264,748,879	3%	$ 349,273

Totales	2837	100%	$ 8,211,566,563	100%	$ 4,578,616

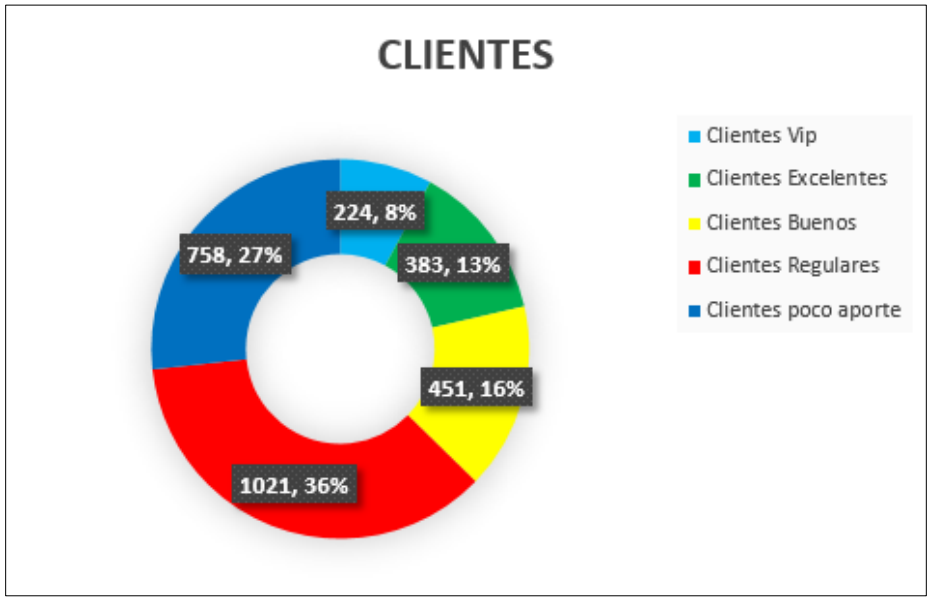


Ilustración 27 Resultados de segmentación

Con estos resultados se puede observar el comportamiento de los clientes en cada uno de los segmentos que se identificaron al implementar el modelo RFM y las personas encargadas de mercadeo y marketing de la empresa comercializadora de alimentos lácteos ya pueden utilizar estos segmentos para realizar campañas, promociones y eventos con sus clientes permitiendo mejorar la venta de sus productos.

- Según la clasificación de los clientes el segmento de los clientes de poco aporte es el 27% de la muestra, pero solo corresponde al 3 % de la venta de la empresa comercializadora de productos lácteos.
- El segmento de los clientes vip corresponde al 8% de la muestra y generan el 42 % de los ingresos a la empresa, siendo el segmento que genera mayores ingresos.
- El segmento de los clientes excelentes corresponde al 14% de la muestra y representan el 24% de la venta de los productos de la empresa comercializadora de productos lácteos.

- El segmento de los clientes buenos representa el 16% de la muestra y participan en el 17% de la venta de la empresa.
- El segmento de los clientes regulares es el 36% de la muestra al igual que el segmento de clientes de poco aporte generan el 15% de la venta de la empresa comercializadora de productos lácteos.
- Se identificaron 117 clientes con puntajes de Recencia = 1 y Frecuencia = 1, los cuales corresponden a clientes que compraron hace mucho tiempo y no volvieron a comprar, por lo que se clasifican como clientes potencialmente PERDIDOS.
- Se identificaron 276 clientes con puntajes de Recencia = 5 y Frecuencia = 1, los cuales corresponden a clientes que compraron hace muy poco tiempo y que no habían comprado antes o si lo hicieron fue hace mucho tiempo, por lo que se clasifican como clientes potencialmente nuevos.

5.2 Resultados del modelo K-Means

En este apartado se presentan los resultados del análisis del proceso de clustering con K-Means. El proceso de analizar los clústers se realizó en Excel con los valores de Recencia, Frecuencia, Monto y Clúster respecto a cada cliente, esto con el fin de analizar los resultados para identificar las características poblacionales que representa cada clúster.

A continuación, se presenta la gráfica que con los clústers y sus respectivos valores de Recencia, Frecuencia y Monto.

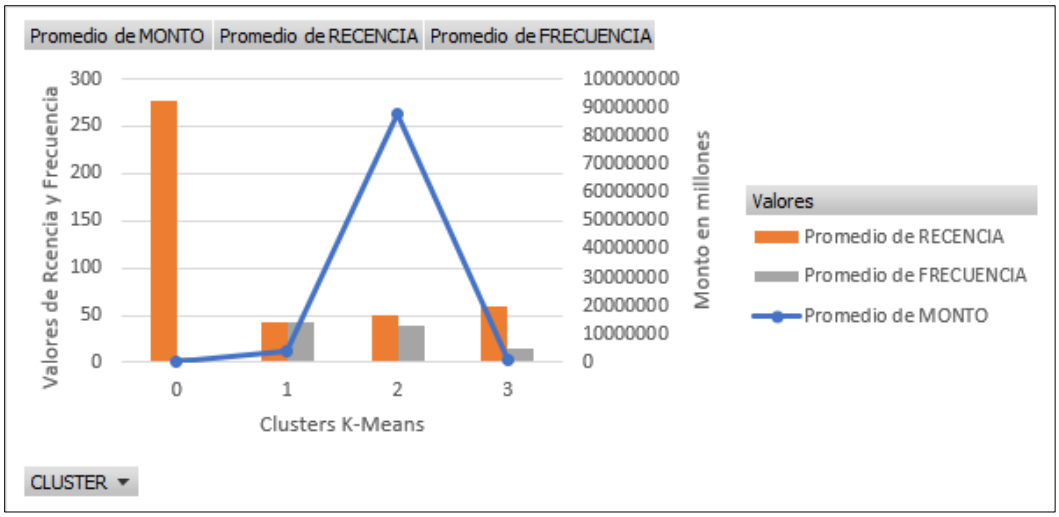


Ilustración 28 Grafica de barras con clústers

5.2.1 Caracterización de los clientes

Dando cumplimiento al objetivo general en el cual nos comprometimos a: **"Caracterizar los clientes de la empresa comercializadora de lácteos en la ciudad de Popayán, implementado unsupervised machine learning y análisis RFM."**

En esta sección se presenta la caracterización de clientes y el análisis de los resultados de la implementación del algoritmo de K-Means para segmentar los clientes de la empresa comercializadora de productos lácteos en la ciudad de Popayán.

Tabla 12 Resultados de segmentación con K-Means

SEGMENTACION DE CLIENTES CON K-MEANS								
Clúster	**Clientes**	**Porcentaje**	**Monto T**	**Monto T%**	**Prom(R)**	**Prom(F)**	**Prom(M)**	**Clasificacion**
0	187	7%	46.900.369,00 COP	1%	276	5	250.804,00 COP	Clientes Poco Aporte
1	1717	61%	6.354.262.864,00 COP	77%	42	42	3.700.793,00 COP	Clientes Buenos
2	12	0%	1.054.201.070,00 COP	13%	50	39	87.850.089,00 COP	Clientes VIP
3	921	32%	756.220.060,00 COP	9%	59	15	821.066,00 COP	Clientes Regulares
Total	**2837**	**100%**	**8.211.584.363,00 COP**	**100%**				

a) **Clúster 0:** Se identifico como **clientes de poco aporte**, debido a que son clientes que compraron hace mucho tiempo, han comprado muy pocas veces e invierten poco dinero.

b) **Clúster 1:** Como segmento son los más importantes, puesto que representan 61% de los clientes de la empresa y generan el 77% de los ingresos. A nivel de caracterización como clientes, gastan en promedio 308.000 ménsulas y son considerados como **clientes Buenos.**

c) **Clúster 2:** Conformado por un selecto grupo de 12 clientes, quienes generan el 13% de las ganancias de la empresa. A nivel de caracterización de clientes este segmento es denominado como **clientes VIP**.
d) **Clúster 3:** Este segmento está conformado por 921 clientes equivalente a poco más de la mitad de la cantidad de clientes del clúster 1, aun así; el aporte que generan es poco más de 1/8 parte del aporte generado por los clientes buenos. Adicionalmente a esto tienen peor frecuencia y recencia por lo que se clasificaron como **clientes Regulares**.

6. Conclusiones

Bajo el entorno competitivo de comercio actual, la minería de datos junto con sus algoritmos constituye un conjunto de técnicas de análisis de datos, que realmente pueden ayudar a generar estrategias que aporten valor a las empresas, e incluso a los clientes de las mismas.
Al implementar un modelo de segmentación de cliente basado bien sea en el análisis RFM o clustering con K-menas, el encargado del área de marketing debe estar en condiciones de responder entre otras, las siguientes preguntas:

- ¿Cuáles son mis mejores clientes?
- ¿Quiénes están cerca de abandonar la empresa?
- ¿Cuáles son los clientes considerados como *perdidos* a los que no debes prestar mucha atención?
- ¿En qué clientes se debe hacer un esfuerzo extra para conservarlos?
- ¿Cuáles son los clientes más leales?
- ¿Qué grupo (segmento) de clientes reaccionará de forma favorable ante la próxima campaña de publicidad o la actual?

Este conocimiento, enfocado en campañas de marketing diferenciado puede generar los siguientes beneficios para la empresa:

- Una mayor retención de clientes.
- Aumento de la tasa de respuesta.
- Aumento de la tasa de conversión
- Aumento de ingresos.

En discordancia con esto, hemos podido constatar en el transcurso del desarrollo del proyecto, que hasta las grandes empresas como la que es objeto de este estudio, prescinden de los beneficios que pueden obtener al implementar este tipo de tecnologías, ya sea por desconocimiento o simple tradicionalismo donde se desacreditan este tipo de técnicas para optar por métodos rudimentarios, poco automatizados y no necesariamente precisos. Por tal motivo, y más allá de los objetivos establecidos en este proyecto, este estudio busca de forma inherente acercar este tipo de conocimiento al contexto regional, presentando dos alternativas con características distintas a la hora de segmentar clientes, las cuales resaltan por la facilidad de su implementación y la calidad de los resultados, ofreciendo al empresario las ventajas comerciales que le aporta el conocer las diferentes características y necesidades de sus clientes.

6.1 Modelo RFM

- Teniendo en cuenta los resultados se puede observar que el modelo RFM es un modelo muy práctico para segmentar clientes cuando se cuenta únicamente con datos de las transaccionales de las ventas, Así mismo este método tiene gran adaptabilidad para ser enfocado a necesidades más específicas de la empresa.
- La implementación del modelo RFM nos permitió elegir 5 segmentos completamente diferentes para los clientes de la empresa comercializadora de productos lácteos por medio de la puntuación de RFM y estos resultados los pueden interpretar el personal de marketing para generar campañas de fidelización con sus clientes.
- La elección de la cantidad de segmentos y del tipo de población que los conforma está muy ligado a la interpretación del desarrollador del modelo con base en la tipología de negocio.

- En desarrollo del modelo RFM es muy impórtate tener en cuenta las características del negocio para determinar el peso de las variables de Frecuencia, Recencia y Monto, ya que este paso determinara en gran medida el puntaje RFM y por ende la asignación de los clientes a los diferentes segmentos. Un ejemplo claro de esto es el peso que le daría a la frecuencia un negocio sustentado en el esparcimiento familia en comparación con un banco, el valor del peso de la frecuencia no será el mismo en los dos casos.

6.1.1 Pros y contras del modelo RFM en la segmentación de clientes

Algunas de las apreciaciones que nos ha dejado la implementación del modelo RFM como método de segmentación de clientes son las siguientes:

A. Ventajas

- Es un modelo fácil de aprender y ampliamente difundido.
- Casi toda empresa cuenta con las variables necesarias para su implementación.
- Dado que los modelos de abandono son complicados de ajustar en el trabajo continuo sobre la recencia de compra es un medio excelente de reducción de abandono a medio plazo.
- Ofrece gran flexibilidad dado que podemos hacer casi cualquier definición de lo que significan los niveles de las variables para poder analizar el valor potencial y presente de un cliente. Si bien no es un modelo predictivo (en su versión estándar), sirve como "alimentador" para los modelos predictivos ya sea en su entrenamiento o en su interpretación al ser muy ligero y fáciles de leer.
- Por todo lo anterior, se presenta como un magnífico punto de partida cuando nos enfrentamos a la segmentación de clientes sin un aprendizaje previo, y con la necesidad de aplicarla inmediatamente.

B. Desventajas

- Cada cliente nuevo se debe segmentar de forma manual.

- El modelo RFM estándar solo admite las variables de recencia frecuencia y monto, por lo que si se cuenta con datos más diversos no se podrá agregar nueva información al modelo.
- El modelo ignora el comportamiento histórico de un cliente, el cual puede estar influenciado por las actividades de mercadeo realizadas por la empresa.
- Sus resultados están más influenciados por las decisiones que tome el desarrollador en cuanto a la forma de calcular los puntajes, el peso de las variables, el número de segmentos, entre otras consideraciones, por lo que se debe conocer bien las particularidades de la empresa para que así mismo el resultado del modelo RFM sea el esperado.

6.2 Algoritmo de K-Means

- A nivel técnico, la evaluación del modelo de K-means a través de las diferentes métricas de evaluación interna, arrojaron buenos resultados en cuanto a cohesión intra-clúster y separación inter-clúster, presentado resultados consecuentes entre las diferentes métricas para el número de clústers implementado.
- Los grupos obtenidos mediante la aplicación de técnicas de Minería de Datos sobre las variables RFM de los clientes de la empresa en estudio, revelaron segmentos conformados por: Clientes VIP, Clientes Buenos, Clientes Regulares y Clientes de Poco Aporte, estos resultados le permitirán a la empresa elaborar estrategias de retención hacia sus clientes, en lugar de pagar un alto costo por la atracción de nuevos clientes.
- A nivel funcional, se pudo identificar con facilidad qué tipo de características poblacionales representa cada agrupamiento realizado por K-means, dejando marcadas diferencias en cuanto a la Recencia, Frecuencia, Monto y las combinaciones de estas variables.
- Los segmentos identificados por el algoritmo de K-means respecto a la recencia, frecuencia y monto, ofrecen resultados bastante confiables a la hora de identificar el valor de los clientes para la empresa, por lo que esta técnica puede ser

implementada por cualquier empresa que maneje un registro de valor de venta y fecha junto con un mínimo de clientes.

6.2.1 Pros y Contras del modelo K-Means en la segmentación de clientes

Algunas de las apreciaciones que nos ha dejado la implementación de K-means como método de segmentación de clientes son:

A. Ventajas:

- Gracias a su autonomía, cada vez que se tenga un nuevo cliente será posible predecir a que segmento pertenece y realizarle las acciones de marketing para dicho segmento de forma rápida y eficaz sin tener repetir el proceso de segmentación.
- Es totalmente autónomo en la selección de los criterios utilizados para segmentar los clientes.
- Se obtienen resultados confiables al implementarlo con variables basadas en RFM por lo que puede ser implementado en cualquier empresa que maneje un registro tradicional de ventas.
- El software y las tecnologías necesarias para su desarrollo no conllevan necesariamente un gasto adicional.
- Al ser un modelo basado en aprendizaje de maquina no supervisado, no requiere que el operario del sistema de segmentación tenga conocimientos previos sobre su funcionamiento.

B. Desventajas:

- Necesitamos decirle al algoritmo el número de cluster (K), no puede inferir el número de clusters por sí mismo.
- El algoritmo no descarta puntos, es decir, todos los puntos pertenecen a un clúster, aunque haya una distancia abismal hacia dicho cluster.
- Es especialmente sensible con datos vacíos y outliers.

7. Trabajos futuros y recomendaciones

Para proyectos futuros relacionados con la segmentación de clientes se recomienda que:

- Realizar el análisis de las características de cada segmento conformado por el algoritmo de K-Means con el objetivo de identificar el grupo poblacional que representan.
- Evaluar la incidencia de una distribución de datos sesgada en la segmentación de clientes.
- Analizar el desempeño de otros algoritmos como **DBSCAN, AGNES, Mean Shift** para la segmentación de clientes con una limitada cantidad de variables.

Bibliografía

Alpaydin, E. (2020). *Introduction to machine learning.* massachusetts.

Alvarez,M.A.(2013,11,19).*Desarrolloweb.com*.Retrievedfrom https://desarrolloweb.com/articulos/1325.php

BBVA. (2017, septiembre 19). *La importancia de la segmentación de mercado al desplegar una estrategiaempresarial.* Retrieved from https://www.bbva.es/finanzas-vistazo/ef/empresas/segmentacion-de-mercado.html

BBVA. (2019, 11 8). Retrieved from https://www.bbva.com/es/machine-learning-que-es-y-como-funciona/

Benesty, J., Chen, J., & Huang, Y. &. (2009). Coeficiente de correlación de Pearson. In *Noise reduction in speech processing* (pp. 1- 4). Berlin: Springer, Berlin, Heidelberg.

Betanzos, C. E., Berdinas, B., Betanzos, A., & Antonio, B. (2015). Segmentacion de mercado explicable sobre datos de alta dimensiones. *II Workshop en Big Data y Analisis de Datos Escalable*, 6.

Blanco, E. J., & Hermida, S. (2016). *Algoritmos de clustering y aprendizaje.* Barcelona.

Carrizo, D., & Ortiz, C. (2016). Modelos del proceso de educción de requisitos: Un mapeo sistemático. *ingenieria y desarrollo*, 1-20.

casariego, N. (2019). *Metodologia de analisis y segmentacion de clientes usando secuencias de comportamiento.* Madrid.

Chirinos, R. (2016). Big Data para la segmentacion de mercados en redes sociales en accesorios de moda emergente. *Marketing Visionario*, 1-30.

Chun-Hau, L. (2012). *DISENO E IMPLEMENTACIÓN DE ALGORITMOS APROXIMADOS DE CLUSTERING BALANCEADO EN PSO.* SANTIAGO.

Córdoba, G. (2011, 2 10). *Análisis RFM en retail. Empezando a segmentar clientes (I)*. Retrieved from https://www.unica360.com/analisis-rfm-en-retail-empezando-a-segmentar-clientes-i

Cornejo Arce, M. L. (2019). *solucion basada en inteligencia de negocio para apoyar a la toma de decisiones en el area de ventas de una empresa en la ciudad de chiclayo.* Chiclayo.

Cuadros López, A. j., & Gonzales Caicedo, C. a. (2017). Análisis multivariado para la segmentacion de clientes basado en RFM. *revistas udistrital*, 1-11.

Cubides Proaños, C. M. (2013). *Aplicación de mineria de datos para la segmentación de clientes y desarrollo de estrategías de comunicación para la empresa DPC Studio S.A.S.* Bogota.

Ecured. (2017). Retrieved from https://www.ecured.cu/Clustering

El naga, I., & Murpy, M. j. (2015). *Machine Learning in Radiation Oncology.* New York: springer international publishing switzerland.

Evans, D. (2011). *Internet de las cosas Cómo la próxima evolución de Internet lo cambia todo.* Cisco Internet Business Solutions Grou.

Gago Utreta, R. (2017). *Uso de algoritmos de aprendizaje automático a base de datos genericos.* catalunya: España creative commons.

Garrido Agenjo, O. A. (2017). *Aplicación de técnicas de clúster al análisis de responsabilidad de los conductores en accidentes de tráfico.* Madrid.

Grabusts, P. (2011). The choice of metrics for clustering algorithms. *Letonia.ISBN 978-9984-44-071-2.*, 1-7.

IBM. (2018). Retrieved from https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html

IBM. (2018). Retrieved 3 18, 2020, from https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html

Jacome Ortega, O., & Mariella, J. O. (2014). programa para la identificacion del comportamiento del cliente de MIPYES con base en la recencia, frecuencia y magnitud de las transacciones. *alternativas*, 1-8.

León Guzmán, E. (2019). *Métricas para la validación de clustering.* Bogota.

Maimon, O., & Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook, 2nd ed. Springer Science+Business Media. *Edited by Maimon and Rokach, Tel-Aviv University, Israel. ISBN 978-0-387-09822-7.*

manzana, L. g. (2019). Retrieved from https://lagranmanzana.net/que-es-el-marketing-personalizado/

Marroquin, J., & Girosi, F. (1993). Some Extensions of the K-Means Algorithm for Image Segmentation and Pattern Classification. *MASSACHUSETTS INSTITUTE OF TECHNOLOGY*, 1-23.

Morelo Tapias, K. a. (2014). *Sistema para caracterización de perfiles de clientes de la empresa zona T.* Cartagena.

Morelo Tapias, K. A. (2014). *Sistema para caracterización de perfiles de clientes de la empresa zona T.* Cartagena.

Plazas Cardenas, L. P., & Plazas Cardenas, J. E. (2013). *Aplicación de mineria de datos para la segmentación de clientes que compran materias primasderivadas del maíz para la generación de estrategías de comunicación.* Bogota.

Raffino, R. M. (2020). *Concepto.de*. Retrieved from https://concepto.de/base-de-datos/

Rincon Boneth, J. (2016). *estudio del tanger objetivo de la empresa Madecentro Colombia S.A.S sucursal Santander.* Bucaramanga.

Rivera, J., Peña, Y., & Martinez, P. (2016). propuesta de platarfoma de procesamiento de datos para marketing directo. *universitaria y sociedad*, 65-71.

Rogers, S., & Schroedl, S. (2001). Constrained K-means Clustering with Background Knowledge. *Proceedings of the Eighteenth International Conference on Machine Learning*, 577–584.

Román Villena, J. (2016, 09 2). *sngular*. Retrieved from https://www.sngular.com/es/data-science-crisp-dm-metodologia/

Sampieri, R., Collado, C., & Lucio, a. P. (1996). *Metodologia de La Investigación.*

Sanchéz Galán, J. (2019). *economipedia*. Retrieved from https://economipedia.com/definiciones/segmentacion-de-mercado.html

Sancho Caparrini, F. (2017). *Clustering por K-means*. Retrieved from http://www.cs.us.es/~fsancho/?e=43

SAS. (2019). *Big Data*. Retrieved from https://www.sas.com/es_co/insights/big-data/what-is-big-data.html

significados. (2015). *Significados*. Retrieved from https://www.significados.com/cliente/

Sinnexus. (2018). *Sinnexus*. Retrieved from https://www.sinnexus.com/business_intelligence/datamining.aspx

Vargas Rojas, R. (2006). *Herramientas para realizar una investigación.* Cochabamba: Preparation for MSc Thesis Research.

Vergara, C. (2019). *Revista PYM*. Retrieved from https://revistapym.com.co/destacados/definicion-mercadeo-lo-que-lo-que-fue-lo-que-puede-ser/

Wagstaff, K., & Cardie, C. (2000). Clustering with Instance-level Constraints. *Proceedings of the Seventeenth International Conference on Machine Learning*, 1103-1110.

Yánez Peter, D. G. (2012). *Venta Cruzada de Productos.* Quito.