FACULTAD DE POSTGRADO
TRABAJO FINAL DE GRADUACIÓN

IMPLEMENTACIÓN DE UN MODELO DE CLUSTERIZACIÓN MEDIANTE LA SEGMENTACIÓN DE PERFIL DE CLIENTES PARA CORPORACIÓN MULTI INVERSIONES

SUSTENTADO POR:

ANA CAROLINA CARRILLO GARCÍA
EMILI GISELLE FLORES VELÁSQUEZ

PREVIA INVESTIDURA AL TÍTULO DE

MÁSTER EN
ANALÍTICA DE NEGOCIOS

TEGUCIGALPA, FRANCISCO MORAZÁN,
HONDURAS, C.A.

FEBRERO, 2024

UNIVERSIDAD TECNOLÓGICA CENTROAMERICANA
UNITEC

FACULTAD DE POSTGRADO

AUTORIDADES UNIVERSITARIAS

PRESIDENTE EJECUTIVO /

RECTORA

ROSALPINA RODRÍGUEZ

SECRETARIO GENERAL /

PRORRECTOR

ROGER MARTÍNEZ MIRALDA

VICERRECTOR ACADÉMICO NACIONAL

JAVIER ABRAHAM SALGADO LEZAMA

DIRECTORA NACIONAL DE POSTGRADO

ANA DEL CARMEN RETTALLY VARGAS

IMPLEMENTACIÓN DE UN MODELO DE CLUSTERIZACIÓN PARA LA SEGMENTACIÓN DE PERFIL DE CLIENTES PARA CORPORACIÓN MULTI INVERSIONES

TRABAJO PRESENTADO EN CUMPLIMIENTO DE LOS REQUISITOS EXIGIDOS PARA OPTAR AL TÍTULO DE

MÁSTER EN

ANALÍTICA DE NEGOCIOS

ASESORA METODOLÓGICA:

ALBA GABRIELA GARAY ROMERO

ASESOR TEMÁTICO:

JULIO ESTEBAN RAMOS MEDINA

MIEMBROS DE LA TERNA:

DANIEL ANTONIO LUNA RODRÍGUEZ
CARLOS ROBERTO AMADOR
KEVIN EDUARDO FÚNEZ FÚNEZ

DERECHOS DE AUTOR

Ana Carolina Carrillo García
Emili Giselle Flores Velásquez

FACULTAD DE POSTGRADO

IMPLEMENTACIÓN DE UN MODELO DE CLUSTERIZACIÓN MEDIANTE LA SEGMENTACIÓN DE PERFIL DE CLIENTES PARA CORPORACIÓN MULTI INVERSIONES

AUTORES:
Ana Carolina Carrillo García
Emili Giselle Flores Velásquez

Resumen

Esta investigación presenta la implementación de un modelo de clusterización para Corporación Multi Inversiones, cuyo objetivo principal es identificar segmentos o clústeres de clientes con patrones de consumo y preferencias similares. Esto permitirá crear estrategias de marketing dirigidas hacia los diferentes tipos de clientes, focalizar eficientemente los recursos y en última instancia, contribuir a la mejora de la experiencia de compra logrando así la retención de estos. Con el aumento de los datos que generan las empresas hoy en día la aplicación y generación de modelos mediante algoritmos de aprendizaje automático son herramientas que se están volviendo cada día más útiles para las empresas en la actualidad ya que su implementación genera una ventaja competitiva. Para la clasificación de los clientes se utilizó la técnica RFM en función de las variables recencia, frecuencia y valor monetario. Posteriormente se aplicó el algoritmo de clusterización no supervisado K-Means mediante la herramienta Knime. Es importante que antes de someter a los datos a la aplicación de algoritmos de aprendizaje automático se realice un preprocesamiento exhaustivo de los datos tal como limpieza, transformación y modelado de los mismos.

Palabras claves: Análisis RFM, Aprendizaje Automático, K-Means, Perfiles de Clientes, Segmentación de Clientes.

GRADUATE SCHOOL

IMPLEMENTATION OF A CLUSTERIZATION MODEL THROUGH CUSTOMER PROFILE SEGMENTATION FOR A MULTI-INVESTMENT CORPORATION

AUTORES:
ANA CAROLINA CARRILLO GARCÍA
EMILI GISELLE FLORES VELÁSQUEZ

Abstract

This research presents the implementation of a clustering model for Corporation Multi Inversions, whose main objective is to identify segments or clusters of customers with similar consumption patterns and preferences. This will allow the creation of marketing strategies targeted at different types of customers, efficiently focus resources, and contribute to improving the shopping experience, thereby achieving customer retention. With the increase in data generated by companies today, the application and generation of models using machine learning algorithms are tools that are becoming increasingly useful for businesses, as their implementation generates a competitive advantage. The RFM technique was used for customer classification based on the variables of recency, frequency, and monetary value. Subsequently, the unsupervised clustering algorithm K-Means was applied using the Knime tool. It is important that before subjecting the data to the application of machine learning algorithms, an exhaustive preprocessing of the data is carried out, such as cleaning, transformation, and modeling.

Palabras claves: Customer Profiles, Customer Segmentation, K-Means, Machine Learning, RFM Analysis.

DEDICATORIA

Dedico este proyecto primeramente a Dios quien me ha brindado fortaleza, salud y sabiduría para poder alcanzar esta meta.

A mis padres quienes me han dado la dirección con sus consejos, apoyo incondicional y motivación para seguir creciendo como una persona correcta y un buen profesional.

A mis hermanas, amigas y pareja por su apoyo, comprensión, paciencia y ánimos durante toda la etapa académica.

"El mejor modo de predecir el futuro es inventándolo"

Alan Key

Emili Giselle Flores Velásquez

DEDICATORIA

Dedico este proyecto a Dios Padre quien siempre me da la fortaleza, salud y sabiduría para poder alcanzar las metas y objetivos que me propongo en la vida.

A mi familia mis padres Julio César Carrillo y Ana Luz García, quien siempre a lo largo de toda mi vida me han dado ese apoyo incondicional en cada una de las etapas y me han enseñado que cuando se quiere lograr algo hay que esforzarse.

A mis hermanas Karen, Fany y Maryan porque cuando me veían cansada y sin fuerzas siempre me motivaron para seguir adelante.

Ellos han sido el motor de impulso y fuente de inspiración en este logro.

"El futuro pertenece a los que creen en la belleza de sus sueños"

Eleanor Roosevelt

Ana Carolina Carrillo García

AGRADECIMIENTOS

Primeramente, agradecemos a Dios por habernos dados la guía y la sabiduría necesaria para alcanzar este objetivo personal y profesional.

A la Universidad Tecnológica Centroamericana por la formación académica brindada a través de sus docentes y directivos quienes se encargaron de impartirnos el conocimiento y guía a lo largo de la carrera.

A Corporación Multi Inversiones por habernos brindado su confianza y apoyo, abriéndonos las puertas para llevar a cabo este proyecto de investigación, especialmente a Allan Ventura por haber creído en nosotras para elaborar este proyecto de investigación.

A nuestro Asesor Temático Julio Esteban Ramos por su apoyo, paciencia, recomendaciones, tiempo y conocimiento impartido durante todo el proceso de elaboración de nuestra tesis.

A nuestros Compañeros Samuel Zelaya, Oscar Zúniga y Martha Benítez por haber sido parte un excelente equipo y compañeros que contribuyeron a nuestra formación y crecimiento profesional a lo largo de la Maestría.

A nuestro Amigo Roberto Arturo Mejía por su apoyo incondicional y dedicación por su orientación en el desarrollo de esta investigación.

A nuestra Colega Stephany Osorto quien a través de su experiencia nos orientó para poder desarrollar el modelo implementado en esta investigación.

Ana Carolina Carrillo García

Emili Giselle Flores Velásquez

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO I. PLANTEAMIENTO DE LA INVESTIGACIÓN

1.1 INTRODUCCIÓN

Corporación Multi Inversiones (CMI) es una corporación agroindustrial multinacional fundada en 1920 con sede en Guatemala. CMI cuenta con presencia en más de 16 países la mayor parte ubicados en Latinoamérica, en Honduras su presencia data hace más de 30 años donde sus primeras inversiones las realizaron en el rubro de las comidas rápidas y en la industria avícola. La primera franquicia fue vendida a una empresa de capital hondureño mientras que la segunda sigue operando hasta el día de hoy en el procesamiento y comercialización de producto a base de proteína cárnica de origen avícola y porcícola (Corporación Multi Inversiones, 2023).

CMI cuenta con amplio portafolio de marcas y categorías de productos brindando atención a diferentes tipos de clientes y canales. Hoy en día CMI generan diariamente una gran cantidad de datos de sus clientes esto combinada con la rápida evolución de la tecnología ha generado el interés y la razón de ser de esta investigación por profundizar y desarrollar algoritmos para agrupar los clientes que atiende actualmente.

Este agrupamiento es conocido como modelo de clusterización, básicamente estos modelos están centrados en la categorización de los clientes, brindan a las empresas la capacidad de descubrir perfiles y patrones relacionados con los servicios o las compras. Estos hallazgos, a su vez, proporcionan las bases necesarias para diseñar estrategias que respalden la toma de decisiones más fundamentadas ya sea campañas publicitarias o estrategias de TRADE Marketing y Ventas.

CMI tiene como estrategia líder agregar valor al cliente y la utilización de la clusterización es una herramienta indispensable para este enfoque empresarial. La clusterización es una técnica que consiste en segmentar e identificar grupos de clientes en base a los datos de una empresa (da Silva, 2020). Esta metodología implica la aplicación de algoritmos de ciencia de datos para detectar patrones dentro de la base de datos, permitiendo así la formación de conjuntos distintivos de clientes que presenten similitudes entre sí. Permitiendo crear subcategorías dentro de nuestro segmento objetivo para direccionar aún mejor el mensaje que se quiere transmitir.

1.2 ANTECEDENTES DEL PROBLEMA

Se realizó una revisión de la literatura para encontrar investigaciones previas sobre la problemática planteada en la presente investigación. Cuadros López et al., (2017) abordan en su estudio la segmentación de clientes utilizando un método extendido de RFM (Recencia, Frecuencia y Monto) con nuevas variables seleccionadas a través del análisis multivariado. La microempresa ubicada en Cali, dedicada a la manufactura de productos de plástico, trataba de buscar una forma efectiva de segmentar a sus clientes para entender mejor su comportamiento y poder tomar decisiones estratégicas basadas en datos, por lo que, existía la necesidad de identificar qué clientes aportan más valor a la empresa y cómo se pueden establecer relaciones más cercanas con ellos.

Para lograr el objetivo de la investigación, se implementó el análisis RFM (Recencia, Frecuencia y Monto), analizando las transacciones de 304 clientes durante 8 meses. Además del modelo RFM tradicional, se incorporaron variables adicionales como la Ganancia (G) para obtener una visión más detallada. A través de técnicas de análisis multivariado, se validó la independencia y relevancia de estas variables. Como resultado, se identificaron cinco grupos distintos de clientes, cada uno con características y comportamientos de compra únicos (Cuadros López et al., 2017).

Por ejemplo, el grupo 1 está compuesto principalmente por clientes de las afueras de la ciudad que compran productos con alto margen de contribución, mientras que el grupo 2, aunque es el más numeroso, no muestra el mismo potencial de compra que otros grupos. A pesar de la riqueza de datos y la segmentación obtenida, se identificó que algunos clientes de alto valor no han realizado compras en más de dos meses. El análisis RFM, complementado con otras técnicas y variables, demostró ser una herramienta valiosa para la empresa, proporcionando insights clave para el diseño de estrategias futuras y la toma de decisiones basada en datos (Cuadros López et al., 2017).

En otro estudio llevado a cabo por Madariaga Fernández et al., (2022) identifica la necesidad de clasificar a los clientes de manera efectiva para optimizar la planeación y toma de decisiones en la empresa ACINOX Holguín comercializadora. Se propone una metodología multicriterio para la clasificación de clientes, utilizando algoritmos KNN como base para la planificación agregada. Esta metodología clasifica a los clientes según seis variables: fidelidad a la empresa, frecuencia de compra, activos del cliente, variedad de artículos comprados, proximidad física y horizonte temporal de compra.

Como resultado de la aplicación de esta herramienta metodológica, se diseñó una escala jerárquica de variables para clasificar a los clientes. Además, se proporcionó una clasificación general de los clientes, complementada con su clasificación según las seis variables del estudio. Esto permitió un análisis personalizado de su comportamiento. La utilización de esta herramienta metodológica en la empresa ACINOX Holguín, validó su efectividad para resolver problemas de clasificación de clientes. A partir de su aplicación, se proporcionó a los directivos de la institución un conjunto de conglomerados individuales de cada variable que respaldó la planificación agregada, facilitando la toma de decisiones y optimizando el proceso de venta desde una perspectiva general (Maradiaga Fernández et al., 2022).

En el estudio llevado a cabo por Bocanegra y Quispe (2012) identifica la problemática de las empresas actuales que buscan explotar la información de sus clientes para segmentarlos eficientemente y dirigir recursos de manera individualizada. En particular, se menciona a la caja Prymera, que necesita identificar grupos de clientes con características similares. Para abordar este desafío, se propone una técnica basada en el algoritmo K-Means en combinación con el índice de Rose Turi, seleccionada por su bajo costo computacional y facilidad de implementación. Esta combinación permite obtener la cantidad óptima de clúster o grupos.

En cuanto a la metodología, se llevó a cabo un estudio para seleccionar la técnica más adecuada para la segmentación de clientes. Se optó por el algoritmo K-Means en complementación con el índice de Rose Turi debido a su eficiencia y facilidad de implementación. Los resultados obtenidos indican que la técnica propuesta logró segmentar a los clientes en clúster con una eficacia superior en un 25% en comparación con el índice de Davies-Bouldin. Además, en términos de eficiencia en tiempo de procesamiento, la técnica propuesta demostró ser superior en un 17% (Bocanegra & Quispe, 2012).

La combinación del algoritmo K-Means con el índice de Rose Turi se presenta como una solución eficiente y efectiva para la segmentación de clientes, superando otras técnicas en términos de eficacia y tiempo de procesamiento. Esta propuesta puede ser de gran utilidad para empresas que buscan optimizar sus estrategias de marketing y recursos basándose en la segmentación de sus clientes (Bocanegra & Quispe, 2012).

En el estudio de Felipe Cálad Noreña (2015) aborda una problemática relacionada con la calidad de la información suministrada por Tiendacol S.A. para la minería de datos. Esta situación

se destaca en el contexto y caracterización del problema presentado en el documento. En cuanto a la metodología, el estudio se basa en un diagnóstico de la calidad de la información proporcionada por Tiendacol S.A., seguido de la creación de un modelo de datos que representa el resultado del procesamiento de dichos datos. Además, se describe un procedimiento específico para la aplicación de la técnica K-Means clustering (Noreña, 2015).

Los resultados del estudio se manifiestan en la formación de clústeres detallados, respaldados por el modelo final de minería de datos. En la investigación se realiza un análisis profundo de la minería de datos en el contexto de la empresa Tiendacol S.A., con un enfoque en la calidad de la información y la segmentación de clientes ayudando a resolver la problemática que presenta la empresa. (Noreña, 2015).

En el estudio realizado por Arrelucea Zapata (2020) se puede detectar que la problemática se centra en la implementación de un modelo de clusterización para segmentar el perfil del cliente en el área comercial de supermercados. La necesidad surge de la importancia de comprender y categorizar adecuadamente a los clientes para optimizar las estrategias comerciales y mejorar la experiencia del cliente.

Se empleó un enfoque cuantitativo, utilizando técnicas de clusterización para segmentar a los clientes. Se recopilaron datos relacionados con las interacciones de los clientes, y se aplicaron algoritmos de clusterización para identificar patrones y categorías dentro de la base de clientes (Arrelucea Zapata, 2020).

Los resultados indican una relación significativa entre el modelo de clusterización y los indicadores de correlación. Se identificó que las variables de Recencia, Frecuencia y Monto de venta son cruciales para identificar comportamientos de clientes en el área comercial. Además, se pudo determinar que ciertos segmentos de clientes, como los "Mejores clientes", "Clientes leales" y "Clientes más gastadores", tienen una mayor tendencia a realizar compras recurrentes. El modelo también sugiere que la precisión de la segmentación podría mejorarse con la inclusión de más variables numéricas. (Arrelucea Zapata, 2020)

El estudio concluye que la hipótesis planteada inicialmente, que sugiere una relación correcta entre el modelo de clusterización y el perfil de clientes basado en indicadores de correlación, cliente y predicción, es correcta. El modelo desarrollado permite una identificación óptima de los comportamientos de los clientes, lo que puede ser de gran utilidad para el área

comercial de supermercados. Sin embargo, se sugiere que futuras investigaciones podrían explorar la incorporación de más variables para mejorar la precisión del modelo.

1.3 DEFINICIÓN DEL PROBLEMA

1.3.1 ENUNCIADO DEL PROBLEMA

En el contexto actual de la industria del consumo masivo, la cual se caracteriza por una intensa competencia, las empresas enfrentan el desafío de personalizar sus servicios y ofertas para satisfacer las necesidades de un cliente cada vez más informado y exigente. La capacidad de una empresa para identificar y comprender las diversas necesidades y preferencias de sus clientes es fundamental para diseñar estrategias de marketing efectivas y mejorar la retención de clientes.

Corporación Multi Inversiones cuenta con una extensa cantidad de datos de clientes, pero carece de un método estructurado y científicamente validado para analizar y segmentar su base de clientes. La segmentación tradicional basada en criterios demográficos y de comportamiento no ha resultado en una diferenciación significativa que permita una focalización efectiva de las campañas. Por lo tanto, se plantea la necesidad de explorar y adoptar técnicas de análisis de datos avanzadas, como la clusterización, que podrían revelar patrones ocultos y segmentos de clientes con características homogéneas.

El presente estudio propone la implementación de un modelo de clusterización para segmentar la base de clientes de Corporación Multi Inversiones. El objetivo es desarrollar un modelo que permita identificar grupos de clientes con patrones de consumo y preferencias similares, que no son aparentes a través de métodos de segmentación convencionales. Este modelo buscará maximizar la homogeneidad dentro de los clústeres y la heterogeneidad entre ellos, utilizando algoritmos de aprendizaje automático y técnicas de minería de datos.

La investigación abordará las siguientes preguntas clave: ¿Qué metodología de clusterización es más adecuada para los datos de clientes de Corporación Multi Inversiones? ¿Cómo se pueden interpretar los segmentos de clientes resultantes para informar la toma de decisiones en estrategias de marketing?

La resolución de este problema no solo beneficiará a la empresa Corporación Multi Inversiones en términos de eficiencia y efectividad del marketing, sino que también contribuirá al

conocimiento del análisis de datos para la segmentación de clientes, ofreciendo un caso de estudio sobre la aplicación práctica de técnicas de clusterización en la industria del consumo masivo.

1.3.2 FORMULACIÓN DEL PROBLEMA

En la era de la información, las empresas acumulan grandes volúmenes de datos sobre sus clientes, pero a menudo carecen de las herramientas analíticas necesarias para transformar estos datos en información accionable para la toma de decisiones estratégicas.

Corporación Multi Inversiones ha identificado una oportunidad de mejora en su enfoque de marketing y gestión de la relación con el cliente, ya que se detectó una fuga anual de un 30% de la cartera de clientes, por lo que la empresa necesita analizar cuáles son los segmentos de clientes que se están fugando y cuáles son sus características para poder ofrecer estrategias para retenerlos. Al carecer de un enfoque sistemático y analítico para segmentar su base de clientes no está detectando a tiempo cuales son estos clientes que se están fugando.

La segmentación actual, basada en criterios simples y estáticos, no permite una diferenciación suficiente para una focalización precisa y una personalización de las ofertas. Esto ha resultado en campañas de marketing genéricas, una asignación ineficiente de recursos y, en última instancia, una experiencia del cliente que no cumple con las expectativas de personalización del mercado actual.

La implementación de un modelo de clusterización avanzado promete abordar este problema al identificar segmentos de clientes más definidos y homogéneos dentro de la base de datos de la empresa.

La investigación se centrará en la formulación y aplicación de un modelo de clusterización que pueda manejar grandes conjuntos de datos, identificar patrones no evidentes y agrupar a los clientes en segmentos basados en múltiples dimensiones de sus datos.

1.3.3 PREGUNTAS DE INVESTIGACIÓN

Pregunta general

¿Puede un modelo de clusterización ayudar a CMI a identificar y segmentar eficientemente los perfiles de sus clientes?

Preguntas especificas

a) ¿Se pueden crear clústeres de clientes que permitan perfilar los mismos en base a su comportamiento de compra?
b) ¿Cuál es el nivel de lealtad identificado en los clientes de CMI?

c) En base al modelo de clusterización ¿Se puede crear un sistema de recomendación para mejorar la experiencia de los(as) clientes actuales?

1.4 OBJETIVOS DEL PROYECTO

1.4.1 OBJETIVO GENERAL

Implementar un modelo de clusterización que identifique y segmente eficientemente los perfiles de sus clientes.

1.4.2 OBJETIVOS ESPECÍFICOS

1. Crear clústeres de clientes que permitan perfilar los mismos en base a su comportamiento de compra.
2. Identificar el nivel de lealtad en los clientes de CMI.
3. Proponer un sistema de recomendación en base al perfil del cliente para mejorar la experiencia de los clientes actuales.

1.5 JUSTIFICACIÓN

Este proyecto de investigación radica de la necesidad de la implementación de un modelo basándose en algoritmos de clusterización para la segmentación de la cartera actual de clientes del área comercial de CMI, ya que en la actualidad no existe esta subcategorización.

El algoritmo que se recomienda para esta aplicabilidad es el algoritmo de K-Means por su precisión y confiabilidad en resultados. Y se recomienda previamente aplicar la técnica de RFM.

Esto permitirá asignar los recursos de manera más eficiente, dirigiendo esfuerzos e inversiones hacia los segmentos más rentables o con mayor potencial de crecimiento, así mismo, ayudará a entender mejor a sus clientes y ofrecer experiencias personalizadas, lo que puede

traducirse en una mayor satisfacción y lealtad del cliente y en un mejor retorno de la inversión en marketing.

La implementación de un modelo de clusterización proporcionará a la empresa insights analíticos valiosos que facilitarán la toma de decisiones, adoptando técnicas avanzadas de segmentación que pueden ayudarle a posicionarse como líderes innovadores en su industria.

Además de ello la clusterización es un modelo que tiene un alcance mayor, ya que también puede incluso segmentar los clientes por los diferentes tipos de canales de atención, es decir, clusterizar los clientes por canales específicos como Tradicional, Terceros, Food Service y Moderno como una sub clusterización. Asimismo, también se podría implementar un sistema de recomendación como están utilizando muchas empresas hoy en día, es decir, utilizar algoritmos de aprendizaje automático que rápidamente identifiquen aquellos clientes que nos han dejado de comprar cada cierto tiempo o incluso recomendar un producto o promociones a determinados clientes por sus patrones de compra.

Estos procesos o actividades de recomendación si se hacen de forma manual implican mucha inversión en recursos y tiempo, sin embargo, los algoritmos de aprendizaje automático ayudan a ser más eficientes los recursos apoyando a la fuerza de venta, lo que conlleva una ventaja sobre la competencia. Así mismo, a través de este tipo de análisis lograr identificar más rápido el porcentaje de clientes que se fugan e investigar los motivos de la fuga, para así, crear estrategias de contingencia para evitar la pérdida de clientes y lograr un mayor número de fidelización y retención de clientes.

CAPITULO II. MARCO TEÓRICO

2.1 ANÁLISIS DE LA SITUACIÓN ACTUAL

2.1.1 ANÁLISIS DEL MACROENTORNO

En todo el mundo diariamente se genera, captura, almacena y analiza una avalancha de datos en las organizaciones y, por ende, se ha dado lugar a una nueva tendencia de hoy a nivel mundial llamada Big Data. Lo que a su vez también ha generado el gran problema que viene presentando la mayor parte de las empresas hoy en día, no saber qué hacer con tanta información que manejan, lo que conlleva a perdida de oportunidades y no tener un respaldo en la toma de decisiones basado en datos o estadísticas (Mamaní Rodríguez y otros, 2017).

Los grandes volúmenes de datos han ido creciendo de forma exponencial. En un estudio que hizo en el 2012 la consultora IDC llamado “El universo digital de los datos” señala que la proliferación de dispositivos como computadoras personales y teléfonos inteligentes, junto con un aumento significativo en el acceso a Internet en los mercados emergentes y el incremento en la generación de datos provenientes de máquinas, como cámaras de videovigilancia y sensores inteligentes, ha llevado a una duplicación del universo digital en los últimos años, alcanzando la asombrosa cifra de 2,8 ZB.

Uno de los aspectos más destacados del estudio es que menciona que aproximadamente el 23% del universo digital, equivalente a 643 exabytes, podría haberse utilizado para Big Data si se hubiera organizado y examinado adecuadamente. No obstante, solo el 3% de estos datos potencialmente valiosos está organizado, y un porcentaje aún menor está siendo sometido a análisis (Dell Technologies Forum, 2012).

2.1.1.1 EJEMPLOS EMPRESAS APLICANDO CLUSTERIZACIÓN

La plataforma de streaming Netflix hace uso de segmentación de clientes, en un momento se enfrentó a un problema significativo: los usuarios se sentían abrumados por la gran cantidad de contenido disponible. Para solucionarlo, implementaron el clustering, pero con un enfoque único: en lugar de agrupar a los usuarios por edad, género o ubicación, los organizaron según sus gustos y preferencias, también aplicando este criterio a los títulos disponibles.

Gracias a esta estrategia, lograron mostrar a cada usuario solo entre 40 y 50 títulos que se alineaban con sus intereses en la pantalla de inicio. Esto no solo evitó que los usuarios se sintieran

abrumados, sino que también los atrajo hacia las recomendaciones, aumentando así las posibilidades de retenerlos en la plataforma.

La clusterización resultó fundamental para Netflix, ya que tienen apenas unos segundos para convencer a un usuario de quedarse o abandonar el servicio. Por lo tanto, la personalización lograda mediante este método fue esencial en su camino hacia el éxito (Munar, Perez, 2023).

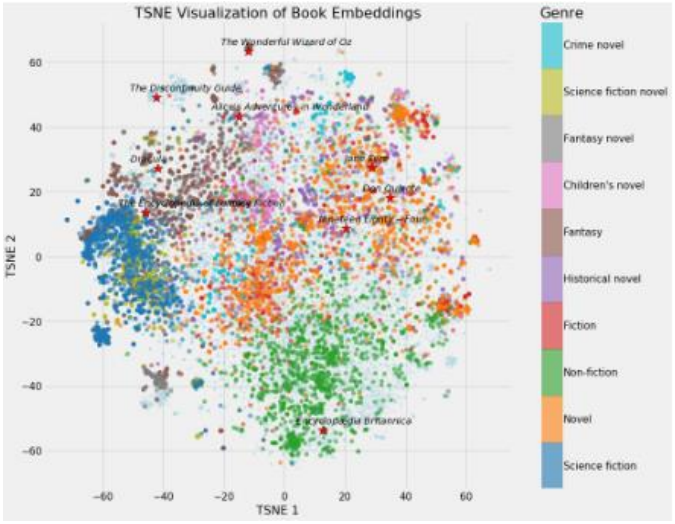


Figura 1. Visualización de Clústeres de Netflix.
Fuente: (Meyer, Sandro, 2020)

Por otra parte, Spotify utiliza el clusterización para recomendar las canciones a los usuarios, aunque lo hace de manera más general que la plataforma de Netflix.

Una empresa como ser Silicon Valley que representa el ejemplo paradigmático de un clúster empresarial. En este lugar se encuentran todas las grandes compañías de internet, y se podría decir que, para destacar en el mundo digital, es casi obligatorio haber estado en San Francisco. A pesar de la competencia entre estas empresas, la concentración geográfica crea numerosas sinergias. Desde empleados que pueden cambiar fácilmente de una empresa a otra, hasta eventos como charlas y conferencias, todo contribuye a una imagen compartida de innovación. El valor fundamental de un clúster empresarial radica precisamente en esa proximidad geográfica que facilita estas interacciones y colaboraciones.

Así mismo el 03 de junio del 2022 se presentó un trabajo de investigación acerca de la segmentación de clientes de una empresa utilizando recencia, frecuencia, monto (RFM) y métodos

de clusterización. Cuyo propósito radica en la creación de una solución informática destinada al análisis y organización de conjuntos de datos, con el objetivo de desentrañar su potencial máximo.

Los investigadores se enfocaron en un escrutinio metódico de datos relativos a transacciones de venta, con la finalidad de generar una segmentación que se fundamenta en los patrones de conducta de cada cliente, aplicando un algoritmo de agrupamiento (clusterización). Sin embargo, antes de esta etapa, se puso en práctica el método RFM para discretizar las variables, lo que simplifico la implementación de la clusterización, ya que resulta más sencillo determinar los puntos de corte en variables discretas (Mulford y otros, 2022).

Los autores de la investigación llegaron a la conclusión de que una de las ventajas más notables al emplear la segmentación de usuarios basada en el método RFM radica en que se emplea información cotidiana y esencial relativa a las operaciones de venta de la empresa. Esto conlleva a la viabilidad de su implementación sin requerir un extenso proceso de recopilación de datos. La relevancia de contar con una segmentación de la clientela reside en fortalecer la relación entre el cliente y la empresa, con la finalidad de ofrecer un servicio más personalizado, abordando de manera oportuna las necesidades de cada cliente.

También agregan que una segmentación inicial basada en variables generales también sienta las bases para una segmentación más específica, que podría incorporar variables históricas, como la edad del cliente, el tipo de producto adquirido y la periodicidad de las compras, entre otras (Mulford y otros, 2022).

2.1.2 ANÁLISIS DEL MICROENTORNO

Grupo OPSA se considera la primera empresa en Centroamérica que cuenta con su propia plataforma First Party Data, considerando que son aquellos datos que una empresa adquiere de manera directa, es decir, la información recopilada a través de sus propias fuentes como su sitio web, plataformas de redes sociales, sistema de gestión de relaciones con los clientes (CRM) y servicio de atención al cliente, abarcan una diversidad de aspectos. Estos comprenden, entre otros, las preferencias de sus clientes, sus patrones de interacción con la marca y la empresa en el ámbito virtual, así como diversos otros indicadores de comportamiento en línea (La Prensa, 2022).

Mediante la aplicación de la estrategia de agrupamiento de audiencia, se logra desvelar información previamente no identificada, una hazaña que sería inalcanzable sin esta táctica. Este tipo de datos adquiere un papel crítico en las estrategias de marketing, ya que permite el descubrimiento de nuevos segmentos de consumidores, así como la revelación de patrones hasta entonces desconocidos dentro de los segmentos preestablecidos por la empresa (La Prensa, 2022).

Grupo OPSA demuestra su entendimiento de la relevancia de la clusterización de audiencias, al disponer de su propia herramienta de big data, denominada MIDRI. Gracias a esta inversión en tecnología, Grupo OPSA ostentó el título de ser la primera empresa en Centroamérica en 2022 en contar con su propia plataforma de datos de primera mano.

Este recurso se encuentra a disposición de todos sus clientes, proporcionando una vía fundamental para la recopilación de datos esenciales y la identificación de los diversos agrupamientos de audiencia que sus clientes desean abordar (La Prensa, 2022). MIDRI presenta tres distinguidos agrupamientos de audiencias, los cuales se describen de la siguiente manera (El Heraldo, 2023):

Agrupación de Conciencia (Clusters Awareness): Esta categoría engloba a las audiencias que consumen contenido de las plataformas de OPSA en una escala más reducida, lo que las convierte en un blanco idóneo para campañas de Branding.

Agrupación de Interés (Clusters Interesting): En este caso, se refiere a las audiencias que frecuentan el contenido de las plataformas OPSA de manera más recurrente, lo que las posiciona como un conjunto propicio para campañas que combinan Branding y Rendimiento (El Heraldo, 2023).

Agrupación de Intención (Clusters Intention): Esta agrupación identifica a las audiencias que consumen contenidos y participan activamente en campañas publicitarias de manera sustancial, destacándose como el segmento ideal para campañas orientadas al rendimiento (El Heraldo, 2023).

2.2 CONCEPTUALIZACIÓN

2.2.1 ANÁLISIS EDA

El Exploratory Data Analysis (EDA), o Análisis Exploratorio de Datos, es una herramienta que se utiliza para examinar y analizar conjuntos de datos, resumiendo sus características clave, a menudo a través de técnicas de visualización (IBM Corp., 2023).

Su enfoque es descubrir errores evidentes, comprender mejor los patrones presentes en los datos, identificar valores anómalos o eventos inusuales, y revelar relaciones significativas entre distintas variables. La finalidad del EDA es asegurar que los resultados obtenidos sean válidos y relevantes para las conclusiones y metas comerciales propuestas. (IBM Corp., 2023)

2.2.2 ANÁLISIS RFM

Según Rodríguez y Gallardo (2020) es una técnica de segmentación de clientes que organiza a los clientes en diferentes niveles dentro de una jerarquía de valor. Este método implica la clasificación de los clientes en función de tres variables principales: la recencia, la frecuencia y el valor monetario.

RECENCIA (R): El tiempo transcurrido desde la última compra de un cliente. Este factor se considera esencial, ya que los clientes que han realizado compras recientemente tienen una probabilidad más alta de realizar futuras compras y responder positivamente a promociones en comparación con aquellos que hicieron compras hace mucho tiempo (Analytats, 2019).

FRECUENCIA (F): La cantidad de compras que un cliente realiza en un período específico. Los clientes que compran con frecuencia se consideran más propensos a repetir compras en el futuro (Analytats, 2019).

MONETARIO (M): El valor total de las compras realizadas por un cliente en un período definido. Los clientes que han gastado más tienen un mayor valor para la organización en comparación con aquellos que han gastado menos (Analytats, 2019).

El análisis RFM ofrece diversas posibilidades de segmentación, con once segmentos propuestos en la literatura. Sin embargo, la elección de los segmentos específicos debe basarse en las necesidades y la realidad de cada empresa (Analytats, 2019).

2.2.3 BIG DATA

Información que no puede ser procesada o analizada mediante procesos tradicionales. Big Data son "cantidades masivas de datos que se acumulan con el tiempo que son difíciles de analizar y manejar utilizando herramientas comunes de gestión de bases de datos" (Camargo Vega et al., 2014).

Es crucial reconocer una serie de etapas fundamentales en la adopción del Big Data:

- Comprender a fondo la empresa y la naturaleza de sus datos es esencial. Este inicial análisis exige una colaboración estrecha con aquellos que actualmente están inmersos en los procesos y manejan la información corporativa (Camargo Vega et al., 2014).
- La segunda etapa implica identificar los desafíos y discernir cómo la información disponible puede ser de utilidad. Es durante la exploración de los procesos operativos donde a menudo emergen las dificultades que afronta la organización o el negocio en cuestión (Camargo Vega et al., 2014).
- Es imperativo fijar expectativas realistas, o sea, establecer objetivos que sean efectivamente alcanzables; esto es vital especialmente si la implementación de una solución no genera mejoras inmediatas, obligando a la búsqueda de alternativas (Camargo Vega et al., 2014).
- Se recomienda enfáticamente que, al emprender un proyecto de Big Data, se debe proceder en simultáneo con el sistema actualmente operativo (Camargo Vega et al., 2014).
- La flexibilidad es clave al desplegar un proyecto de Big Data, tanto en lo referente a metodologías como a herramientas, dado que ambas pueden ser relativamente nuevas y susceptibles de dificultades durante su aplicación. Tales contratiempos se pueden superar mediante la investigación continua y la inversión en esta modalidad tecnológica (Camargo Vega et al., 2014).
- Es vital mantener la visión del propósito de Big Data presente; esto se debe a que el procedimiento puede ser complejo y, aunque no necesariamente tedioso, es posible que los métodos y herramientas especializados en el análisis de datos de Big Data enfrenten inconvenientes. La idea es preservar el enfoque en el objetivo último del proyecto,

evitando la desmotivación prematura (Camargo Vega et al., 2014).

2.2.4 COMPORTAMIENTO DEL CONSUMIDOR

El comportamiento del consumidor se centra en el estudio de los diversos elementos que inciden en la actitud de un individuo o colectivo al adquirir un bien o prestación. Desde una perspectiva más extensa, se refiere a la comprensión de cómo un sujeto opta por emplear sus activos a su disposición (capital, tiempo y dedicación) con el objetivo de colmar sus requerimientos (da Silva, 2022).

2.2.5 CLUSTERIZACIÓN

En el contexto de la segmentación de clientes, la clusterización se refiere al empleo de un modelo matemático con el fin de identificar conjuntos de clientes que compartan similitudes notables, mediante la detección de las mínimas discrepancias entre los distintos clientes en cada agrupación. Estos conjuntos homogéneos se conocen comúnmente como "prototipos de clientes" o simplemente "perfiles de personas" (Rodríguez H. , 2022).

La Clusterización identifica conjuntos de datos en la que los elementos dentro de un conjunto de datos exhiban semejanzas mutuas y distinciones con respecto a los elementos de otros conjuntos de datos (Fúnez, 2023).

El catedrático (Fúnez, 2023), también considera que un buen Clúster deben tener la capacidad de escalabilidad, la capacidad de manejar diversas categorías de variables, la habilidad para lidiar con datos que presenten interferencias o valores atípicos, la insensibilidad al orden de los registros y la generación de resultados altamente interpretables son cualidades esenciales.

2.2.6 CRISP-DM

CRISP-DM su abreviatura de Cross-Industry Standard Process for Data Mining, es un método establecido y eficaz para dirigir proyectos de minería de datos (Copyright IBM Corporation, 2021).

En su función como metodología, proporciona un desglose detallado de las etapas habituales en un proyecto de minería de datos, las tareas requeridas en cada etapa y una descripción de cómo estas tareas se interrelacionan. Como modelo de proceso, CRISP-DM presenta una visión general del ciclo de vida completo de la minería de datos (Copyright IBM Corporation, 2021).

2.2.7 CRM

La administración de las relaciones con los clientes, conocida por sus siglas en inglés como CRM (Customer Relationship Management), constituye un sistema esencial que posibilita el entendimiento estratégico de los consumidores y sus inclinaciones, facilitando además la gestión óptima de sus datos dentro de la entidad (Agudelo y otros, 2013).

Este instrumento se implanta con la decidida intención de propiciar un progreso eficaz en todas las actividades internas de la empresa, las cuales se ven reflejadas en la habilidad para el feedback y la evaluación del desempeño empresarial (Agudelo y otros, 2013).

La implementación de un CRM representa una táctica empresarial dirigida a consolidar una ventaja competitiva sostenida, lograda a través de la provisión de valor excepcional a los consumidores y la captura de valor por parte de la empresa, ambos procesos ocurriendo de forma concurrente (Agudelo y otros, 2013).

2.2.8 DATA MINING O MINERÍA DE DATOS

"El Data Mining es un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos que expliquen el comportamiento de estos datos" (Vallejo Ballesteros et al., 2018).

Si bien podría pensarse que la Minería de Datos es un avance tecnológico de aparición reciente, lo cierto es que su origen se remonta a la década de 1960, surgiendo en paralelo a términos como la pesca de datos o la arqueología de datos. No obstante, no fue sino hasta la década de 1980 que este campo comenzó a afianzarse (Vallejo Ballesteros et al., 2018).

El propósito de la extracción de datos radica en la habilidad de interpretar volúmenes masivos de información, los cuales, a su vez, pueden ser aprovechados para inferir conclusiones que impulsan el desarrollo y expansión corporativa, especialmente en áreas como las ventas y la lealtad del cliente (Vallejo Ballesteros et al., 2018).

Así, los datos se convierten en el instrumento esencial para deducir perspectivas y metamorfosear esta información bruta en Insights valiosos, permitiendo a las organizaciones implementar mejoras y estrategias orientadas a la consecución de sus metas (Vallejo Ballesteros et al., 2018).

2.2.9 DIFERENCIAS ENTRE CLUSTERIZACIÓN Y CLASIFICACIÓN

Existen notables diferencias entre la clusterización y la clasificación, y gran parte de esta distinción radica en el aspecto conceptual. En primer lugar, cabe señalar que la clasificación de datos se fundamenta en clases o categorías específicas y previamente definidas, a las cuales se asignan los objetos con el propósito de agruparlos. En contraste, la clusterización se dedica a la identificación de similitudes entre los distintos objetos presentes en los conjuntos de datos y a partir de las características que comparten entre sí, los agrupa en función de estas similitudes (Rodríguez H. , 2022).

Otra de las diferencias entre ambos conceptos radica en el campo de aprendizaje. Si hablamos de los procesos de clusterización, destacamos que pertenecen al aprendizaje no supervisado. Lo que quiere decir es que con los clústeres solo se dispone un grupo de datos de entrada, sin procesos de etiquetados de los cuales obtener información, sin conocer cuáles son los resultados o datos de salida. La clusterización es aplicada en iniciativas de organizaciones que desean encontrar e investigar elementos o patrones afines en sus grupos de consumidores (Rodríguez H. , 2022).

Por otro lado, la clasificación de datos se ubica en el ámbito del aprendizaje supervisado, lo que implica una dinámica opuesta a la de la clusterización. En este contexto, se adquiere información de los datos de entrada mediante la asignación de etiquetas a los elementos objeto de estudio, lo que permite conocer las posibles salidas del algoritmo (Rodríguez H. , 2022).

2.2.10 DIFERENCIAS ENTRE CLUSTERIZACIÓN Y SEGMENTACIÓN

Según Lafosse (2016), en su investigación enfatiza que la Clusterización implica la división de una fuente de datos en múltiples agrupaciones, una distinción fundamental en comparación con la segmentación. Mientras que la segmentación se centra en identificar grupos con características similares, la clusterización opera al dividir datos en grupos no predefinidos. En otras palabras, la clusterización se refiere a la identificación de grupos que son inherentemente diferentes entre sí, aunque sus elementos compartan similitudes entre ellos.

2.2.11 DISTANCIA EUCLIDIANA

En el campo de las matemáticas la distancia euclidiana, también conocida como distancia euclídea, representa la medida convencional de separación entre dos puntos en un espacio euclidiano. Su cálculo se deriva directamente del teorema de Pitágoras (Fúnez, 2023).

La evaluación de la similitud o afinidad entre ubicaciones se efectúa mediante el cálculo de la Distancia Euclidiana (DEjk) en un entorno multidimensional p de características (Aranda, 2017). La fórmula se define como:

$$DE_{jk} = \left[\sum_{i=1}^{p} W_i \left(C_j^i - C_k^i \right)^2 \right]^{1\,2}$$

Figura 2. Formula de Distancia Euclidiana.
Fuente: (Aranda, 2017)

Donde:

J: representa la localización en estudio comparada con la estación de referencia k. **P:** denota la cantidad de características consideradas en la Distancia Euclidiana.

Wi: es el factor de importancia asignado a cada característica

Cij y Cik: son los valores normalizados de la característica i para las estaciones j y k, respectivamente (Aranda, 2017).

2.2.12 ETL

Es un tipo de integración de datos que implica tres fases clave: extracción, transformación y carga. Es un método comúnmente utilizado para la creación de un almacén de datos. En este proceso, los datos son inicialmente extraídos de su fuente original, luego transformados a un formato adecuado para su almacenamiento, y finalmente almacenados en un data warehouse o en sistemas similares como bases de datos (Arboleda, 2023).

2.2.13 INTELIGENCIA DE NEGOCIOS

Resulta crucial destacar que las herramientas de Inteligencia de Negocios (BI, por sus siglas en inglés) ofrecen perspectivas retrospectivas, contemporáneas y prospectivas acerca del comportamiento del ámbito corporativo (Rivera, 2015).

Entre las aplicaciones habituales de estas tecnologías figuran la generación de reportes, la analítica en tiempo real, la minería de datos, la administración del desempeño de la organización, la detección de tendencias, el procesamiento de textos y las proyecciones futuras (Rivera, 2015).

Este campo está en constante evolución con el advenimiento del siglo actual, emergiendo conceptos innovadores como Analytics y Big Data, los cuales se desarrollarán más adelante (Rivera, 2015).

2.2.14 K-MEANS

Es una técnica frecuentemente empleada en el ámbito de la clusterización. Se trata de un algoritmo matemático que contribuye al refinamiento de la representación de los clientes, a una mejora en la capacidad predictiva y, además, se utiliza para dirigir a los consumidores con ofertas y estímulos adaptados a sus deseos, requerimientos y elecciones individuales (Rodríguez H. , 2022).

Figura 3. Fórmula algoritmo de K-Means.
Fuente: (Aranda, 2017)

2.2.15 KNIME

KNIME (Konstanz Information Miner) es una plataforma avanzada para el análisis de datos que facilita la realización de estadísticas complejas y la minería de datos. Esta herramienta permite a los usuarios examinar tendencias y hacer predicciones sobre resultados futuros. Su interfaz de trabajo visual integra funciones como el acceso y transformación de datos, análisis

exploratorios, análisis predictivos avanzados y herramientas de visualización (Chamba Jiménez, 2015).

Además, Knime ofrece opciones para crear informes personalizados basados en los datos analizados o para automatizar la integración de nuevos conocimientos en sistemas de producción existentes. KNIME es una herramienta de código abierto y se distribuye bajo una licencia específica.

KNIME, construido sobre la plataforma Eclipse y programado principalmente en Java, se caracteriza por ser una herramienta gráfica. Utiliza nodos, que encapsulan diversos algoritmos, y flechas, que simbolizan los flujos de datos, para crear y combinar procesos de manera visual e interactiva (KNIME, 2023).

Estos nodos permiten realizar varias operaciones en tablas de datos, incluyendo: Manipulación de filas y columnas, como selección de muestras, transformaciones y agrupaciones, visualización mediante gráficos como histogramas., desarrollo de modelos estadísticos y de minería de datos, incluyendo árboles de decisión, máquinas de vector soporte y regresiones.

La naturaleza abierta de KNIME permite su ampliación mediante la creación de nuevos nodos con algoritmos personalizados. También ofrece la posibilidad de integrar de forma directa y transparente herramientas como Weka, o incorporar código de R o Python/Jython (KNIME, 2023).

KNIME combina varios componentes para el aprendizaje automático y la minería de datos mediante un enfoque modular de canalización de datos (data pipelining). Su interfaz gráfica facilita la configuración rápida de nodos para el preprocesamiento de datos (ETL: extracción, transformación, carga), análisis de datos, modelado y visualización. Desde 2006, KNIME se ha utilizado en la investigación farmacéutica, pero también encuentra aplicación en otros campos como el análisis de datos de clientes en CRM, inteligencia empresarial y análisis de datos financieros (KNIME, 2023).

2.2.16 MACHING LEARNING

Es una disciplina dentro de la inteligencia artificial que utiliza algoritmos para capacitar a las computadoras para identificar patrones en grandes conjuntos de datos. Esto les permite realizar análisis predictivos de manera autónoma, sin necesidad de programación manual. (Universidad

Internacional de Valencia, 2022).

El Aprendizaje Automático (ML por sus siglas en inglés) aborda problemas de manera autónoma a través del procesamiento y análisis de datos, siendo el volumen de estos último un factor crucial para la calidad del resultado. Para desglosar y examinar la información, se emplean algoritmos capaces de producir nuevos conjuntos de datos adaptados a requisitos específicos (Rojas, 2020).

En este proceso, el ML opera un algoritmo con datos de entrada y, como consecuencia, amplifica la información pertinente al problema investigado. La meta de enriquecer la base de datos se fundamenta en una variedad de metodologías, entre ellas la regresión tanto lineal como polinómica, los árboles de decisión, las redes neuronales, la inferencia bayesiana y las cadenas de Markov. Estos enfoques facultan al ML para identificar tendencias, deducir saberes, descubrir detalles ocultos y formular pronósticos (Rojas, 2020).

El ML no equivale a una auto programación, sino que representa una autoeducación basada en datos y experiencia previa, con el fin de crear patrones y afrontar retos emergentes. Este proceso educativo integra un conjunto de técnicas, recolección de datos, teorías de análisis de información y algoritmos dedicados a la construcción de nuevos patrones o esquemas predictivos (Rojas, 2020).

2.2.17 PERFILAMIENTO DE CLIENTES

Representa la percepción que una empresa tiene sobre el cliente óptimo al que aspira dirigir sus iniciativas de marketing, con el propósito de dinamizar las transacciones. En esencia, se trata de una delineación meticulosa de aquellos posibles consumidores a quienes se pretende captar y persuadir (Málaga, 2003).

El perfilamiento de clientes implica detallar las particularidades de un individuo o un conjunto de consumidores, abarcando aspectos demográficos, ubicación geográfica y rasgos psicológicos, así como sus hábitos de compra, solvencia financiera y registros previos de compras (Moreno, 2022).

Es, esencialmente, un proceso de identificación de los atributos de aquellos individuos que son más propensos a realizar la compra de un producto o servicio y beneficiarse de este. Al precisar estos atributos de tu público meta, es posible dividir tu cartera de clientes en distintos segmentos basados en estos perfiles (Moreno, 2022).

2.2.18 SEGMENTACIÓN DE CLIENTES

También conocida como segmentación de mercado, implica la partición de una base de clientes en subconjuntos homogéneos dentro de un mercado diverso, con un enfoque en marketing. Este proceso permite a las empresas comprender a sus clientes y dirigir sus decisiones hacia grupos claramente definidos, abordando las necesidades específicas de cada uno. Por ejemplo, es factible identificar los grupos de clientes más rentables, lo que permite que la organización se centre en la retención de estos clientes (Analytats, 2019).

El proceso de segmentación de clientes capacita a las corporaciones para clasificar a sus usuarios en grupos distintos, los cuales se definen a partir de rasgos identificados en el patrón de consumo de estos y los datos recabados de sus intercambios con la entidad (Corrales, 2020).

Mediante la información recabada a través de la segmentación de la clientela, es posible discernir los atributos esenciales del consumidor prototípico, permitiendo así determinar las categorías de individuos que conformarán la audiencia meta a ser posteriormente analizada y categorizada (Corrales, 2020).

2.3 TEORÍAS DE SUSTENTO

2.3.1 BASES TEÓRICAS

Se revisó la literatura encontrando algunas teorías basadas en la segmentación de clientes, entre las cuales se puede mencionar:

2.3.1.1 TEORIA DE AFINIDAD

Según la bibliografía revisada, aborda la problemática de la segmentación de mercados y cómo encontrar conjuntos de consumidores homogéneos a través de la teoría de afinidad. Se plantea la necesidad de identificar características comunes entre los consumidores para poder dirigir las promociones de manera más efectiva y optimizar los recursos destinados a la publicidad (Lazzari, 2018).

La metodología propuesta consiste en construir una matriz borrosa de características de clientes utilizando información de bases de datos. Luego, se determina la incidencia de estas características en la posibilidad de compra de determinados artículos. Se emplean técnicas de recuperación de efectos olvidados y se realiza la composición máx.-min de las matrices para

obtener una matriz de clientes y sus posibilidades de adquirir productos. Finalmente, se aplica la teoría de afinidad para obtener la segmentación del mercado (Lazzari, 2018).

El análisis de afinidades realizado con la metodología propuesta permitió llegar a un acuerdo entre los expertos y expresar las valuaciones en un único número del intervalo [0,1]. Se destaca la importancia de reflexionar con los expertos acerca de los factores que inciden en la decisión de compra de los consumidores, ya que pequeñas variaciones en la valuación de las características pueden determinar importantes variaciones en la posibilidad de compra (Lazzari, 2018).

El estudio de segmentación de mercados mediante la aplicación de la teoría de afinidad destaca la posibilidad de aprovechar esta información para dirigir las promociones a un público más afín, optimizando la distribución de los recursos destinados a la publicidad. También se resalta la importancia de comprender los factores que inciden en la decisión de compra de los consumidores, ya que esto permite evitar la saturación publicitaria y realizar un marketing dirigido y de relación. En resumen, la segmentación del mercado a través de la teoría de afinidad ofrece una estrategia efectiva para el desarrollo de las actividades de marketing (Lazzari, 2018).

2.3.1.2 TEORIA DEL CONSUMIDOR

La concepción del consumidor reviste una relevancia crucial en la administración de entidades empresariales, independientemente de su envergadura o sector de actividad. Desde una perspectiva preliminar, esta concepción facilita la interpretación de las maneras en que los individuos buscan y adquieren tus bienes o prestaciones (LATAM, 2023).

Más aún, esta teorización permite descifrar y examinar los múltiples elementos que influyen en las decisiones de tus consumidores al optar por una marca específica. De esta manera, al emplear la concepción del consumidor, amplías tu habilidad para consolidar vínculos con posibles consumidores, con la finalidad primordial de potenciar la conversión (LATAM, 2023).

Desde esta óptica, al explorar las tendencias y decisiones del consumidor, se aspira a lograr un entendimiento integral de los aspectos, costumbres, influjos y motivaciones que pueden incidir en la fase de interacción y transacción comercial (LATAM, 2023).

Al incorporar esta teorización, directivos y conjuntos de ventas pueden instaurar estrategias comerciales enfocadas en su demografía meta, perfeccionando la operación, realzando sus métricas y eludiendo ineficiencias en tiempo y medios al atraer a un público no idóneo (LATAM, 2023).

Aplicando de manera adecuada la concepción del consumidor, se propicia que tu organización profundice en el conocimiento de su demografía y esboce el arquetipo del cliente perfecto. Esto, consecuentemente, te posiciona para proponer una combinación pertinente de bienes o servicios a dicho público, agilizando el ciclo de ventas (LATAM, 2023).

Por añadidura, esta teorización asegura que otorgues mayor consideración a consumidores pasivos, ya que te brinda las herramientas para discernir sus patrones y determinar qué gestos de tu parte podrían incentivar una adquisición subsiguiente. De igual forma, tu escuadra comercial puede efectuar seguimientos de ventas con mayor regularidad, previniendo la pérdida de posibles clientes debido a planteamientos desacertados (LATAM, 2023).

2.3.1.3 TEORÍA DEL COMPORTAMIENTO DEL CONSUMIDOR

Según Gómez Munar et al (2023) aborda cómo los individuos toman decisiones sobre qué productos o servicios a consumir basándose en la teoría del comportamiento del consumidor en la cual presenta puntos clave:

Condicionamiento Clásico de Preferencias (CCP): Se refiere a un ámbito específico de la psicología social del consumidor. En la experimentación para el CCP, se observan diferencias en comparación con el Comportamiento Clásico (CC). En el CCP, las respuestas se centran en la actitud de la persona frente a los estímulos presentados. Esta actitud considera tres componentes:

Cognoscitivo: Relacionado con el conocimiento y la comprensión.

Afectivo: Relacionado con los sentimientos y emociones.

Volitivo o Conductual: Relacionado con la acción o comportamiento.

Motivación: Según Chiavenato (2009), la motivación está relacionada con el comportamiento y el desempeño de las personas e involucra metas y objetivos. La motivación se presenta como una reacción que produce una conducta dirigida a alcanzar una meta. En términos de marketing, la motivación es una parte esencial del proceso de decisión de compra. Kotler

menciona que "Una necesidad se convierte en una motivación cuando es lo suficientemente fuerte como para llevar a una persona a la acción" (Gómez Munar et al., 2023).

Modelo de Comportamiento del Consumidor de Kotler: Este modelo explica la relación entre los procesos psicológicos y las características de los individuos. Los estímulos de mercadeo llevan al consumidor a una decisión de compra. Además, el comportamiento individual de los consumidores es influenciado por su entorno, y existen grupos de consumidores que comparten características. La segmentación, que es el proceso de identificar estos grupos, es esencial para las estrategias de mercadeo (Gómez Munar et al., 2023).

Segmentación: Es un proceso fundamental para el mercadeo que consiste en dividir el mercado en grupos específicos llamados segmentos. Estos segmentos comparten características internas, pero son diferentes entre sí. La segmentación puede basarse en variables geográficas, demográficas, psicográficas y conductuales (Gómez Munar et al., 2023).

Psicología del Consumidor: Además de la motivación, dentro de la psicología del consumidor también se encuentra la percepción y el aprendizaje.

Es evidente que el comportamiento del consumidor es un tema complejo que involucra múltiples factores y dimensiones, desde respuestas psicológicas hasta influencias externas del entorno.

2.3.1.4 TEORÍA DE LA PROBABILIDAD

El campo de la probabilidad matemática se dedica al estudio y la cuantificación de sucesos que ocurren al azar, apoyándose en la medición de la probabilidad, que es la estimación cuantitativa de la ocurrencia de un suceso, variando desde 0, para un evento imposible, hasta 1, para uno cierto.

Dentro del ámbito del aprendizaje de máquinas y el análisis de conglomerados, se emplea la teoría probabilística para construir modelos de la distribución inherente a los conjuntos de datos. Se parte de la premisa de que los datos representan instancias originadas por procesos estocásticos y que el objetivo del análisis de conglomerados es revelar el modo en que dichos datos se congregan o se dispersan en un dominio de atributos definido.

Los métodos de agrupamiento, tales como K-Means, modelos de mezclas gaussianas (GMM, por sus siglas en inglés) y el enfoque jerárquico, con frecuencia se fundamentan en conceptos de probabilidad para la creación de grupos:

K-Means: Si bien no se cataloga como un método probabilístico de forma explícita, su lógica puede verse como el intento de optimizar la probabilidad de asociación de los datos con el clúster más próximo, presuponiendo que todas las variables tienen una varianza uniforme y que los clústeres se forman alrededor de puntos centrales.

Modelos de Mezclas Gaussianas (GMM): Estos modelos se basan claramente en principios de probabilidad. Postulan que los conjuntos de datos emergen de la combinación de diversas distribuciones gaussianas, con cada una simbolizando un clúster. La probabilidad de que un punto de datos sea miembro de un clúster se calcula conforme a la probabilidad de que haya sido producido por la distribución gaussiana asociada a dicho clúster.

Agrupamiento Jerárquico: A pesar de que este método no aplica directamente la teoría probabilística para la formación de clústeres, su interpretación podría implicar que la probabilidad de unir dos clústeres es inversamente proporcional a la distancia que los separa.

En síntesis, los algoritmos de agrupamiento aplican la teoría de probabilidad para postular hipótesis respecto al agrupamiento o distribución de los datos. Dichos algoritmos asignan probabilidades a la pertenencia de puntos de datos a distintos grupos y buscan perfeccionar estas asignaciones para desentrañar la configuración subyacente de los datos. A través de la probabilidad, se mide la incertidumbre y se realizan inferencias acerca de qué conglomerados representan de manera más fidedigna la estructura subyacente de los datos.

2.3.2 METODOLOGÍAS DESAROLLADAS

En la investigación realizada por López et al, (2017), se basó en una metodología estructurada para analizar y segmentar a los clientes de una microempresa en Cali dedicada a la manufactura y comercialización de productos desechables y plásticos. Inicialmente, se realizó una preparación de los datos, revisando la información existente de los clientes para identificar posibles errores o datos irrelevantes.

Una vez depurados los datos, se calcularon las variables clásicas del modelo RFM, que son Reciente, Frecuencia y Monetario. Adicionalmente, se introdujeron variables adicionales, como "Ganancia", que refleja el valor real que las ventas aportan a la organización. Estas variables se seleccionaron y calcularon utilizando técnicas de análisis multivariable, evaluando su correlación

y redundancia con respecto al RFM tradicional. Con las variables definidas, se procedió a agrupar a los clientes en segmentos homogéneos basados en sus comportamientos de compra.

Finalmente, se analizaron los resultados de la segmentación, comparando la eficacia de la metodología propuesta con el método RFM tradicional, lo que permitió identificar áreas de mejora y proporcionar información valiosa para el diseño de estrategias de marketing específicas para cada segmento de clientes.

En la investigación realizada por Ching-Hsue Cheng, You-Shyang Chen (2009), en donde expone de manera concisa el esquema metodológico y el procedimiento delineado para la categorización del valor al cliente, la satisfacción de sus necesidades y el fortalecimiento de las relaciones cliente-empresa. En la era reciente, la minería de datos ha ascendido a un estatus prominente, no solo en la esfera académica sino también en la comercial. Actualmente, la incorporación de herramientas de minería de datos en la gestión de relaciones con clientes (CRM) es habitual, empleando técnicas como árboles de decisión (DT), redes neuronales artificiales (ANN), algoritmos genéticos (GA) y reglas de asociación (AR), que son ampliamente aplicadas en sectores como la ingeniería, la ciencia, las finanzas y los negocios para abordar cuestiones vinculadas a la clientela.

La aplicación de la minería de datos en CRM no está exenta de desafíos. Por ejemplo, en DT, una cantidad excesiva de casos puede resultar en árboles de decisión de gran tamaño, lo que reduce la precisión en la clasificación. Las ANN, por su parte, pueden implicar extensos periodos de entrenamiento, especialmente con conjuntos de datos voluminosos, y suelen requerir un enfoque de prueba y error. Los GA se caracterizan por su lenta convergencia, un tiempo de cálculo considerable y una estabilidad reducida. Por último, las AR pueden producir un exceso de reglas, lo que conlleva a redundancias (Cheng & Chen, 2009).

Ante estos retos, este estudio introduce dos metodologías: el algoritmo K-Means y la teoría de conjuntos aproximados (Rough Set Theory, RS). Se propone un novedoso procedimiento que integra el valor cuantitativo de los atributos RFM y el algoritmo K-Means dentro del marco de la teoría RS (mediante el algoritmo LEM2), para derivar reglas significativas y superar las limitaciones mencionadas (Cheng & Chen, 2009).

La metodología implementada se desglosa en las etapas siguientes:

Modelo de Investigación: El modelo propuesto en este análisis se enfoca en segmentar el valor del cliente utilizando los atributos RFM junto con el algoritmo K-Means. El modelo RFM se emplea como variable de entrada, generando un valor cuantitativo que sirve de base para el K-Means.

Procedimiento Sugerido: Se detalla el procedimiento sugerido para la clasificación del valor del cliente, que se divide en cuatro fases:

(1) Selección y preparación del conjunto de datos.

(2) Aplicación de las métricas de recencia, frecuencia y valor monetario para obtener un valor cuantitativo, que se utiliza como variable de entrada en el análisis de conglomerados, resultando en la clasificación del valor del cliente (denominado lealtad del cliente) en 3, 5 o 7 categorías, según una perspectiva subjetiva, mediante el algoritmo K-Means.

(3) División del conjunto de datos en un segmento para entrenamiento y otro para pruebas, seguido de la extracción de reglas a través de la teoría de conjuntos aproximados (algoritmo LEM2).

(4) Evaluación de los resultados experimentales, su comparación con otros métodos y la presentación de dichas comparaciones en diferentes niveles de clasificación. El procedimiento de cálculo se describe paso a paso de la siguiente manera:

Paso 1: Preprocesamiento de Datos.

Inicialmente, se selecciona el conjunto de datos para el caso de estudio empírico. Se procesan los datos para facilitar el descubrimiento de conocimiento, eliminando registros con valores ausentes o erróneos, descartando atributos redundantes y transformando los datos para su procesamiento eficaz en la segmentación del valor del cliente.

Paso 2: Segmentación del Valor del Cliente con K-Means.

Se procede a establecer una escala para los atributos R-F-M basándose en la metodología de Hughes (1994), asignando un valor cuantitativo a los atributos RFM para su uso como variables de entrada, y posteriormente se segmenta el valor del cliente con el algoritmo K-Means. Este paso se articula en dos sub-pasos detallados a continuación.

Paso 2-1: Establecimiento de la Escala de los Atributos R-F-M.

Este subproceso se divide en cinco partes, que se enumeran a continuación:

(1) Los atributos R-F-M se ponderan equitativamente (es decir, 1:1:1).

(2) Se define una escala para los tres atributos R-F-M, asignando valores de 5 a 1, que representan la contribución del cliente a los ingresos de la empresa, siendo '5' la mayor y '1' la menor. (3) Se ordenan los datos de los atributos R-F-M en orden descendente.

(4) Se dividen los atributos R-F-M en cinco segmentos iguales, cada uno representando el 20% del total, asignándoles puntuaciones de 5 a 1 en orden descendente.

2.4 MARCO LEGAL

2.4.1 LA PROTECCIÓN DE DATOS EN HONDURAS

Es evidente la magnitud del desafío que enfrenta Honduras en la promoción y protección efectiva de los derechos de protección de datos. Aunque estos derechos están formalmente reconocidos, el país carece de las herramientas, mecanismos y de una cultura social que sean suficientemente robustos para garantizar su implementación práctica y eficaz, lo cual constituye, el objetivo primordial de cualquier derecho formalizado.

La Constitución de la República garantiza en su artículo 76 el "*derecho al honor, a la intimidad personal, familiar y a la propia imagen*", de misma manera el artículo 100 establece que "*toda persona tiene derecho a la inviolabilidad y al secreto de las comunicaciones… salvo resolución judicial… los documentos personales únicamente estarán sujetos a inspección o fiscalización de la autoridad competente*." Estableciendo de esta manera una base de donde parten los derechos a la protección de datos (Zelaya, 2021).

Desde abril de 2018, la Asamblea Legislativa de la nación hondureña ha postergado el debate final, el cual es esencial para la ratificación de la propuesta normativa concerniente a la Protección de Datos Personales. La importancia de este proyecto legislativo se destaca por la ausencia de un marco legal preexistente en Honduras que ampare los datos personales, ofreciendo así definiciones y procedimientos que constituyen una novedad legislativa. No obstante, algunos de los conceptos propuestos en la legislación son susceptibles a interpretaciones equívocas o no

están suficientemente elaborados en términos conceptuales, lo que podría desembocar en la ejecución de actos arbitrarios.

2.4.2 POLÍTICA DE CIBERSEGURIDAD EN CMI

En CMI existe una política con una última actualización en noviembre 2019 relacionada con la Ciberseguridad Informática. Cuyo alcance va dirigido a todos los colaboradores de la compañía y a terceros con acceso a la red de la empresa.

En cuya política se asegura que el colaborador conozca las acciones y precauciones que son responsabilidad de este garantizar el buen uso para protección de la información, datos y equipo tecnológico de la empresa (Corporación Multi Inversiones, 2019).

Estas responsabilidades se relacionan con el control de accesos, equipos de cómputo, correo electrónico y uso del internet, con la finalidad de proteger los recursos intangibles, es decir, información y datos de la empresa (Corporación Multi Inversiones, 2019).

Así mismo existen Normativas de Uso de Equipo de Cómputo y Normativas de Gestión Documental y Archivos, las cuales tiene como objetivo, la primera, establecer el manejo adecuado de computadores laptop, desktop, monitores, teclados, mouses, servidores entre otros y en la segunda, establecer lineamientos de recepción, custodia física y digital de expedientes, asegurando la confidencialidad y resguardo de documentación.

CAPITULO III. METODOLOGÍA

3.1 CONGRUENCIA METODOLOGICA

3.1.1 MATRIZ METODOLÓGICA

IMPLEMENTACIÓN DE UN MODELO DE CLUSTERIZACIÓN MEDIANTE LA SEGMENTACIÓN DE PERFIL DE CLIENTES PARA CORPORACIÓN MULTI INVERSIONES

Problema

Fuga del 30% de clientes anual.

Preguntas

Pregunta Primaria

¿Puede un modelo de clusterización ayudar a CMI a identificar y segmentar eficientemente los perfiles de sus clientes?

Preguntas Secundarias

a) ¿Se pueden crear clústeres de clientes que permitan perfilar los mismos en base a su comportamiento de compra?

b) ¿Cuál es el nivel de lealtad identificado en los clientes de CMI?

c) En base al modelo de clusterización ¿Se puede crear un sistema de recomendación para mejorar la experiencia de los clientes actuales?

Objetivos

Objetivo General

Implementar un modelo de clusterización que identifique y segmente eficientemente los perfiles de sus clientes.

Objetivos Específicos

1. Crear clústeres de clientes que permitan perfilar los mismos en base a su comportamiento de compra.

2. Identificar el nivel de lealtad en los clientes de CMI.

3. Proponer un sistema de recomendación en base al perfil del cliente para mejorar la experiencia de los clientes actuales.

Variables Independientes

Segmentación del Perfil del Cliente

Variable Dependiente

Modelo de Clusterización

3.1.2 ESQUEMA DE VARIABLES DE ESTUDIO

Figura 4. Variables de estudio.
Fuente: Elaboración propia.

Pérez (2007), asegura que la variable independiente es la explicación de ocurrencia de otro fenómeno, además indica que esta es la variable sujeta a la manipulación por parte del científico. Así mismo Pérez (2007), indica que la variable dependiente es el resultado o efecto que debe explicarse. Por lo tanto, en este apartado se presenta el diagrama de las variables de investigación de la siguiente manera:

3.1.3 OPERACIONALIZACIÓN DE LAS VARIABLES

Tabla 1. Operacionalización de las Variables.

Variable Independiente	Definición Conceptual	Definición Operacional	Dimensiones	Indicadores	ítems
Segmentación de Perfil de Cliente	Implica la partición de una base de clientes en subconjuntos homogéneos de acuerdo con una técnica ampliamente utilizada en identificación de sus clientes más significativos.	Aplicación de la técnica de clasificación de RFM	• Recencia • Frecuencia • Monto	Segmento "Clientes Campeones"	• Recencia (4,5) • Frecuencia (4,5) • Monto (4,5)
				Segmento "Clientes Fieles"	• Recencia >=2 • Frecuencia >=3 • Monto >=3
				Segmento "Potencialmente Fieles"	• Recencia >=3 • Frecuencia <=3 • Monto <=3
				Segmento "Nuevos Clientes"	• Recencia (4,5) • Frecuencia (1,2) • Monto (2,3)
				Segmento "Necesitan Atención"	• Recencia (2,3) • Frecuencia (2,3) • Monto (2,3)
				Segmento "A Punto de Riesgo"	• Recencia (2,3) • Frecuencia <=2 • Monto >=4
				Segmento "En Riesgo"	• Recencia >=2 • Frecuencia >=1 • Monto >=1
				Segmento "Perdidos"	• Recencia >=2 • Frecuencia <=2 • Monto <=2

Fuente: Elaboración propia.

Variable Dependiente	Definición Conceptual	Definición Operacional	Dimensiones	Indicadores
Modelo de Clusterización	Es la segmentación de los datos mediante la similitud de sus comportamientos o patrones, agrupándolos con precisión de acuerdo con la configuración de los indicadores.	Aplicación del Algoritmo de K-Means	• Inicialización	• Determinación de K Clusters
			• Asignación	• Centroide de Inicialización
			• Iteración	• Alcance de número máximo de iteraciones

Fuente: Elaboración propia.

3.1.4 HIPÓTESIS

Una hipótesis constituye una suposición preliminar sobre el elemento bajo investigación, presentando una eventualidad que puede conducir al descubrimiento o construcción de un fenómeno específico. Son proposiciones provisionales acerca del tema de estudio, cuya veracidad aún está por determinarse a través de la corroboración o refutación de los hechos objeto de análisis. (Hernández Sampieri et al., 2014), Por lo tanto, las hipótesis planteadas en esta investigación son:

***H1*:** Un modelo de clusterización puede segmentar eficientemente los perfiles de clientes en la empresa CMI, permitiendo identificar grupos de clientes claramente diferenciados y patrones significativos en sus comportamientos y preferencias.

***H0*:** Un modelo de clusterización no es capaz de segmentar eficientemente los perfiles de clientes en la empresa CMI.

3.2 ENFOQUE Y MÉTODOS

La presente investigación se enmarca bajo un enfoque cuantitativo, según Hernández Sampieri et al, (2014) explica que en este enfoque la recolección de datos es indispensable para probar así la hipótesis con base en la medición numérica y el análisis estadístico, con el fin de establecer pautas de comportamiento y probar teorías. Además, se utilizará para el análisis herramientas de minería de datos en este caso la plataforma Knime.

El presente estudio es cuantitativo porque los datos obtenidos o dataset contiene una serie de variables continuas, que facilitaran la medición de la técnica en RFM a través de recencia, frecuencia y valor monetario.

El alcance la investigación es descriptiva, ya que según Hernández Sampieri et al, (2014) en los estudios descriptivos se busca especificar las propiedades, las características y los perfiles de personas, grupos, comunidades, procesos, objetos o cualquier otro fenómeno que se someta a un análisis.

En el presente estudio se va a investigar la estructura subyacente de los datos a través de un modelo de clusterización el cual agrupará los clústeres similares de clientes basándose en los indicadores predefinidos.

Esta es una investigación con un diseño no experimental, ya que se enfoca en la observación y análisis de fenómenos en su entorno natural sin manipular intencionalmente variables independientes. A diferencia de los enfoques experimentales, donde se buscan cambios deliberados, este tipo de investigación se abstiene de tales modificaciones (Agudelo Viana & Aigneren Aburto, 2008). En el diseño no experimental no se manipulará ninguna variable de manera directa ya que estas se agrupan según las similitudes del comportamiento de los clientes y es el algoritmo a través de un aprendizaje no supervisado el que se encarga de identificar los clústeres.

Este estudio se destaca por su carácter longitudinal, en donde se recopilan datos de la misma muestra repetidamente durante un periodo prolongado de tiempo (Rodríguez & Mendivelso, 2018).

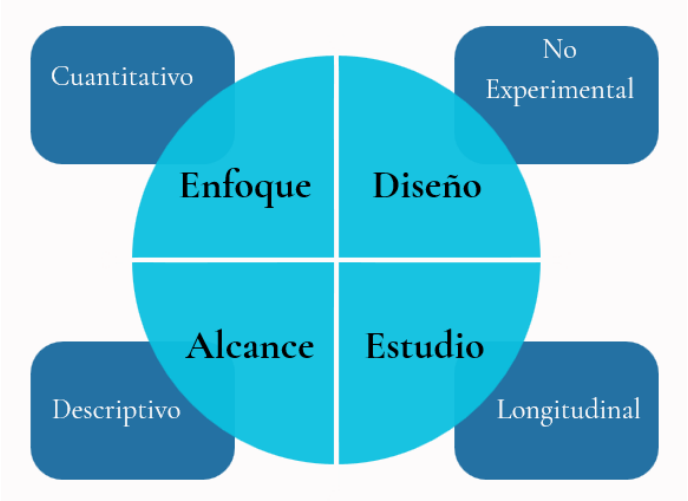


Figura 5. Enfoques y Métodos.
Fuente: Elaboración propia.

3.3 DISEÑO DE LA INVESTIGACIÓN

3.3.1 POBLACIÓN

Considerando que una población es el conjunto de todos los casos que se concuerdan con una serie de especificaciones (Hernández Sampieri et al., 2014), en esta investigación la población está compuesta por todas las facturas de los clientes de CMI Honduras en el periodo comprendido entre el **1 de noviembre del 2021 al 31 de octubre 2023** siendo un total de 1,302,114 registros.

3.3.2 MUESTRA

La población incluye la facturación de todos los canales de clientes los cuales son: Food Service, Industria, Moderno, Terceros y Tradicional, sin embargo, para este estudio se estarán excluyendo los clientes industriales ya que el producto que adquieren no es de las categorías de consumo masivo sino materia prima para preprocesamiento.

La muestra se define según Hernández-Sampieri (Hernández Sampieri et al., 2014) como "un subconjunto de elementos que pertenecen a un conjunto definido en sus características al que llamamos población". Para efectos de esta investigación la muestra está compuesta por 1,277,858 registros excluyendo 24,256 registros. En el proceso de transformación y limpieza de los datos se suprimieron los "Missing Values" o valores nulos, clientes desconocidos o facturas de prueba y facturas con valor cero. Representa una muestra no probabilística.

3.3.3 TÉCNICAS DE MUESTREO

Para este estudio la técnica de muestreo que se utilizó fue por selección intencionada o muestreo de conveniencia, el cual consiste en la selección por un procedimiento no aleatorio de una muestra cuyas características de la población sean muy similares o parecidas a la de la población objetivo (Casal & Mateu, 2003).

3.4 TÉCNICAS, INSTRUMENTOS Y PROCEDIMIENTOS APLICADOS

3.4.1 PROCEDIMIENTOS APLICADOS

En esta investigación se utilizaron técnicas y herramientas para segmentar clientes, utilizando la técnica RFM, es decir, recencia, frecuencia y monto de la base de clientes de CMI, seguidamente se aplicaron algoritmos de aprendizaje automático como K-Means. para identificar

cual es el algoritmo con mejor precisión y que se ajusta al objetivo. Para ello se estandarizaron o normalizaron los datos antes de aplicar la técnica RFM y los algoritmos, puesto de no hacer así, los diferentes rangos y escalas pueden distorsionar la importancia relativa de las variables (Cheng & Chen, 2009). Por lo tanto, se ha definido los siguientes pasos para desarrollar el proceso:

1. Obtención del dataset que fue brindado por parte de Corporación Multi Inversiones.
2. Limpieza, transformación y normalización de los datos.
3. Análisis EDA (Análisis Exploratorio de Datos).
4. Aplicación de la técnica RFM.
5. Segmentación mediante el algoritmo K-Means.
6. Análisis e interpretación de los resultados.

3.5 FUENTES DE INFORMACIÓN

La información recopilada en esta investigación procede de las siguientes fuentes de información:

3.5.1 FUENTES PRIMARIAS

La fuente primaria de esta investigación es la obtención de acceso a la información o base de datos transaccional de clientes de Corporación Multi Inversiones, los datos proporcionados necesarios para el análisis corresponden a clientes y ventas realizadas entre el periodo del 01 de noviembre del 2021 al 31 de octubre del 2023.

Los datos recopilados se dividen en dos categorías:

- **La categoría de "Cliente"** el cual comprende los siguientes campos:

- Cliente principal
- Nombre del cliente
- Macro Canal
- Tipo de negocio
- Meses de Antigüedad Primera Compra
- Años de Antigüedad Primera Compra
- Meses de Antigüedad Ultima Compra
- Forma de pago

- Municipio
- Departamento

- **La categoría de Ventas:** La cual comprende los registros diarios de compras realizadas durante el periodo del 2021 al 2023.

3.5.2 FUENTES SECUNDARIAS

El estudio no considera fuentes secundarias.

CAPITULO IV. RESULTADOS Y ANÁLISIS

En el presente capitulo se estarán presentando los resultados y análisis que se obtuvieron en base a los objetivos planteados. Para la clasificación de los consumidores, según sus patrones de adquisición se utilizó la base de datos transaccional originados en el área de ventas de Corporación Multi Inversiones.

El modelo RFM resulta ser una herramienta eficaz para la segmentación de los clientes, basándose en sus hábitos de compra. Esta categorización de los consumidores, fundamentada en criterios de comportamiento, facilitará la determinación de sus grados de fidelidad. De este modo, la empresa estará en condiciones de desarrollar diversas tácticas de marketing, específicamente orientadas hacia los distintos grupos de clientes identificados.

4.1 EXPLORACIÓN DE LOS DATOS

Se realizó un análisis EDA por sus siglas en ingles Exploratory Data Analysis, o Análisis Exploratorio de Datos la cual es una herramienta que se utiliza para examinar y analizar conjuntos de datos asegurando que los resultados obtenidos sean válidos y relevantes (IBM Corp., 2023).

Dentro de los insights relevantes se observa que el 2021 tiene un total de 104,528 transacciones tomando en cuenta que el periodo tomado fue dos meses, el promedio de transacciones mensuales de ese periodo fue de 52,264. El 2022 tiene un total de 632,984 transacciones totales en todo el año esto equivale a un promedio de 52,749 transacciones mensuales. Y el 2023 tiene un total de 540,336 transacciones tomando en cuenta que el periodo tomado fue diez meses, el promedio de transacciones mensuales equivale a 54,034.

Es decir, si comparamos año 2023 con año 2022 el promedio de transacciones mensuales a crecido un 2.4% en el 2023. Se observará mejor en la figura 6.

Figura 6. Transacciones por año.
Fuente: Elaboración propia.

Por otro lado, el 61% de las transacciones provienen del Canal Tradicional, seguido por un 29% del Canal Food Service y en tercer lugar con un 7% del Canal Terceros. Siendo esto los 3 canales que mayoritariamente tienen transacciones.

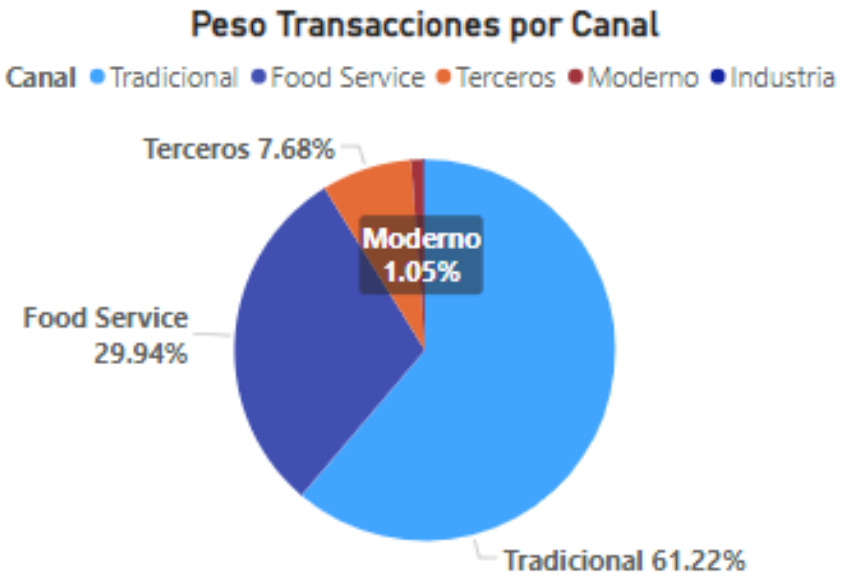


Figura 7. Peso Transacciones por canal.
Fuente: Elaboración propia.

En la figura 8 se puede observar que el canal que tiene el mayor número de transacciones es Tradicional con más de 782k transacciones en el periodo.

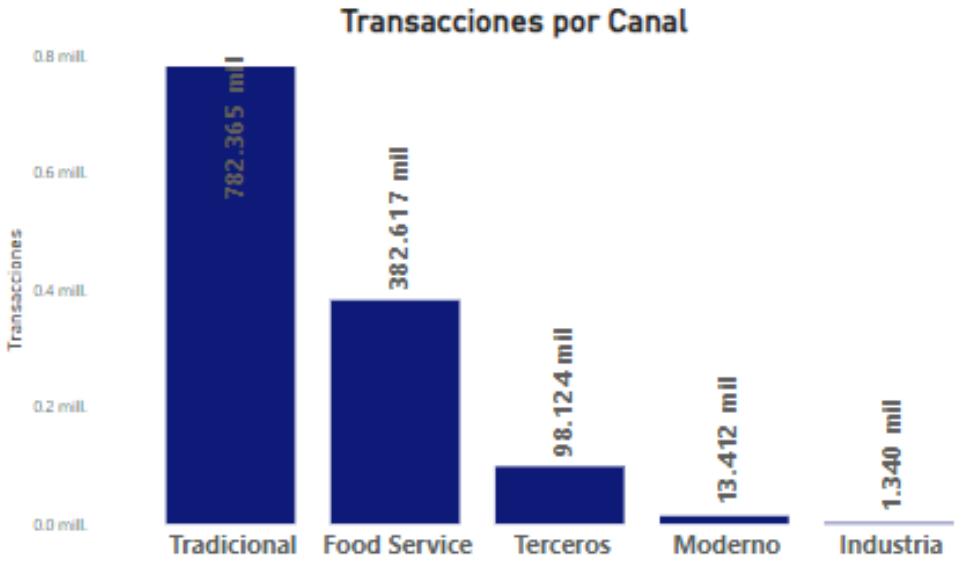


Figura 8. Transacciones por Canal.
Fuente: Elaboración propia.

Sin embargo, cuando hacemos el análisis por el monto de cada canal, se detecta que el canal que más aporta al monto de la venta total es el Canal Food Service con un 39%, seguido de Canal Terceros con un 27% y en tercer lugar Canal Tradicional con un 11%.

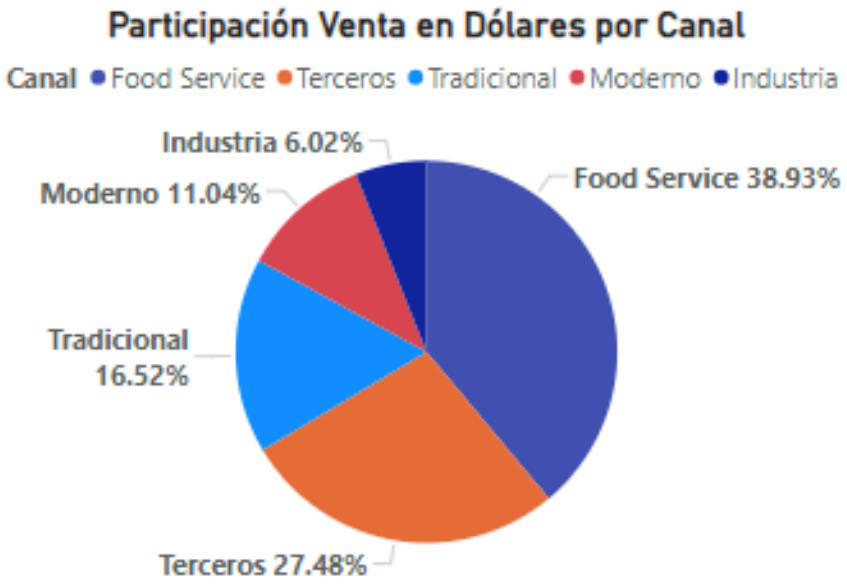


Figura 9. Peso monto de ventas por canal.
Fuente: Elaboración propia.

Otro dato relevante obtenido en base al análisis EDA y que se observa en la figura 10 es que la mayor concentración de clientes se encuentra en el Canal Tradicional con un total de 14,767 clientes, el segundo canal con más clientes es Food Sevice con 3,987 y, en tercer lugar, Canal Terceros con 423 clientes.

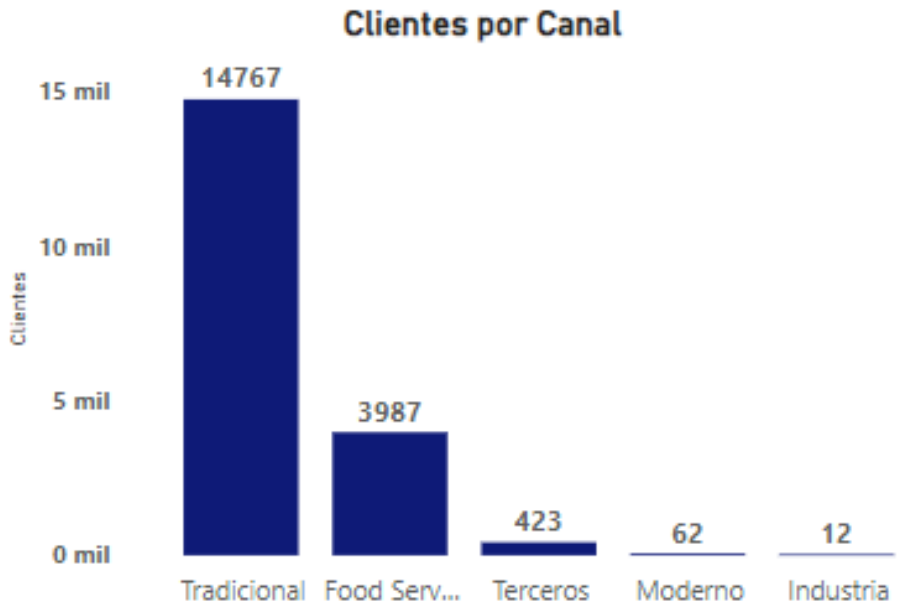


Figura 10. Clientes por macro canal.
Fuente: Elaboración propia.

Los clientes más antiguos se concentran en el Canal Industrial con un promedio de antigüedad de 93 meses, seguidamente por el Canal Moderno con un promedio de antigüedad de 87 meses desde su primera compra.

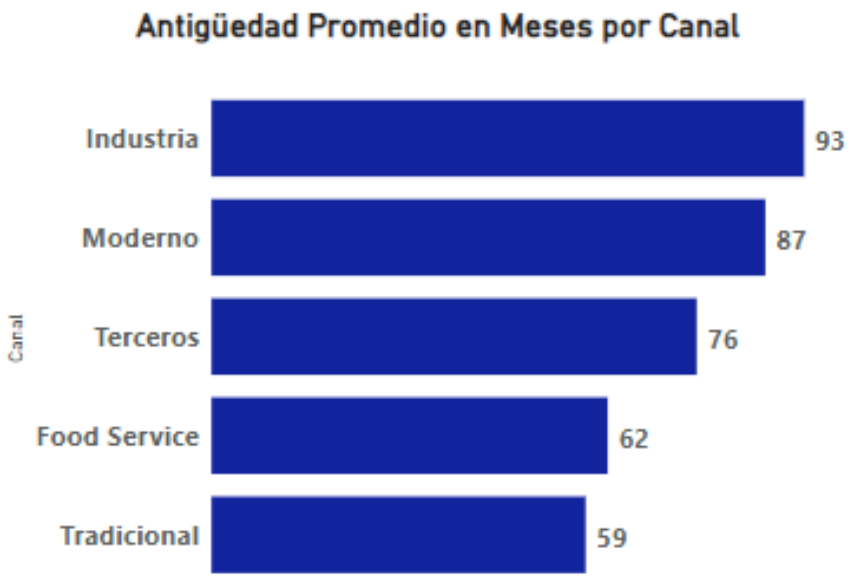


Figura 11. Antigüedad promedio de clientes por canal.
Fuente: Elaboración propia.

La recencia promedio, es decir, el canal que tiene compras más recientes es Canal Moderno con 0.3 meses (aproximadamente 7.8 días), segundo lugar, Canal Terceros con una recencia de 0.4 meses (aproximadamente 12 días) y el canal que tiene compras menos recientes en Food Service con 1.4 meses (a aproximadamente 42 días).

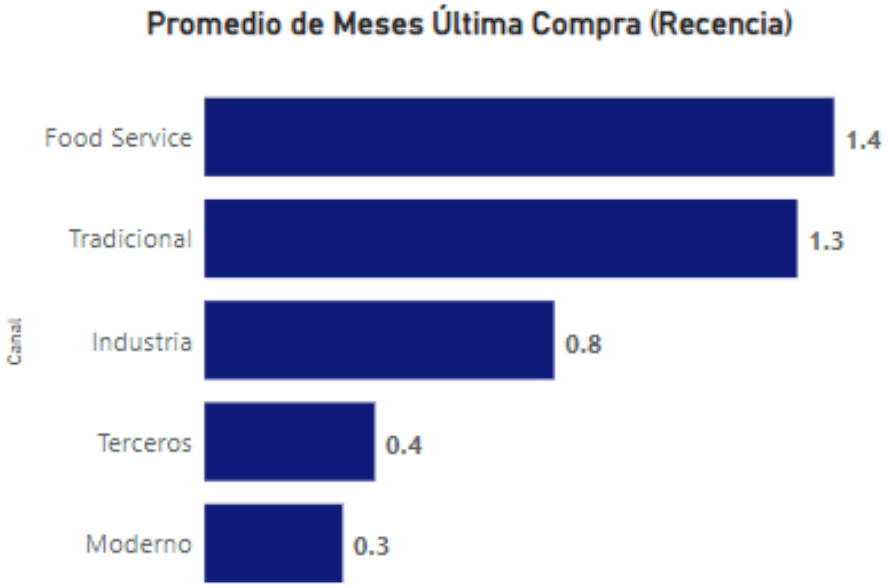


Figura 12. Recencia promedio de compra por canal
Fuente: Elaboración propia.

Para realizar el análisis de los resultados se utilizó la herramienta de Knime, tal como se observa en la siguiente figura:

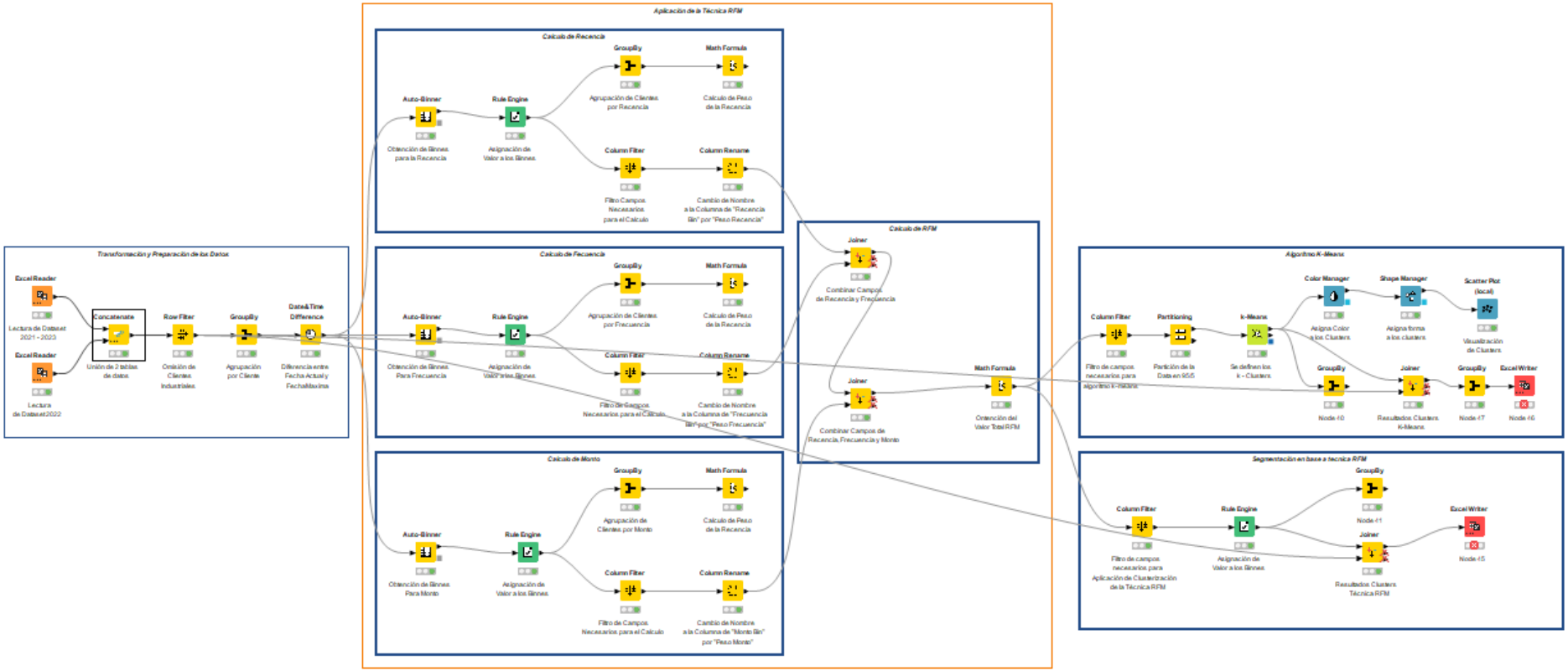


Figura 13. Modelo en Knime.
Fuente: Elaboración propia.

Mediante el cual se obtuvieron los siguientes análisis de resultados:

4.2 ANALISIS Y RESULTADOS DE "SEGMENTACIÓN EN BASE A LA TECNICA RFM"

La diferencia entre la Clusterización mediante el algoritmo K-Means y la segmentación en base a la técnica RFM es que en esta sección la agrupación de clientes se realizara de forma manual, en base a los indicadores definidos en la operacionalización de las variables. Mientras que en el algoritmo K-Means se define el número de k-clústeres en base a criterios analíticos, para la segmentación mediante la distancia euclidiana del centroide y las similitudes de comportamiento de las variables.

Si se analiza la segmentación de los clientes aplicando únicamente la técnica RFM se obtienen los siguientes resultados de acuerdo con su recencia, frecuencia y monto:

Tabla 2. Análisis RFM.

Grupo De Cliente	Recencia (días)	Frecuencia (Transacciones Mensual)	Monto Promedio (Mensual)
Campeones	1 - 6	2 - 28	$70 - $891,231
Clientes Fieles	1 - 334	1 - 24	$30 - $20,215
En riesgo	28 - 730	0 - 12	-$178 - $7,078
A punto de riesgo	7 -331	0 - 1	$70 - $2,266
Nuevos clientes	1 - 6	0 - 1	$70 - $1,942
Potencialmente fieles	1 - 27	0 - 2	-$24 - $70
Necesitan atención	28 - 334	0 - 2	$9 - $69
Prometedores	1 - 23	2 - 6	$2 - $30

Fuente: Elaboración propia.

Al momento de aplicar la técnica RFM se observa que no hay precisión en los grupos de clientes resultantes, es una técnica que se enfoca en clasificar, pero no en clisterizar. De los indicadores anteriores se representa sus resultados en de manera visual mediante el siguiente grafico de columnas apiladas:

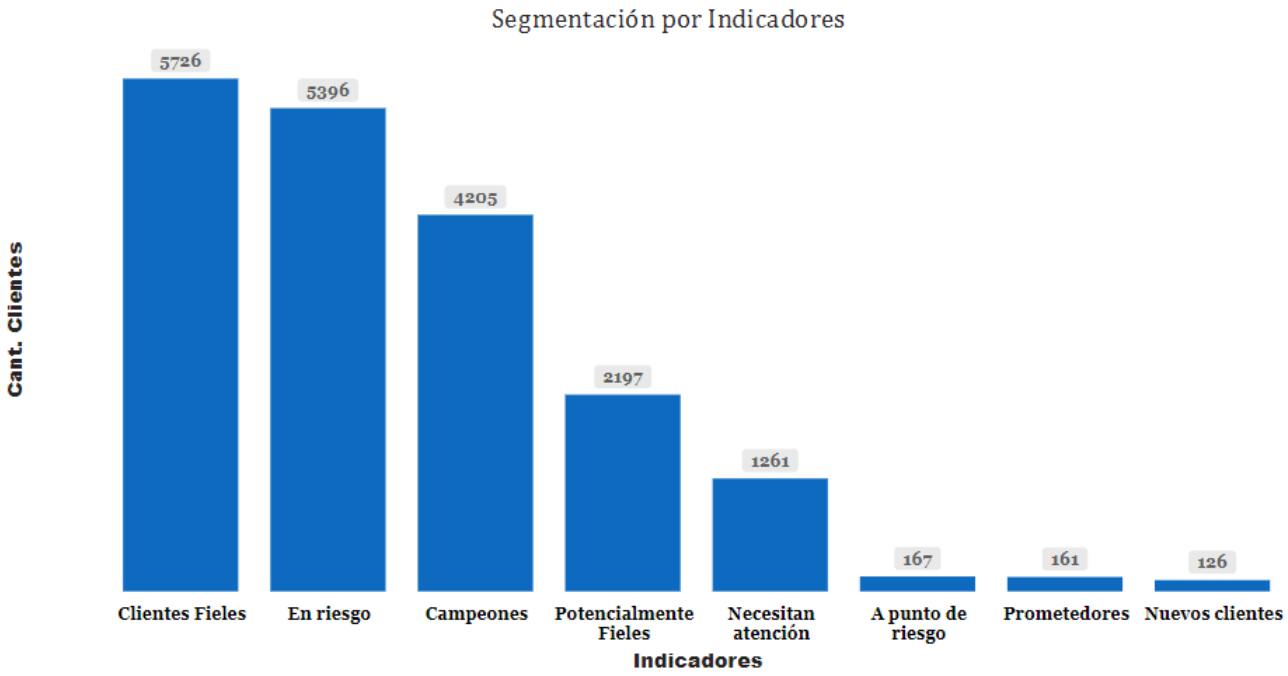


Figura 14. Segmento clientes RFM.
Fuente: Elaboración propia.

Como se puede observar en la visualización de segmentación por indicadores, mediante la evaluación de peso en la técnica RFM se agruparon los clientes, obteniendo como resultado que la cantidad de clientes fieles tienen una mayor agrupación de clientes en comparación al resto de indicadores sin embargo el top 3 viene siendo los indicadores de clientes: “Fieles”, “En Riesgo”, “Campeones”, el resto de los indicadores anda por debajo de la media.

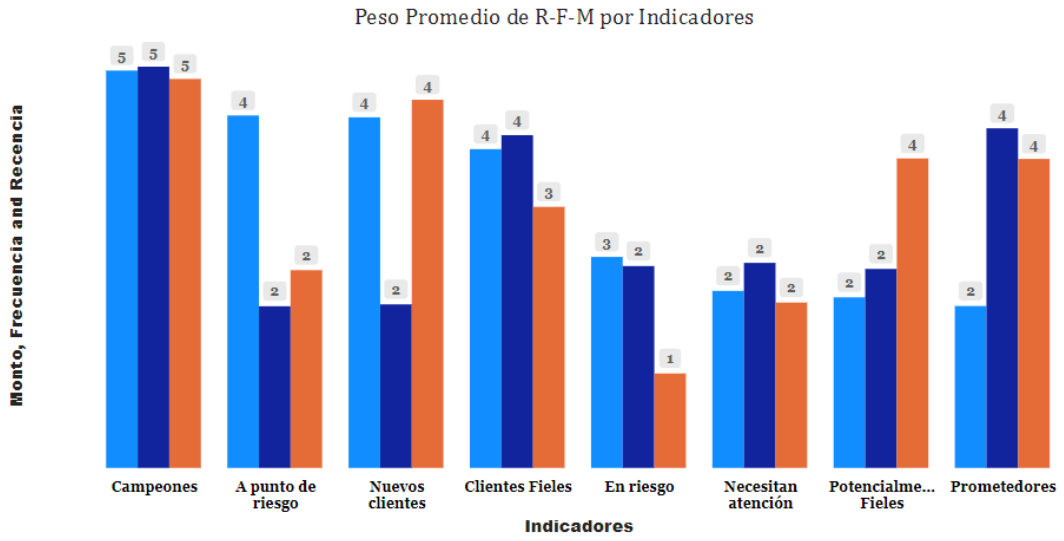


Figura 15. Indicadores por segmentos.
Fuente: Elaboración propia.

En la figura 15 se observa el resultado del peso promedio de agrupación para cada uno de los indicadores, este promedio está siendo igual al definido en las reglas de peso en los indicadores que fueron asignados anteriormente, por lo tanto, se puede validar que esta agrupación está respetando las reglas de peso.

4.3 ANALISIS Y RESULTADOS DE "CLUSTERIZACIÓN MEDIANTE ALGORITMO K-MEANS"

El algoritmo K-Means se aplicó sobre las variables RFM para realizar una clusterización basada en distancias, ya que este procedimiento ayuda a entender mejor las reglas de segmentación. Al aplicar el algoritmo K-Means se obtuvieron cinco clústeres de acuerdo con su comportamiento de compra:

- **Recencia:** Es el intervalo de tiempo entre la última fecha de compra de cada cliente y la fecha tomada al momento de realizar el proyecto de investigación (01-11-2023).
- **Frecuencia:** Corresponde número de transacciones que cada cliente ha realizado durante el periodo 01/11/2021 al 31/10/2023.
- **Monto:** Corresponde al monto total de compra por cada cliente en todas sus compras durante el periodo 01/11/2021 al 31/10/2023.

Para la aplicación de este algoritmo se eligió con 5 Clusters, pues, a través de análisis de prueba y error se detectó que a menor número de clústeres el número de agrupación tiene mayor precisión.

En la figura 17 se observa la cantidad de clientes agrupados por Clúster:

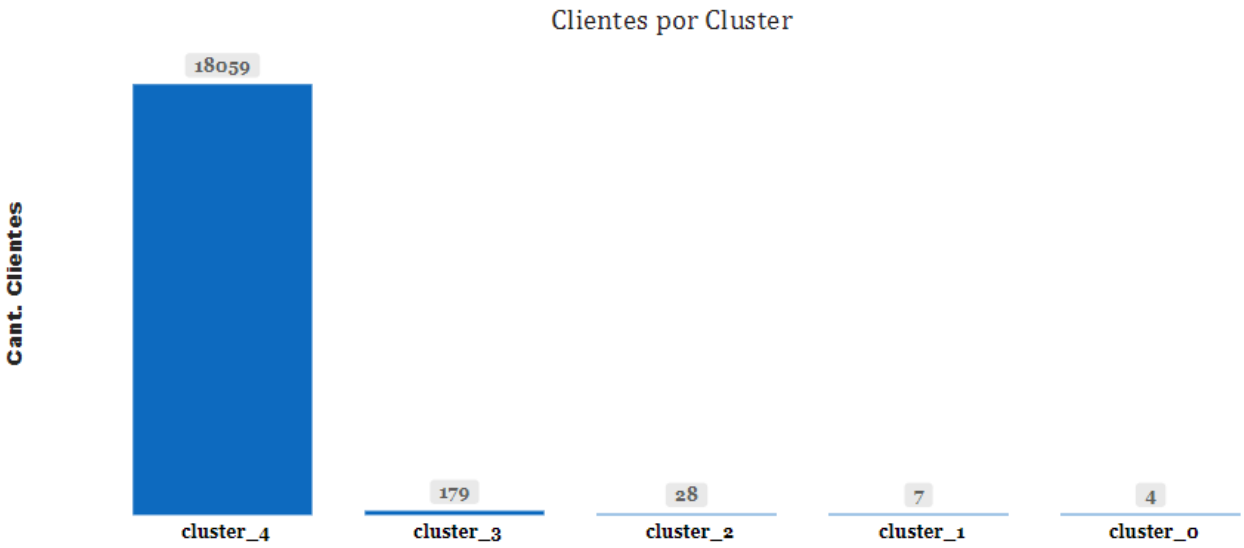


Figura 16. Clientes por clúster.
Fuente: Elaboración propia.

El Cluster_0: se segmentó por un grupo de 4 clientes con una precisión de 0.022 total de clientes. Este segmento define aquellos cuyo resultado de variables RFM devolvieron un excelente Peso de Recencia, Frecuencia y Monto. Estos clientes se pueden identificar con un nivel de lealtad "Muy Alto" para Corporación Multi Inversiones, pues fueron agrupados por las siguientes características numéricas:

Tabla 3. Características Cluster 0.

Cluster	Recencia	Frecuencia	Monto (Facturación)
Cluster_0	1 (días)	633.5 (Transac)	18.9 Mill.

Fuente: Elaboración propia.

(Es importante recordar que a menor recencia mayor "Peso Recencia")

El Cluster_1: lo segmento en un grupo de 7 clientes con una precisión de 0.038 del total de clientes. Estos podrían identificarse como aquellos clientes con un nivel de lealtad de "Alto", pues obtuvieron una calificación muy alta en el RFM, pero el algoritmo detecto que sus similitudes son distintas a los clientes del Cluster_0, tal como se puede observar en la siguiente tabla de caracterización Numérica:

Tabla 4. Características Cluster 1.

Cluster	Recencia	Frecuencia	Monto (Facturación)
Cluster_1	1.143 (días)	466.429 (Transac)	9.3 Mill.

Fuente: Elaboración propia.

El Cluster_2: este clúster es segmentado por un grupo de 31 clientes con una precisión de 0.17 del total de los clientes, estos clientes se podrían identificar con un nivel de lealtad "Medio", pues según su similitud se comprende en la siguiente tabla de caracterización:

Cluster	Recencia	Frecuencia	Monto (Facturación)
Cluster_2	1.355 (días)	507.419 (Transac)	3.5 Mill.

Fuente: Elaboración propia.

Tabla *5.* Características Cluster 2.

Pues como se observa en la tabla 5 el cluster_2 cuenta con una muy buena recencia y frecuencia, sin embargo, su monto es más bajo en comparación a los Cluster_0 y Cluster_1.

El Cluster_3: este segmento agrupa un total de 158 clientes con una precisión de 0.864 del total de clientes, cuyo puntaje RFM es intermedio tal como se observa en su caracterización numérica:

Tabla 5. Características cluster 3.

Cluster	Recencia	Frecuencia	Monto (Facturación)
Cluster_3	2.424 (días)	444.481 (Transac)	8.6 MM.

Fuente: Elaboración propia.

Definitivamente al ver los resultados estos clientes se consideran con un nivel de lealtad "Bajo", pues, aunque su Frecuencia es Alta, su Recencia y Monto son mucho más bajo en consideración a los clústeres anteriores.

El Cluster_4: este cluster segmenta una agrupación total de 18,088 con una precisión de 88.906 del total de los clientes. Estos se podrían identificar con un nivel de lealtad "Muy Bajo", pues en este segmento se agrupan el resto de los clientes cuyas similitudes se identifican conjuntamente según su puntuación de RFM, tal como se puede observar en la tabla de caracterización:

Tabla 6. Características cluster 4.

Cluster	Recencia	Frecuencia	Monto (Facturación)
Cluster_4	154.869 (días)	62.083 (Transac)	8.5 Mil.

Fuente: Elaboración propia.

Según sus características los clientes segmentados en el Cluster_4, son aquellos clientes cuya recencia puede rondar a más de 154.869 días sin una compra y su frecuencia y monto son muy bajos en comparación a los clústeres anteriores por lo tanto se identifican con un nivel de lealtad "Muy Bajo".

A continuación, se representa el modelo de clusterización mediante el algoritmo K-Means en un diagrama de dispersión para identificar el grado de distanciamiento entre 2 variables, en este caso la "Frecuencia y la Recencia", analizando su relación y que tanto se afectan o independientes pueden ser la una de la otra.

Parametrización

Eje X: Frecuencia

Eje Y: Recencia

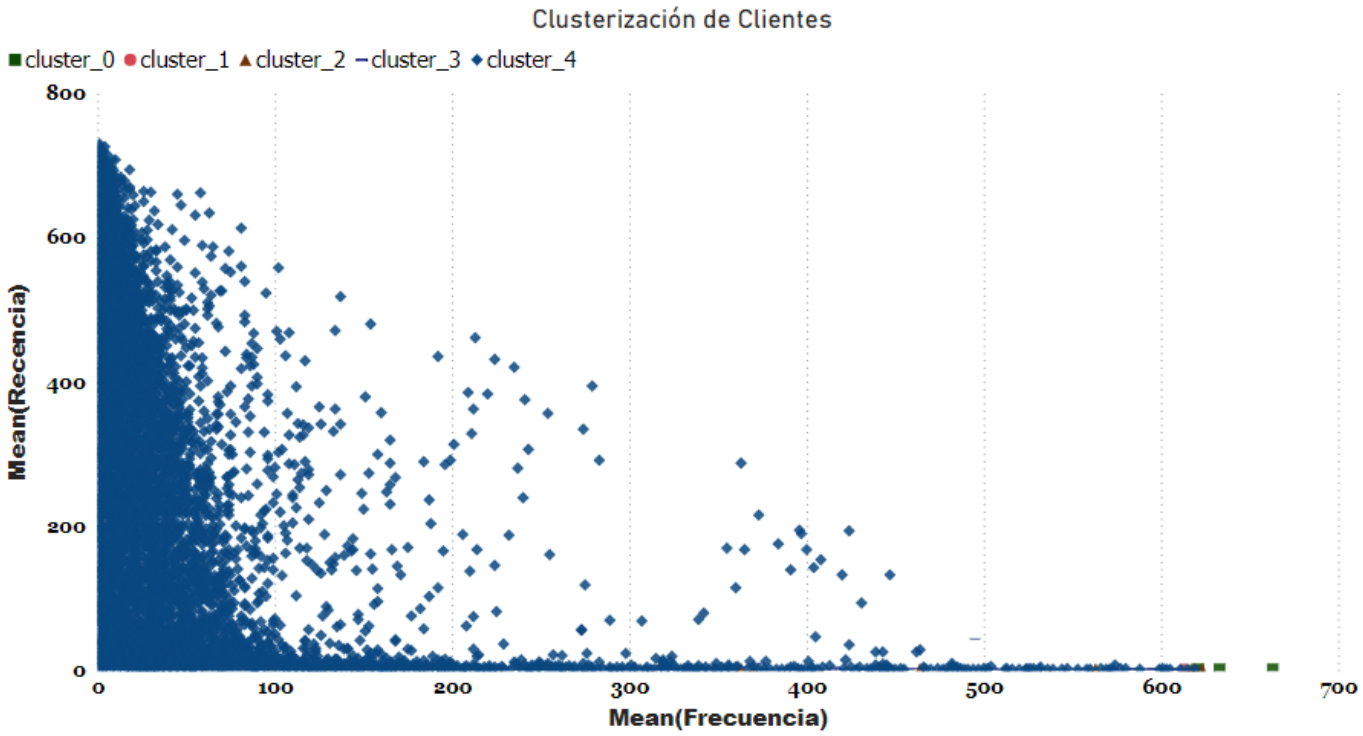


Figura 17. Resultado de Clusterización - Algoritmo K-Means.
Fuente: Elaboración propia.

El resultado del diagrama de dispersión anterior representa una relación inversa proporcional, es decir que lo positivo en este caso es que la variable de recencia disminuya para que la variable de frecuencia incremente. Ejemplo el Cluster_0 tiene una baja recencia, pero a menor recencia mayor frecuencia de compra. (Recuerde que la baja recencia representa una alta asignación de peso recencia).

Por lo tanto, se validó que evidentemente hay una correlación negativa entre la relación de cercanía pues el coeficiente de relación r = -0.37. Lo que indica que si la recencia incrementa la frecuencia disminuye y viceversa. Es por ello por lo que el clúster con mejor correlación es el Cluster_0.

Luego de obtener la segmentación de clústeres, se establecieron escalas de lealtad para cada uno de ellos, las cuales se pueden describir en la siguiente tabla:

Tabla 7. Escalas de lealtad.

Escala de Lealtad	Recencia (días)	Frecuencia (Transacciones)	Monto Promedio (Mensual)
Muy Alto	1 - 1	26 - 28	$724,322 - $891,231
Alto	1 - 1	21 - 26	$236,179 - $402,480
Medio	1 - 4	9 - 26	$91,202 - $215,510
Bajo	1 - 43	5 - 26	$10,447 - $84,928
Muy Bajo	1 - 730	0 - 26	($178) - $17,059

Fuente: Elaboración propia.

Como se puede observar aquellos clientes en escala de lealtad “muy alta” suelen hacer compras todos los días del mes, es decir, tienen una alta recencia lo que conlleva a una alta frecuencia y por ende montos de compras promedio relativamente altos. Un dato importante de destacar es que dentro de esta categoría o escala únicamente están clasificados cuatro clientes del universo total de clientes.

Con patrones de compra muy similares al grupo anterior están aquellos clientes con una lealtad “alta” suelen también hacer compras todos los días del mes, solo que su frecuencia y monto es relativamente más bajo. Y así sucesivamente vemos las categorías de lealtad hasta llegar a la categoría lealtad “muy bajo” en esta categoría entran incluso clientes no rentables con los que la compañía incluso está generando pérdidas, los cuales tiene recencia muy baja suponiendo que muchos de ellos hacen compras una o dos veces al año, por lo que su nivel de transaccionalidad o frecuencia también es muy bajo.

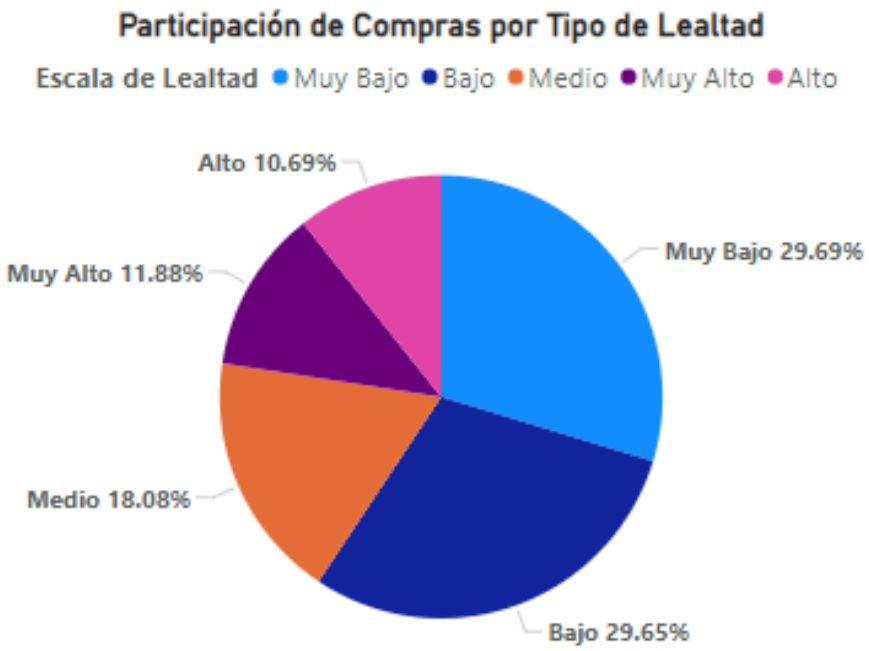


Figura 18. Peso de compras de clientes.
Fuente: Elaboración propia.

Sin embargo, cuando se observa la figura 18 se detecta que casi el 80% del monto de las ventas proviene de clientes con lealtad de “Media” a “Muy Baja”, en este caso se recomiendo al área comercial que pueda crear estrategias de marketing como descuentos o bonos, planes de crecimiento que incentiven a este tipo de clientes para a hacer compras regulares y aumentar el valor monetario de las mismas. Por otro lado, aquellos clientes que tienen una fidelidad “Muy alta” o “Alta” se les debe premiar con monetario o con los mismos productos para así mantener su lealtad.

CAPITULO V. CONCLUSIONES Y RECOMENDACIONES

A continuación, se presentan las conclusiones que responden a la problemática investigada y de los resultados que se obtuvieron.

5.1 CONCLUSIONES

Dada una prueba a definición mediante los resultados de los parámetros definidos en cada uno de los indicadores y la aplicación de estadística pura se confirma la Hipótesis H1 en donde, la implementación de un modelo de clusterización para la segmentación de clientes en Corporación Multi Inversiones, Si es eficiente ya que permite identificar grupos de clientes claramente diferenciados y patrones significativos en sus comportamientos de compra.

Debido a la eficacia de ejecución, el tiempo de procesamiento eficiente del algoritmo K-Means y además la precisión de su clusterización devolviendo agrupaciones mediante la evaluación matemática de los comportamiento y patrones con similitudes de todos los clientes evaluando un lapso de tiempo se obtuvieron resultados objetivos y por lo tanto, se ha llegado a la conclusión, de que la aplicación de este algoritmo representa una gran mejora del proceso de clasificación clientes para CMI, pues actualmente su clasificación es subjetiva (Tradicional) y estas tienen a equivocarse.

Mediante la aplicación de este modelo se puede llegar alcanzar y expandir la implementación de clusterización a subcategorías como ser el portafolio de productos infaltables o los productos más rentables, para la realización de una clasificación de clientes mediante la delimitación de productos.

Dadas las diferentes pruebas realizadas durante el desarrollo de este modelo y consecuentemente en los resultados que este devolvió, que la técnica RFM por si sola (sin la aplicación del algoritmo de k-means) no sería del todo eficiente, pues solo está categorizando los clientes de forma numérica, sin embargo, con la aplicación del algoritmo K-Means este juega un papel sumamente importante en los resultados de todo el modelo.

El algoritmo K-Medoid en este modelo resulto no ser factible, para esta investigación, debido a su prolongada ejecución y procesamiento de los datos y, además, los resultados que este algoritmo devolvió no tenían la precisión que se encontró al aplicar en el algoritmo de K-Means.

Para el caso de CMI en la clusterización de sus clientes representa una relación inversa proporcional, puesto que, aunque la relación sea negativa entre la relación de cercanía resulta ser matemáticamente positivo ya que en este caso la recencia deberá disminuir para que la frecuencia incremente, este resultado es preciso lo que se busca para los clientes de CMI.

5.2 RECOMENDACIONES

Se recomienda una reestructuración en el proceso de gobernanza, desarrollo de inventario y diccionario de datos, así mismo, se recomienda la implementación y ejecución de una política y calidad en los datos. Esto como resultado, que se encontraron varias inconsistencias en los datos, como campos sin clasificación o identificados como “Código Desconocido”, clientes no identificados “Nulos” o “Vacíos” y clientes identificados como ”Código Para Muestra De Laboratorio”. Lo anterior se sugiere para obtener una mayor confiabilidad de los resultados.

Debido a que se encontró varias inconsistencias en la data, se sugiere realizar un preprocesamiento exhaustivo de los datos, tal como selección, limpieza, transformación y modelado, antes de ser sometidos a la aplicación de algoritmos de aprendizaje automático pues, de esto dependerá la calidad y seguridad de los resultados.

En base a las escalas de lealtad que se encontraron, se recomienda que el área de Marketing enfoque sus esfuerzos de inversión según el porcentaje de participación de compras para cada tipo de cliente, para que pueda monetizar o rentabilizar cada una de las estrategias ejecutadas en sus clientes.

Se recomienda que previo a la implementación de este proyecto todo el equipo de Inteligencia Comercial, el área de Comercialización y Ventas, reciban una capacitación de la utilidad y beneficios de aplicación de Clustering en las empresas, y como se beneficiará la compañía al tomar mejores decisiones basadas en herramientas de Analítica, ayudando a mejorar la atención y retención de los clientes.

CAPITULO 6. APLICABILIDAD

En este capítulo para poder definir la aplicabilidad de la investigación se hizo una evaluación de la situación de la empresa, se definieron los objetivos de la propuesta se hizo un plan para desarrollar la misma.

6.1 NOMBRE DE LA PROPUESTA:

"Implementación de un Clustering mediante la Segmentación de Clientes basada en la Técnica RFM".

6.2 JUSTIFICACIÓN DE LA PROPUESTA

Al momento de desarrollar estrategias para retener clientes, muchas empresas se enfrentan al desafío de identificarlos de manera efectiva para poder alcanzarlos. A pesar de disponer de una amplia cantidad de información sobre sus clientes, la gestión eficiente de estos datos es compleja sin las técnicas, herramientas y procedimientos adecuados. Actualmente, la mayoría de las empresas han logrado clasificar a sus clientes por la estructura de la atención, tal es el caso de CMI, estos pueden pertenecer a los cinco diferentes tipos de macro canales que son: Moderno, Tradicional, Food Service, Terceros e Industria, y dividirse en subgrupos de clientes por su tipología y grupo de cliente, pero en esa clasificación no se toman en cuenta otros indicadores relevantes como su recencia, frecuencia y monto para comprender mejor el valor de sus clientes finales y determinar su nivel de fidelidad.

Es por ello, que en esta propuesta consiste en brindar una guía general aquellas empresas de comercialización que cuenten con una base de datos de sus clientes puedan segmentarlos en base a herramientas de analítica con el fin de implementar un sistema de recomendaciones y aplicar mejores estrategias de retención de clientes.

6.3 ALCANCE DE LA PROPUESTA

La propuesta presentada es implementación de un clustering mediante la segmentación de clientes basada en la técnica RFM tomando en cuenta la transaccionalidad total (portafolio total) en un periodo de dos años exactos, sin embargo, el alcance del proyecto es amplio en un futuro este análisis puede hacerse delimitando el portafolio de productos infaltables o con los productos más rentables, eso quedaría a decisión de la empresa.

6.3.1 OBJETIVOS DE LA IMPLEMENTACIÓN

En base a la problemática, misión y visión de la empresa, se plantearon los siguientes objetivos al momento de desarrollar la propuesta:

- Crear grupos de clientes finales en base a su comportamiento de compra.
- Identificar el nivel de lealtad que los clientes finales tienen hacia la empresa.
- Desarrollar estrategias de marketing empresarial dirigido a los clientes finales.

6.4 DESCRIPCIÓN Y DESARROLLO A DETALLE DE LA PROPUESTA

Tabla 8. Plan de desarrollo de la propuesta.

Fases	Tareas	Entradas	Salidas
Comprensión del Negocio	• Evaluación de la situación. • Objetivos del negocio. • Plan del proyecto	Información sobre la empresa	Análisis de la información del negocio.
Compresión de los datos	• Recopilación de los datos. • Descripción de los datos. • Exploración de los datos. • Verificación de la calidad de los datos	Recolección inicial de los datos de clientes y transacciones de la empresa.	Obtención de la base de datos con las que se va a trabajar.
Preparación de los datos	• Selección de datos a analizar (muestreo). • Limpieza de los datos. • Integración y formato de los datos.	Preprocesamiento de los datos proporcionados por la empresa.	Datos seleccionados y preparados para aplicar la Minería de Datos.
Modelado	• Seleccionar las técnicas de modelado. • Diseñar un modelo de comprobación. • Construir el modelo. • Evaluar el modelo.	Desarrollo de técnica RFM y algoritmos de clusterización.	Generación del modelo de clusterización.
Evaluación	• Evaluar los resultados obtenidos.	Modelado de Minería de Datos.	Evaluación e interpretación de los resultados obtenidos.

Fuente: Elaboración propia.

6.4.1 DESARROLLO DE ELEMENTOS

A continuación, se hará una descripción de todas las herramientas, instrumentos y procesos utilizados para el desarrollo de la propuesta:

6.4.2 HERRAMIENTAS

Las Herramientas utilizadas en este proyecto no llevan un mayor costo, puesto que son herramientas de Open Source, como ser:

Power BI: Es una plataforma empresarial integral y expansible en el ámbito de la inteligencia de negocios (BI), caracterizada por atributos de autoservicio que resultan idóneos para empresas de gran envergadura. Establece una conexión con los datos, proporciona capacidades visuales y facilita la integración fluida de elementos visuales en las aplicaciones de uso cotidiano **Fuente especificada no válida.**.

Figura 19. Logotipo Power BI.
Fuente: **Fuente especificada no válida.**

Knime: representa una vanguardia en la categoría emergente de herramientas identificadas por Gartner como Plataformas de Ciencia de Datos y Aprendizaje Automático. Estas herramientas facultan a profesionales de la ciencia de datos, analistas y usuarios empresariales para interactuar con sus datos, dando lugar a la creación, implementación y gestión de modelos de análisis avanzado **Fuente especificada no válida.**.

En cuanto a sus destacadas características, se puede resaltar su facilidad de uso. La programación de aplicaciones en KNIME se torna altamente intuitiva gracias a una interfaz visual. La conexión visual de nodos que encapsulan diversas funciones, junto con la integración de módulos automatizados de Aprendizaje Automático y modelos predefinidos, simplifica la analítica avanzada para usuarios empresariales que carecen de experiencia en este ámbito.

Simultáneamente, proporciona un entorno óptimo para desarrolladores avanzados al permitir la integración de programación en Python o R **Fuente especificada no válida.**.

Figura 20 Logo Knime
Fuente: **Fuente especificada no válida.**

Power Query: conocido como (Obtener & transformar), puede importar datos externos o conectarse a ellos y darles forma, por ejemplo, quitar una columna, cambiar un tipo de datos o combinar tablas de maneras que satisfagan sus necesidades (Microsoft, 2023).

Figura 21. Herramienta Power Query.
Fuente: (Microsoft, 2023)

6.4.3 PROCESOS

La recolección inicial de los datos de los clientes se hizo a través de la base de datos transaccional de los clientes de Corporación Multi Inversiones, los datos proporcionados necesarios para el análisis corresponden a clientes y ventas realizadas entre las siguientes fechas:

Tabla 9. Tiempo de recolección de datos.

Periodo Inicial	Periodo Final
01 de noviembre del 2021	31 de octubre del 2023

Fuente: Elaboración propia.

Es decir, comprendiendo todas las transacciones de compras de clientes de dos años exactos. Los datos recopilados se dividen en dos categorías:

- **Clientes:** el cual comprende el ID del cliente, el nombre del cliente, canal al que pertenece, tipo de negocio, meses de antigüedad desde la primera compra, años de antigüedad desde la primera compra, meses de antigüedad desde la última compra, forma de pago, municipio y departamento.
- **Ventas:** registros diarios de compras realizadas durante el periodo del 2021 al 2023.

6.4.3.1 DESCRIPCIÓN DE LOS DATOS

Los datos obtenidos en la recolección se detallan de la siguiente manera:

- Cliente principal
- Nombre del cliente
- Macro Canal
- Tipo de negocio
- Meses de Antigüedad Primera Compra
- Años de Antigüedad Primera Compra
- Meses de Antigüedad Ultima Compra
- Forma de pago
- Municipio
- Departamento

6.4.3.2 EXPLORACIÓN DE LOS DATOS

Como resumen del análisis EDA se puede decir, que la mayor concentración de transacciones y clientes de Corporación Multi Inversiones se encuentra en el Canal Tradicional, sin embargo, cuando analizamos los canales por monto de venta, el canal que mayor monto total aporta al negocio son los Canales Food Service y Terceros.

Pero a nivel de montos promedios por transacción, el canal que tiene clientes con mayores compras en monto es el Canal Industrial y Moderno, estos canales tiene un menor volumen de clientes, sin embargo, las compras promedio de estos clientes son bastante altas en comparación al promedio de toda la cartera de clientes.

Adicional se puede decir que tanto Canal Moderno como Terceros, tiene una recencia alta, prácticamente son clientes muy frecuentes que aportan transacciones al negocio todos los días. Los

canales más estables en termino de antigüedad de clientes se encuentran en Industrial y Moderno con un promedio de relación entre 7 y 8 años desque que realizaron la primera compra.

6.4.3.3 PREPARACIÓN DE LOS DATOS

En esta fase se hizo una selección y limpieza de los datos a analizar, se hizo un preprocesamiento de los datos proporcionados por la empresa y se prepararon para aplicar la minería de datos.

La preparación de los datos se realizó mediante la herramienta Power Query ya que el dataset compartido por Corporación Multi Inversiones fue mediante Excel y era bastante robusto debido a la cantidad de registros que contenía la consulta, por lo tanto, fue compartido en varios archivos de Excel. Los cuales se dividieron en años 2021, 2022 y 2023, es importante mencionar que el dataset contenía una cantidad de 2 años exactos desde el 1 de noviembre de 2021 al 31 de octubre de 2023, esta información se consolidó de la siguiente manera: los años 2021 y 2023 se almacenaron en una tabla y el año 2022 se almaceno en otra tabla.

6.4.3.4 LIMPIEZA DE LOS DATOS

Para la limpieza de los datos se trabajó con las columnas necesarias para el estudio, transformándola para llevar la data lo más limpia posible a Knime. Los pasos para utilizar en esta herramienta se describen a continuación:

Paso 1. Cargar la data a Power Query, las columnas que se utilizaron se mencionan a continuación:

→ Cliente Principal

→ Día

→ Monto Neto

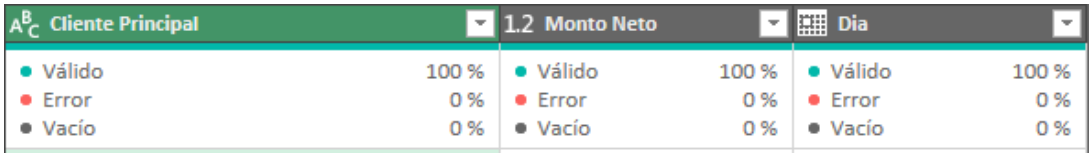


Figura 22. Columnas del Dataset.
Fuente: Elaboración propia.

Paso 2. Transformación de “ID Cliente” y “Nombre Cliente”. Para este paso se realizó una división de la columna por posición, en este caso la columna “Cliente Principal” se compone de la siguiente manera:

Figura 23. Nomenclatura de cliente.
Fuente: Elaboración propia.

En donde:

La nomenclatura por tipo Producto: Esta puede ser de tipo Comercial o Tipo Industrial.

ID del Cliente: Es el ID único que tienen los clientes, desde su registro en la base.

Nombre del Cliente: Es el nombre del Cliente.

Por lo tanto:

Al hacer la división por las primeras 4 posiciones tenemos como resultado 2 columnas de “Cliente Principal” en donde la primera contiene los primeros 4 dígitos y la segunda tiene el resto del nombre. En este mismo paso se realizó el proceso de filtración de columna “Cliente Principal 1”, en la que se filtra únicamente por los clientes comerciales ya que la empresa quiere conocer la segmentación de estos y no la de clientes industriales. Una vez que se tiene limpia la columna “Cliente Principal 1”, se procede con la división por delimitador de la columna “Cliente Principal 2”, pues se pretende dividir el “ID Cliente” con el “Nombre del Cliente” y así mismo, se elimina la columna “Cliente Principal 1”, ya que no es relevante para la aplicación de esta investigación, obteniendo al final el siguiente resultado:

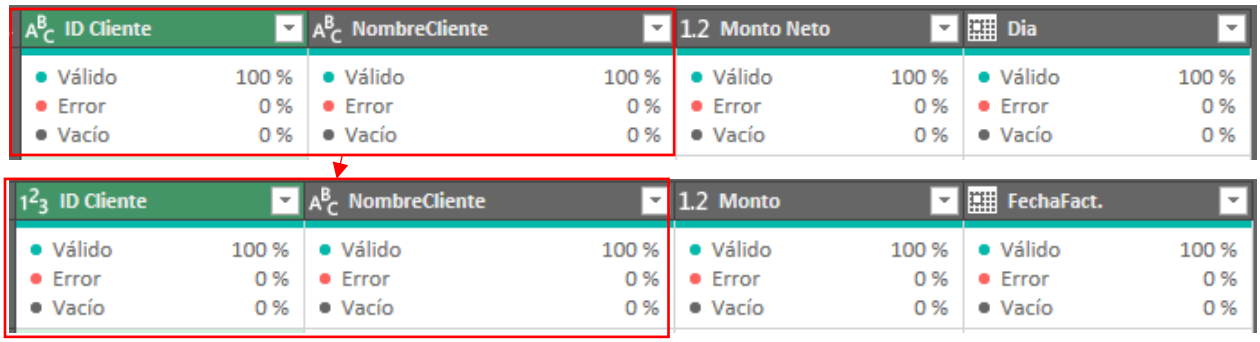


Figura 24. Transformación de Columnas.
Fuente: Elaboración propia.

Es importante mencionar que en este paso se le dio formato a la columna "ID Cliente" de String a Integer.

Paso 3. Limpieza de los datos. En este paso se realizaron las siguientes omisiones de data, pues se detecta que la base de datos contenía datos de prueba, o inválidos que podían sesgar la información y generar o valores atípicos. Por lo tanto, se aplicaron los siguientes filtros:

1. El Monto deberá ser diferente de "0"
2. El ID Cliente deberá ser diferente de "Null" o "Vacío"
3. El Nombre Cliente no deberá ser "CODIGO PARA MUESTRA DE LABORATORIO"
4. El Nombre Cliente no deberá ser "CODIGO DESCONOCIDO"
5. El Nombre del cliente deberá ser diferente de "Null" o "Vacío"

Para la aplicabilidad del proceso antes mencionado se ejecutó el siguiente código de lenguaje M en el editor avanzado en ambos archivos de Excel "Años 2022" y "Años 2021 – 2023".

Años 2021 - 2023

Opciones de presentación

```
let
    Origen = Table.Combine({#"Año 2021", #"Año 2023"}),
    #"Dividir columna por posición" = Table.SplitColumn(Origen, "Cliente Principal", Splitter.SplitTextByPositions({0, 4}, false), {"Cliente Principal.1", "Cliente Principal.2"}),
    #"Filas filtradas" = Table.SelectRows(#"Dividir columna por posición", each ([Cliente Principal.1] = "1290")),
    #"Dividir columna por delimitador" = Table.SplitColumn(#"Filas filtradas", "Cliente Principal.2", Splitter.SplitTextByDelimiter("-", QuoteStyle.Csv),
     {"Cliente Principal.2.1", "Cliente Principal.2.2", "Cliente Principal.2.3"}),
    #"Columnas quitadas" = Table.RemoveColumns(#"Dividir columna por delimitador",{"Cliente Principal.1", "Cliente Principal.2.1"}),
    #"Columnas con nombre cambiado" = Table.RenameColumns(#"Columnas quitadas",{{"Cliente Principal.2.2", "ID Cliente"}, {"Cliente Principal.2.3", "NombreCliente"}}),
    #"Tipo cambiado" = Table.TransformColumnTypes(#"Columnas con nombre cambiado",{{"ID Cliente", Int64.Type}}),
    #"Columnas con nombre cambiado2" = Table.RenameColumns(#"Tipo cambiado",{{"Monto Neto", "Monto"}, {"Dia", "FechaFact."}}),
    #"Columnas reordenadas" = Table.ReorderColumns(#"Columnas con nombre cambiado2",{"ID Cliente", "NombreCliente", "FechaFact.", "Monto"}),
    #"Filas filtradas4" = Table.SelectRows(#"Columnas reordenadas", each ([Monto] <> 0)),
    #"Filas filtradas1" = Table.SelectRows(#"Filas filtradas4", each [ID Cliente] <> null and [ID Cliente] <> ""),
    #"Valor reemplazado" = Table.ReplaceValue(#"Filas filtradas1","CODIGO PARA MUESTRA DE LABORATORIO","",Replacer.ReplaceText,{"NombreCliente"}),
    #"Valor reemplazado1" = Table.ReplaceValue(#"Valor reemplazado","CODIGO DESCONOCIDO","",Replacer.ReplaceText,{"NombreCliente"}),
    #"Filas filtradas3" = Table.SelectRows(#"Valor reemplazado1", each [NombreCliente] <> null and [NombreCliente] <> "")
in
    #"Filas filtradas3"
```

Figura 25. Código utilizado en lenguaje M.
Fuente: Elaboración propia.

Tabla 10. Resultados Limpieza de Datos.

Indicador	Pre-Limpieza	Resultados Limpieza
Número de transacciones	1,302,114	1,277,858
Número de clientes	21,550	19,239
Numero de canales de atención	5	4

Fuente: Elaboración propia.

6.4.3.5 INTEGRACIÓN Y FORMATO DE LOS DATOS

Para la integración y formato de los datos se hizo una serie de nodos y pasos que se detallan a continuación:

Transformación y Preparación de los Datos.

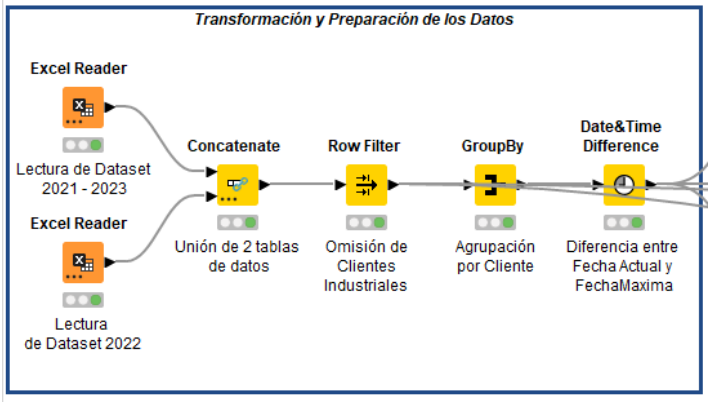


Figura 26. Transformación y Preparación.
Fuente: Elaboración propia.

En este paso se implementaron 2 nodos de lectura de Excel, ya que como se menciona en la fase anterior, debido a la cantidad de registros transaccionales en la data esta se particionó en 2, Dataset "Año 2022" y Dataset "Año 2021 – 2023". Como se observa en la figura 26, se agregaron 2 nodos que hacen la lectura de cada uno de los dataset. Es de suma importancia recordar validar el formato de las columnas o campos antes de ejecutar los nodos, con el fin de evitar problemas al momento de ejecutar el modelo.

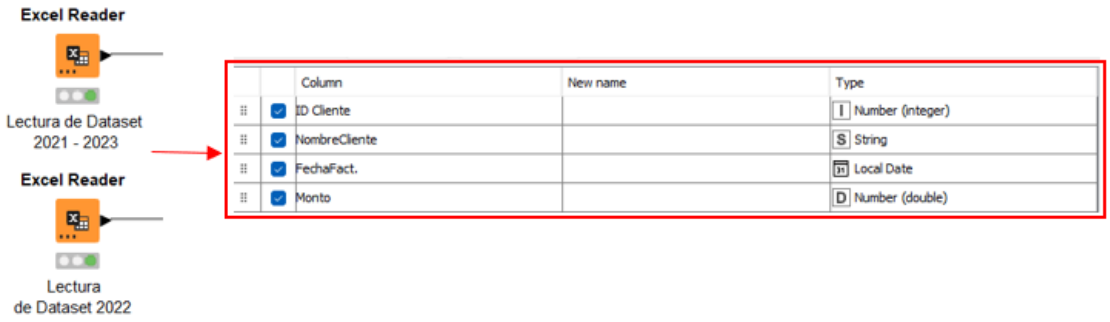


Figura 27. Nodos de Lectura y Configuración.
Fuente: Elaboración propia.

A continuación, se presenta el nodo "Concatenate", por medio del cual se realizó la función de anexar 2 tablas o dataset diferentes y convertir estas en un solo dataset global. Es importante mencionar que la configuración de este nodo deberá ser una unión de columnas permitiendo mantener datos de "ID Cliente" que estén duplicados, tal como se muestra en la figura 28.

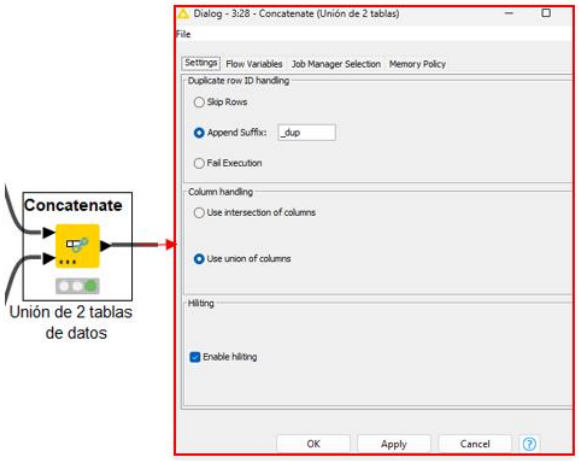


Figura 28. Nodo Concatenate y su configuración.
Fuente: Elaboración propia.

Una vez teniendo anexadas las tablas en un único dataset, se procedió con la Exclusión de las filas que contengan la palabra "Industria" en la columna "Macro Canal", es importante hacer mención que este paso se agregó, debido que a pesar de la limpieza realizada en la data en la sección 4.2.2 de "limpieza de datos", se detectó que aún había 12 clientes "Industriales", por lo que se tuvieron que omitir. Este paso se realizó con el nodo "Row Filter", a continuación, se muestra su configuración:

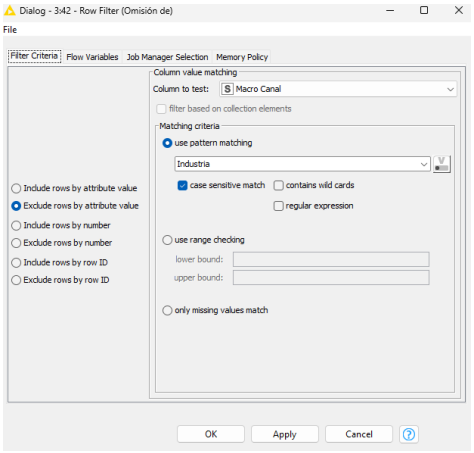


Figura 29. Configuración del nodo "Row Filter" - Preparación de los Datos.
Fuente: Elaboración propia.

Seguidamente se procedió con la agrupación de los clientes y para este proceso se utilizó el nodo “Group by” agrupando los clientes por su ID único, así mismo en la pestaña “Manual Aggregation” de este nodo se realizó la configuración de conteo de “Fechafact” con el fin de identificar el número de facturas por cliente, así mismo para la el campo “Fechafact” se configuro una agregación más de Máximo Valor, con el fin de identificar la última fecha de compra del cliente y en el campo “Monto” se configuro una agregación de Suma para obtener la suma del monto total por cliente, para este proceso se utilizó el nodo “Group by” de Knime, tal como se muestra en la figura 30.

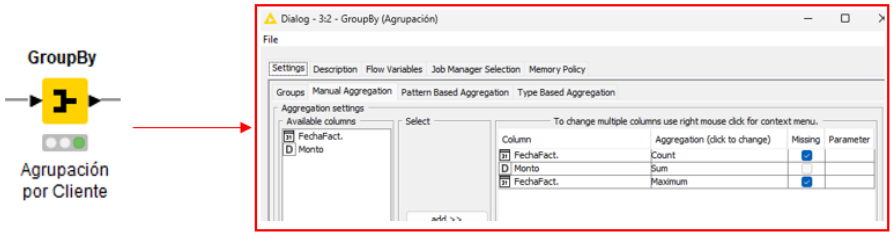


Figura 30. Nodo "Group by" y su configuración.
Fuente: Elaboración propia.

De la configuración anterior se obtuvo el siguiente resultado:

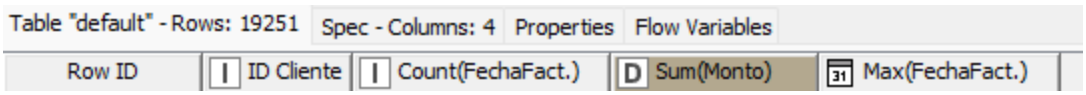


Figura 31. Columnas Resultado del nodo Group By.
Fuente: Elaboración propia.

Una vez identificada la última fecha de compra del cliente con el nodo anterior, es necesario identificar la cantidad de días que han pasado desde su ultimo registro de compra, para obtener ese dato se utilizó el nodo "Date&TimeDifference", en el cual se utilizó como columna base la columna "Max(Fechafact), realizando el cálculo de diferencia contra la fecha del día actual (Es importante mencionar que existen 2 opciones de uso si para identificar la fecha actual, la opción de Fecha/hora Actual de ejecución que esta tomara por default la fecha y hora que el equipo tenga configurado, así mismo está la fecha y hora fijada por el analista, ambas funcionan a la perfección pero la elección será definida por preferencia según sea el caso del analista), en el caso de esta investigación se utilizó la opción de Fecha y Hora fijada y la granularidad sería en base a días, ya que lo se busca obtener es el conteo de los días transcurridos entre esas 2 fechas.

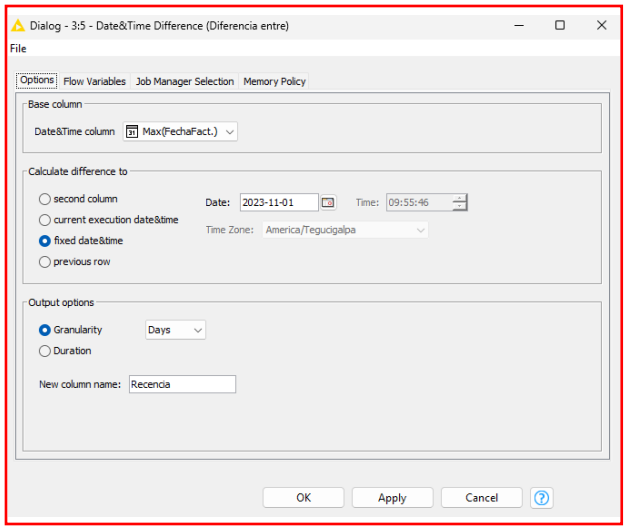


Figura 32. Nodo "Date&TimeDifference" y su configuración.
Fuente: Elaboración propia.

6.4.3.6 MODELADO

Para la aplicación del modelo primero se realizó en primera instancia la Técnica RFM llevando a cabo una serie de pasos para lograr la misma los que se detallan a continuación:

Cálculo de la recencia. En este módulo se modela los datos a manera de obtener el peso de la frecuencia mediante el uso de nodos que aporten solución al dato esperado, tal como se observa en la figura 33.

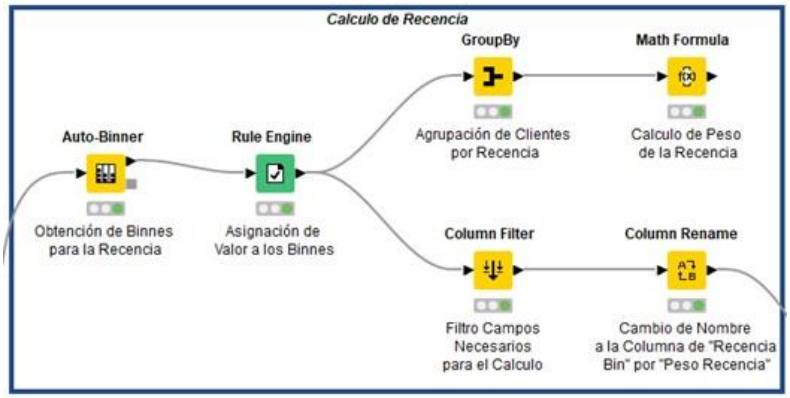


Figura 33. Cálculo de la Recencia.
Fuente: Elaboración propia.

Como se puede observar en la figura 34, el primer nodo utilizado es el "Auto-Binner", para la obtención de Binnes para la recencia, para ello se aplica la configuración únicamente en el campo "Recencia" (que es la diferencia de días entre la última fecha de compra y la fecha actual), asignando un numero de 5 Binnes frecuenciales con formato numerado, tal como se observa en a figura 34.

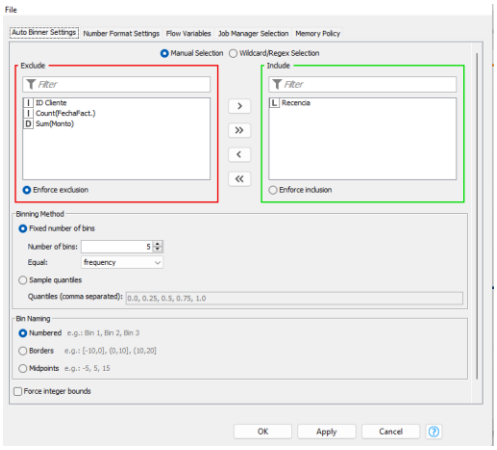


Figura 34. Configuración del nodo "Auto-Binner" – Calculo de la Recencia.
Fuente: Elaboración propia.

A continuación, se procede con la asignación de valor a los Binnes, mediante el nodo Rule Engine", para la configuración de este modo, se dio un valor del 1 al 5 a los Binnes, es decir entre menor es el Bin mayor puntaje tiene la recencia, pues entre menos días ausente de su última compra mayor es el puntaje de su recencia y entre mayor sea cantidad de días ausente desde su última compra menor será el puntaje de recencia, por lo tanto, la configuración de este nodo queda de la siguiente forma:

$Recencia [Binned]$="Bin 1"=>5

$Recencia [Binned]$="Bin 2"=>4

$Recencia [Binned]$="Bin 3"=>3

$Recencia [Binned]$="Bin 4"=>2

$Recencia [Binned]$="Bin 5"=>1

El valor devuelto por este nodo se remplazará en la columna "Recencia Binned", tal como se observa en la figura 35.

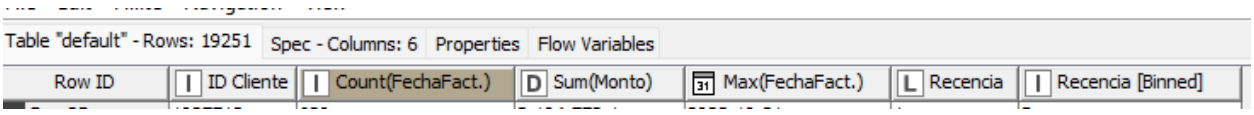


Figura 35. Columnas hábiles en el cálculo de la recencia.
Fuente: Elaboración propia.

Una vez teniendo los resultados mostrados en la figura anterior, se procede a delimitar el dataset únicamente con los campos necesarios para el cálculo de la recencia, estos campos serían: "ID Cliente", "Max (Fechafact)", "Recencia", "Recencia [Binned]", este proceso se realizó con el nodo "Column Filter", siendo la figura 20 la representación de lo antes mencionado.

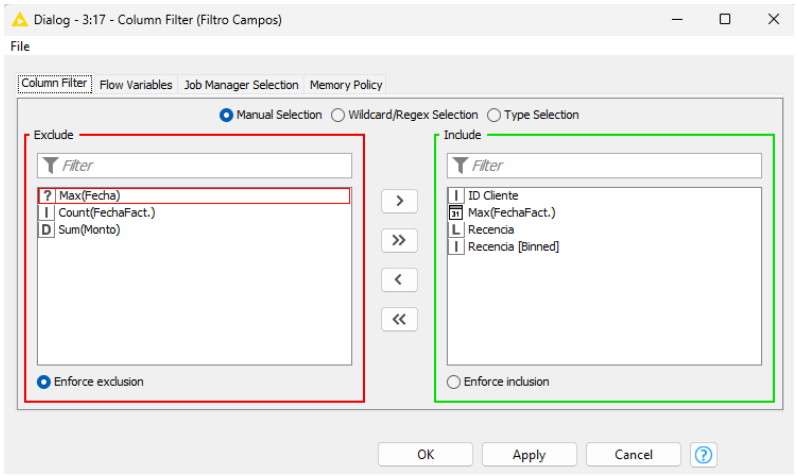


Figura 36. Configuración de nodo "Column Filter" – Calculo de la Recencia.
Fuente: Elaboración propia.

Una vez teniendo los campos necesarios para calculo se procede con el cambio de nombre de la columna “Recencia Binned” a “Peso Recencia” con un formato tipo entero, para este paso se utilizó el nodo “Column Rename” y se configuro tal como se muestra en la figura 37.

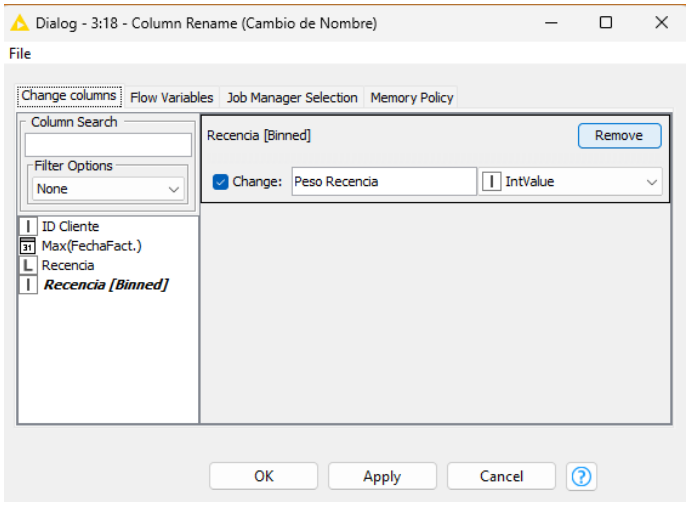


Figura 37. Configuración de Nodo "Column Rename" -Calculo de la Recencia.
Fuente: Elaboración propia.

Con todos los pasos aplicados anteriormente ya está listo el cálculo de la recencia sin embargo en este módulo se aplicó un paso más de observación netamente informativo y con el fin de tener una visión más asertiva del resultado esperado mediante la agrupación de los clientes por “Recencia Binned”, para la observación de la cantidad de clientes que se están segmentando y

como se estaría distribuyendo el Peso de la recencia, este paso no es obligatorio, pero se considera importante. Para realizar lo anterior mencionado se utilizó el nodo “Group By” y el nodo “Math Formula” tal como se logra visualizar en la figura 38.

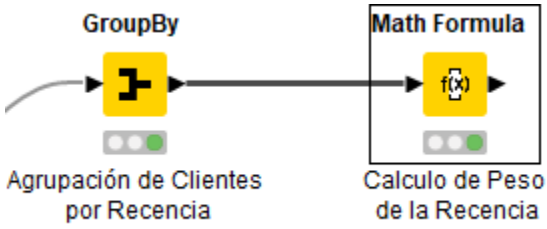


Figura 38. Nodos "Group by" "Math Formula"- Calculo de la Recencia.
Fuente: Elaboración propia.

Obteniendo la siguiente tabla:

Table "default" - Rows: 5 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	Recencia [Binned]	Min*(Recencia)	Max*(Recencia)	Count*(ID Cliente)	Peso Recencia
Row0	1	335	730	4194	1
Row1	2	28	334	3858	2
Row2	3	7	27	3870	3
Row3	4	4	6	3693	4
Row4	5	1	3	3636	5

Figura 39. Resultado del nodo "Math Formula" - Calculo de la Recencia.
Fuente: Elaboración propia.

En donde se logra observar que los clientes con menor recencia, es decir aquellos que tienen mayor cantidad de días ausentes son un total de 4,794 a diferencia de aquellos clientes cuya recencia es de 5 son un total de 3,636, por lo tanto, se puede interpretar de la siguiente manera:

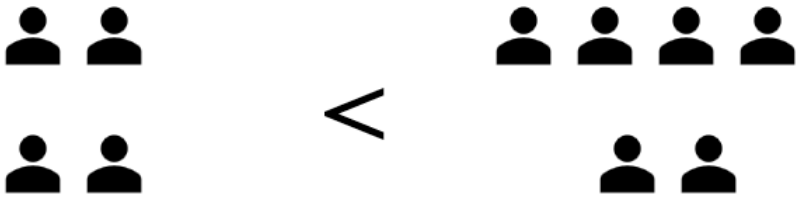


Figura 40. Ejemplo de clientes – Calculo de la Recencia.
Fuente: Elaboración propia.

Los Clientes con menor recencia es menor que los clientes con mayor recencia tal como se logra observar en la figura 40.

Modulo 2. Cálculo de la Frecuencia.

En este módulo se modela la información a manera de obtener como resultado el peso de la frecuencia mediante el uso de nodos aporten solución al dato esperado, tal como se observa en la figura 42.

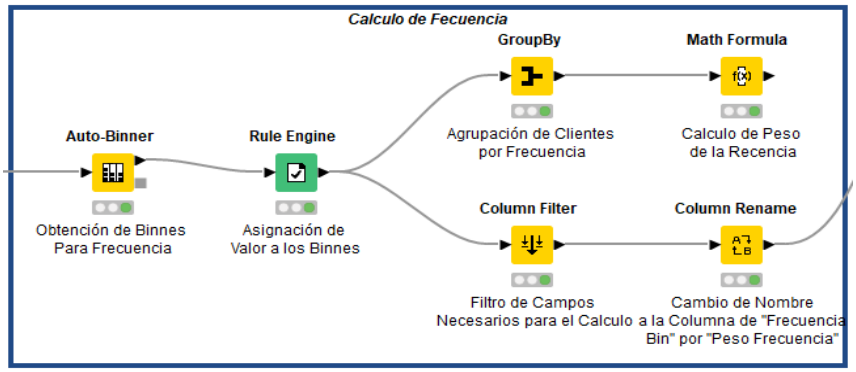


Figura 41. Cálculo de la Frecuencia.
Fuente: Elaboración propia.

Como se puede observar en la figura 41, el primer nodo utilizado es el “Auto-Binner”, para la obtención de Binnes para la frecuencia, para ello se aplica la configuración únicamente en el campo “Count (FechaFact.)” este campo se configuro en la transformación y preparación de los datos realizando un conteo de los registros de compra por fecha y agrupándolos por cliente. Se asigno un numero de 5 Binnes frecuenciales con formato numerado, tal como se observa en a figura 43.

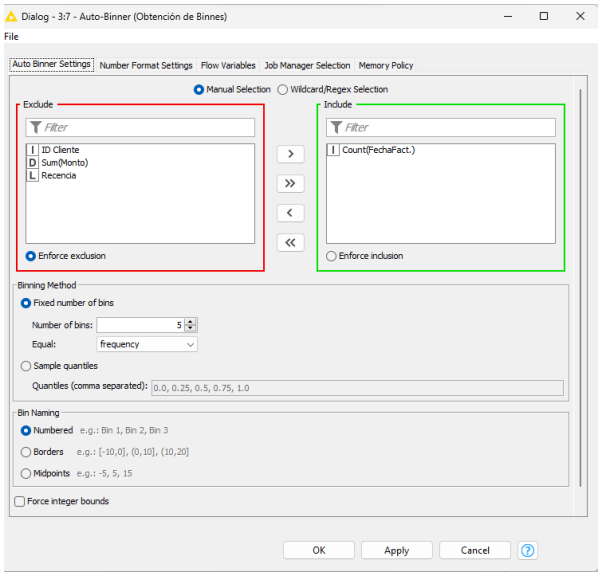


Figura 42. Configuración del nodo "Auto-Binner" - Calculo de la Frecuencia.
Fuente: Elaboración propia.

A continuación, se procede con la asignación de valor a los Binnes, mediante el nodo Rule Engine", para la configuración de este modo, al igual que en el módulo de cálculo de la recencia se dio un valor del 1 al 5 a los Binnes, pero la diferencia es que para el cálculo de la frecuencia el valor se da en orden continuo descendente:

$Count (FechaFact.) [Binned]$="Bin 5"=>5

$Count (FechaFact.) [Binned]$="Bin 4"=>4

$Count (FechaFact.) [Binned]$="Bin 3"=>3

$Count (FechaFact.) [Binned]$="Bin 2"=>2

$Count (FechaFact.) [Binned]$="Bin 1"=>1

El valor devuelto por este nodo se remplazará en la columna "Count(FechaFact.) [Binned]", tal como se observa en la figura 43.

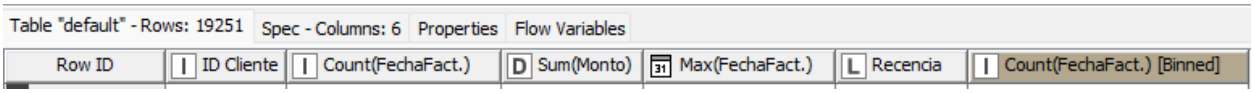


Figura 43. Columnas hábiles en el cálculo de la Frecuencia.
Fuente: Elaboración propia.

Una vez teniendo los resultados mostrados en la figura anterior, se procedió con la delimitación del dataset con los campos necesarios para el cálculo de la frecuencia, estos campos serían: “ID Cliente”, “Count(Fechafact)”, “Max (Fechafact)”, “Count(Fechafact) [Binned]”, este proceso se realizó con el nodo “Column Filter”, siendo la figura 44 la representación de lo antes mencionado.

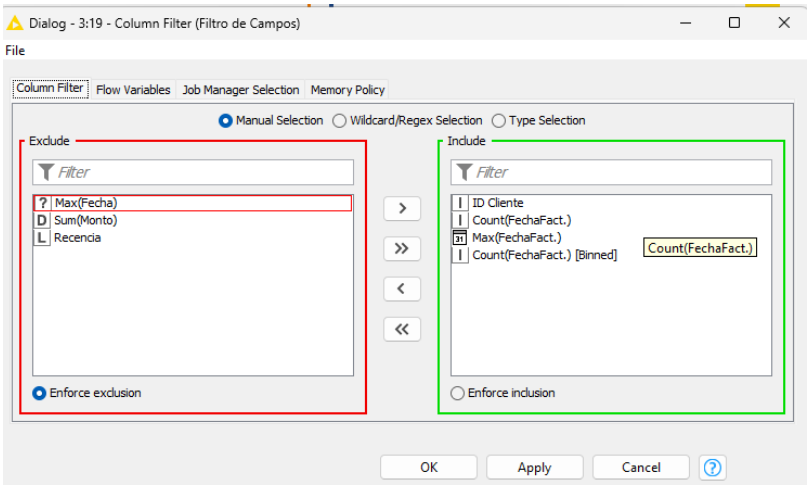


Figura 44. Configuración de nodo "Column Filter" – Calculo de la Frecuencia
Fuente: Elaboración propia.

Una vez teniendo los campos necesarios para calculo se procede con el cambio de nombre de la columna “Count(Fechafact) [Binned]” por “Peso Frecuencia” con un formato tipo entero, para este paso se utilizó el nodo “Column Rename” y se configuro tal como se muestra en la figura 45.

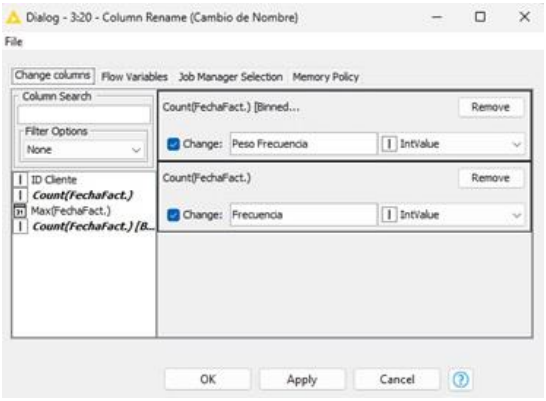


Figura 45. Configuración de Nodo "Column Rename".
Fuente: Elaboración propia.

Al igual que en el cálculo de la Recencia con todos los pasos aplicados anteriormente ya está listo el cálculo de la frecuencia, sin embargo, en este módulo también se aplicó un paso más de observación mediante la agrupación de los clientes por "Frecuencia Binned", para la observación de la cantidad de clientes que se están segmentando y como se estaría distribuyendo el Peso de la frecuencia. Para realizar lo anterior mencionado se utilizó el nodo "Group By" y el nodo "Math Formula" tal como se logra visualizar en la figura 46.

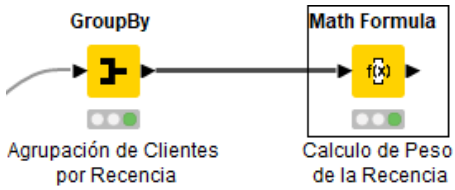


Figura 46. Nodos "Group by" "Math Formula" – Calculo de la Frecuencia.
Fuente: Elaboración propia.

Obteniendo la siguiente tabla:

File Edit Hilite Navigation View

Table "default" - Rows: 5 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	Count(FechaFact.) [Binned]	Count*(ID Cliente)	Min*(Count(FechaFact.))	Max*(Count(FechaFact.))	Peso Frecuencia
Row0	1	3866	1	9	1
Row1	2	3858	10	29	2
Row2	3	3797	30	58	3
Row3	4	3889	59	89	4
Row4	5	3841	90	663	5

Figura 47. Tabla resultado del nodo "Math Formula" - Calculo de la Frecuencia.
Fuente: Elaboración propia.

Los resultados obtenidos en la tabla anterior podrían interpretarse como aquellos clientes agrupados por un peso de frecuencia de 5 suman un total de 3,841 y de los cuales tienen un mínimo de 90 a un máximo 663 compras durante los 2 años que fueron expuestos en este modelo.

También podría interpretarse que los segmentos de clientes según su frecuencia de compra fueron distribuidos en cantidades similares a pesar de que las distancias en las cantidades mínimas y máximas de compra tienen una diferencia significativa.

Modulo 3. Cálculo del Monto.

En este módulo se modela la información a manera de obtener como resultado el peso del “Monto” mediante el uso de nodos aporten solución al dato esperado, tal como se observa en la figura 48.

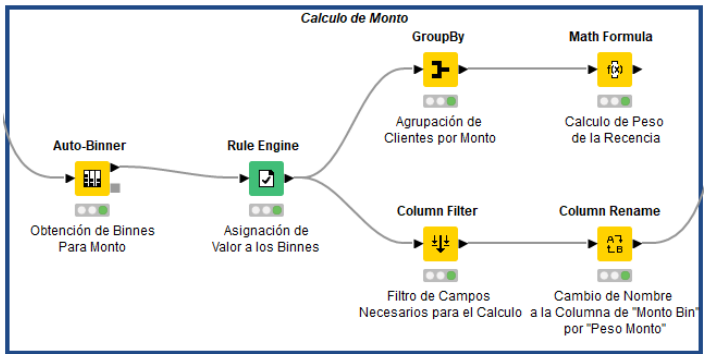


Figura 48. Cálculo del monto.
Fuente: Elaboración propia.

Como se puede observar en la figura 49, el proceso de modelación para el cálculo del “Monto” es similar a los aplicados en los 2 módulos anteriores, sin embargo, en la configuración del nodo “Auto-Binner” se incluirá el campo “sum(Monto)” tal como se observa en la figura 32.

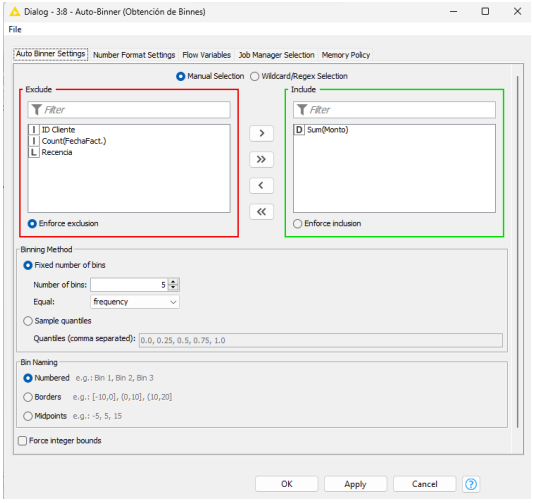


Figura 49. Configuración del nodo "Auto-Binner" - Calculo del Monto.
Fuente: Elaboración propia.

Así mismo para la configuración del nodo “Rule Engine”, se dará paso a asignar binnes a la columna “sum(Monto)”, tal como se observa en las siguientes líneas de código que se utilizaron para configurar el nodo:

$Sum (Monto) [Binned]$="Bin 5"=>5

$Sum (Monto) [Binned]$="Bin 4"=>4

$Sum (Monto) [Binned]$="Bin 3"=>3

$Sum (Monto) [Binned]$="Bin 2"=>2

$sum(Monto) [Binned]$="Bin 1"=>1

El valor devuelto por este nodo se remplazará en la columna “sum(Monto) [Binned]”, tal como se observa en la figura 50.

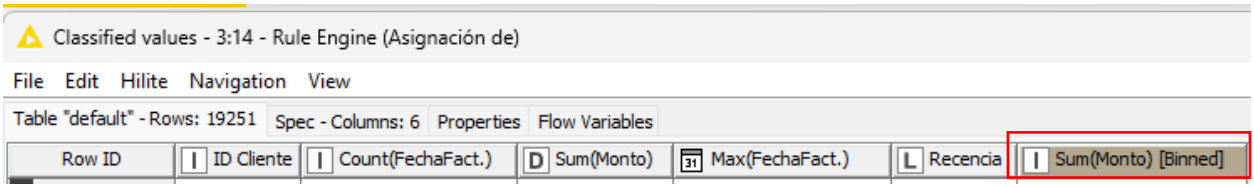


Figura 50. Columnas hábiles en el cálculo del Monto
Fuente: Elaboración propia.

Para el cálculo del monto se delimitaron a los siguientes campos “ID Cliente”, “sum(Monto)”, “Max (Fechafact.)”, “sum(Monto) [Binned]”, utilizando el nodo “Column Filter” tal como se observa su configuración en la figura 51.

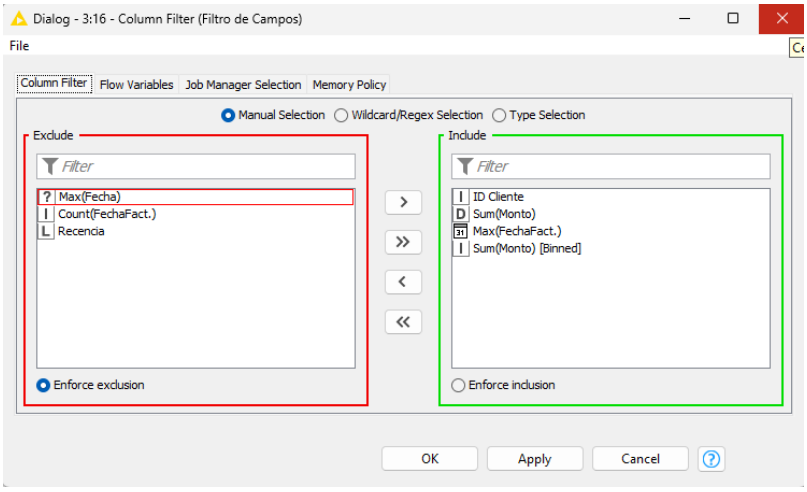


Figura 51. Configuración de nodo "Column Filter" – Calculo Monto.
Fuente: Elaboración propia.

Para la configuración del nodo “Column Rename” se realiza el cambio de nombre de la columna “sum(Monto)” por “Facturación” con un formato tipo decimal, así mismo se realizó cambio de nombre de la columna “sum(Monto) [Binned]” por “Peso Facturación” con un formato tipo Entero, tal como se observa en la figura 52 su respectiva configuración.

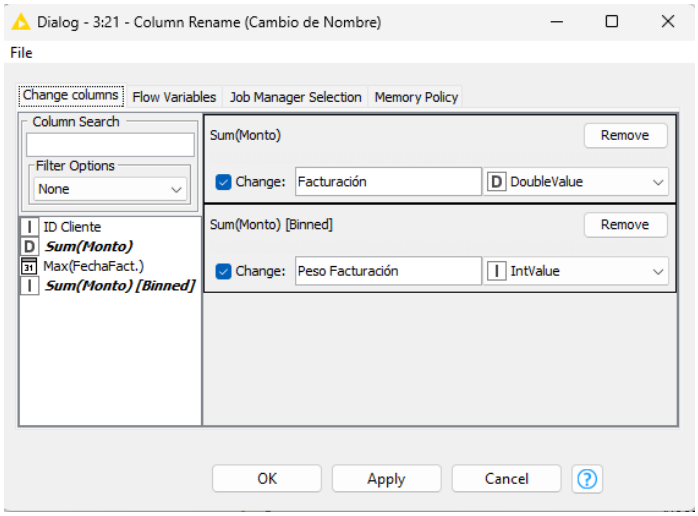


Figura 52 Configuración de Nodo "Column Rename" – Calculo del Monto
Fuente: Elaboración propia.

Al igual que en el cálculo de la Recencia y el cálculo de la Frecuencia se aplicó un paso más de observación mediante la agrupación de los clientes por “Monto Binned”, para la observación de la cantidad de clientes que se están segmentando y como se estaría distribuyendo el Peso del Monto. Para realizar lo anterior mencionado se utilizó el nodo “Group By” y el nodo “Math Formula” tal como se logra visualizar en la figura 53.

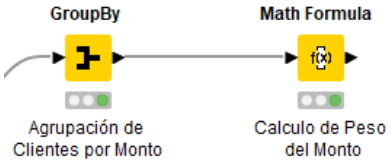


Figura 53. Nodos "Group by" "Math Formula" – Calculo del Monto.
Fuente: Elaboración propia.

Obteniendo la siguiente tabla:

Row ID	Sum(Monto) [Binned]	Count*(ID Cliente)	Min*(Sum(Monto))	Max*(Sum(Monto))	Peso Monto
Row0	1	3850	-4,260.66	215.12	1

Figura 54. Tabla resultado del nodo "Math Formula" – Calculo del Monto
Fuente: Elaboración propia.

De los resultados anteriores podría interpretarse como que aquellos clientes cuyo peso asignado es de 1 es decir, el más bajo, entre sus rangos tienen un mínimo con valores negativos de hasta -4,260.66 y valores positivos de 215.12. Por lo que estos clientes podrían representar una pérdida para Corporación Multi Inversiones.

Modulo 4. Calculo RFM

En este módulo se procede con la combinación de los 3 módulos mencionados y explicados anteriormente: calculo R, F y M para convertirlo en un solo valor total y de esta manera segmentar los clientes. Para la ejecución de este último modulo se utilizan 2 nodos, en 3 pasos tal como se observa en la figura 55.

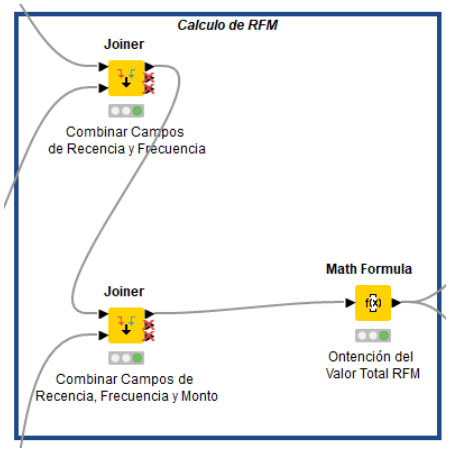


Figura 55. Cálculo de RFM.
Fuente Elaboración propia.

Para la combinación de los campos obtenidos en los 3 módulos anteriores se utilizó el nodo “Joiner”, en 2 pasos distintos, el primer paso se configura para combinar los campos de “Recencia y Frecuencia” obteniendo todas las columnas de ambas tablas mediante una izquierda combinación externa, que combine las columnas llave de cada tabla, en este caso es el “Id Cliente” en cada una, este proceso lo realiza mediante una comparación de valores combinando columnas por valores y

tipos. En la figura 56 se logra observar la configuración utilizada en el nodo para ejecutar las acciones mencionadas anteriormente.

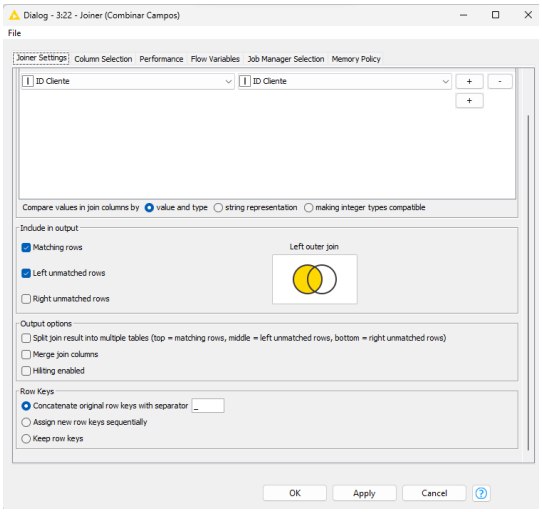


Figura 56. Configuración del nodo "Joiner" - 1 Calculo RFM.
Fuente: Elaboración propia.

Las tablas que fueron combinadas anteriormente son las tablas de “Calculo Recencia” y “Calculo Frecuencia”, desde los resultados obtenidos en los nodos “Column Rename” en ambos módulos.

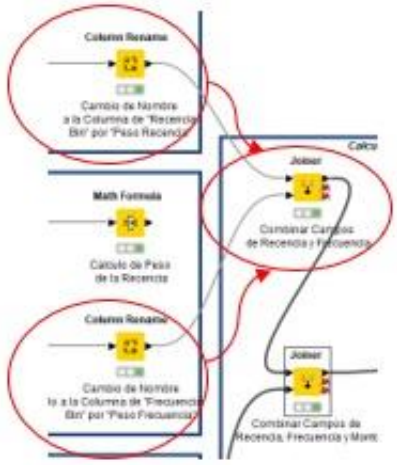


Figura 57. Configuración nodo "Joiner".
Fuente: Elaboración propia.

A diferencia del segundo “Joiner” en este módulo, el cual esta combinando los resultados obtenidos en el primer “Joiner” y el “Column Rename” del módulo del Cálculo del monto, con el objetivo de condensar toda la información en una sola tabla.

En la configuración del segundo “Joiner” utilizado, deberían verse de la siguiente manera la selección de columnas:

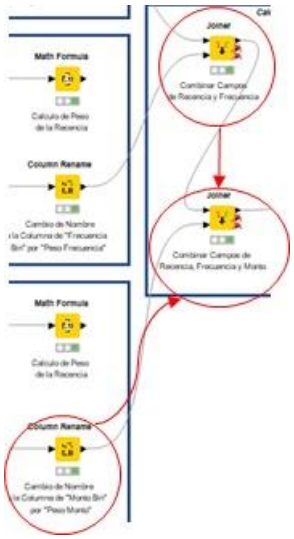


Figura 58.Joiner de Dataset de Frecuencia y Recencia - Cálculo de RFM.
Fuente: Elaboración propia.

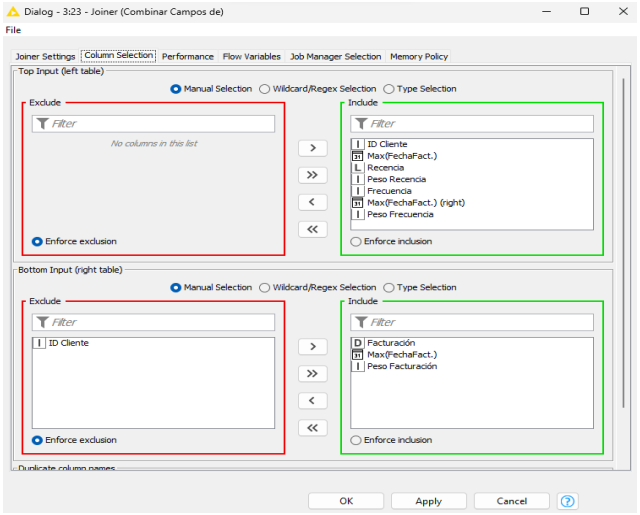


Figura 59. Unión de dataset monto, recencia, frecuencia – Cálculo de RFM.
Fuente: Elaboración propia.

Siempre combinado mediante la columna “Id Cliente”, exactamente tal como se configuro en el primer “Joiner”.

Obteniendo como resultado un total de 10 columnas y 19,251 registros, a continuación, se presenta una figura que representa el resultado obtenido en el segundo “Joiner”.

Table "default" - Rows: 19251 | Spec - Columns: 10 | Properties | Flow Variables

Columns: 10	Column Type	Column Index	Color Handler	Size Handler	Shape Han...	Filter Handler	Lower Bound	Upper Bound
ID Cliente	Number (integer)	0					10000	7017966
Max(FechaFact.)	Local Date	1					2021-11-01	2023-10-31
Recencia	Number (long)	2					1	730
Peso Recencia	Number (integer)	3					1	5
Frecuencia	Number (integer)	4					1	663
Max(FechaFact.) (right)	Local Date	5					2021-11-01	2023-10-31
Peso Frecuencia	Number (integer)	6					1	5
Facturación	Number (double)	7					-4260.6599...	2.13895349...
Max(FechaFact.) (right) (right)	Local Date	8					2021-11-01	2023-10-31
Peso Facturación	Number (integer)	9					1	5

Figura 60. Resultado Joiner 2 - Cálculo de RFM.
Fuente: Elaboración propia.

Una vez validado los resultados obtenidos anteriormente, se procede con la suma de las columnas “Recencia”, “Frecuencia” y “Peso Facturación”, para este paso se utilizará el nodo “Math Formula”, aplicando el siguiente código en su configuración:

$Peso Recencia$+$Peso Frecuencia$+$Peso Facturación$

Anexando el resultado en una nueva columna la cual se llamará “RFM”, la configuración de este nodo se muestra en la siguiente figura.

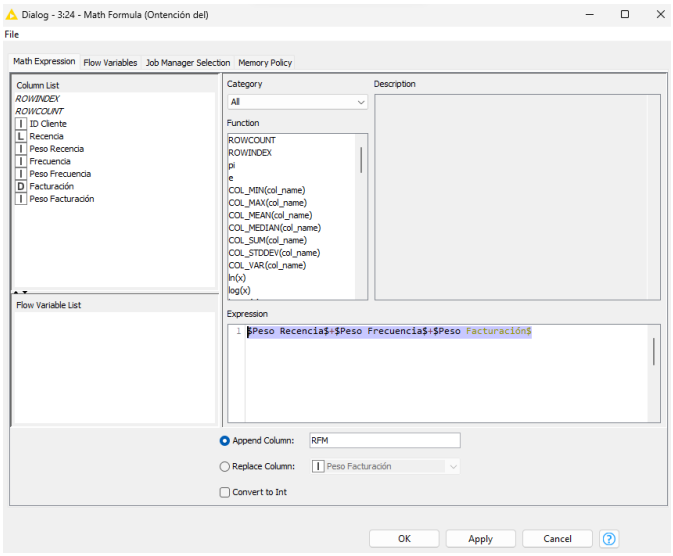
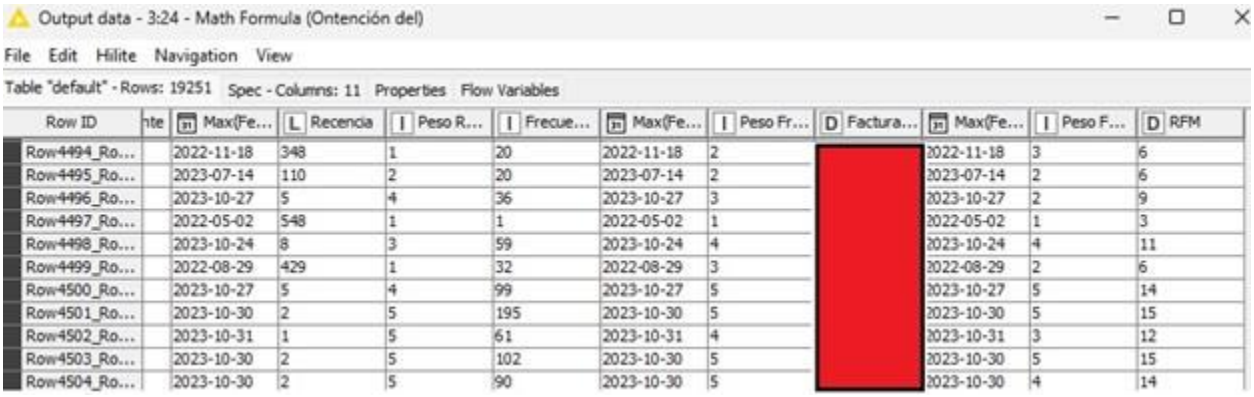


Figura 61. Configuración del Nodo “Math Formula” – Cálculo de RFM.

Fuente: Elaboración propia.

El resultado que se espera obtener del nodo “Math Formula” es basado en la suma de 3 quintiles es decir que si el resultado es 15 significa que ese cliente es un “Excelente Cliente”, pero esta segmentación quedara fijada la técnica de la que se hablara más adelante en este documento. Los valores que se espera tener en el campo “RFM” van desde el “1” como valor mínimo y “15” como valor máximo, se adjunta una figura como ejemplo de algunos resultados que se obtuvieron en esta investigación.


Output data - 3:24 - Math Formula (Ontención del)

File Edit Hilite Navigation View

Table "default" - Rows: 19251 Spec - Columns: 11 Properties Flow Variables

Row ID	nte	Max(Fe...	Recencia	Peso R...	Frecue...	Max(Fe...	Peso Fr...	Factura...	Max(Fe...	Peso F...	RFM
Row4494_Ro...		2022-11-18	348	1	20	2022-11-18	2		2022-11-18	3	6
Row4495_Ro...		2023-07-14	110	2	20	2023-07-14	2		2023-07-14	2	6
Row4496_Ro...		2023-10-27	5	4	36	2023-10-27	3		2023-10-27	2	9
Row4497_Ro...		2022-05-02	548	1	1	2022-05-02	1		2022-05-02	1	3
Row4498_Ro...		2023-10-24	8	3	59	2023-10-24	4		2023-10-24	4	11
Row4499_Ro...		2022-08-29	429	1	32	2022-08-29	3		2022-08-29	2	6
Row4500_Ro...		2023-10-27	5	4	99	2023-10-27	5		2023-10-27	5	14
Row4501_Ro...		2023-10-30	2	5	195	2023-10-30	5		2023-10-30	5	15
Row4502_Ro...		2023-10-31	1	5	61	2023-10-31	4		2023-10-31	3	12
Row4503_Ro...		2023-10-30	2	5	102	2023-10-30	5		2023-10-30	5	15
Row4504_Ro...		2023-10-30	2	5	90	2023-10-30	5		2023-10-30	4	14

Figura 62. Resultados nodo Math formula.
Fuente: Elaboración propia.

Paso 3: Segmentación en base a la técnica RFM

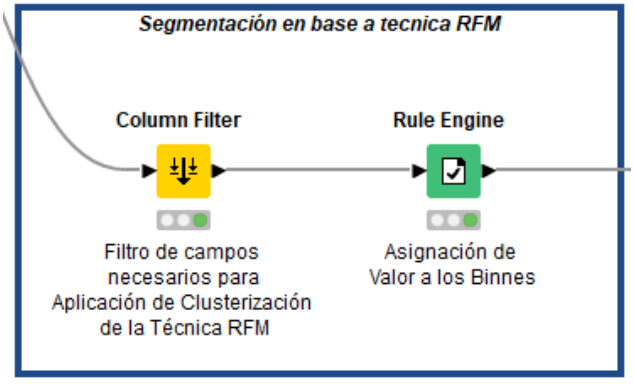


Figura 63. Segmentación en base a técnica RFM.
Fuente: Elaboración propia.

Para la aplicación de la técnica RFM se utilizó el nodo “Math Formula” del paso 2 conectándolo l nodo “Column Filter”.

Para la ejecución de esta técnica, inicialmente se procedió con el filtro de campos necesarios mediante el nodo “Column Filter”, para exclusión de los campos categóricos.

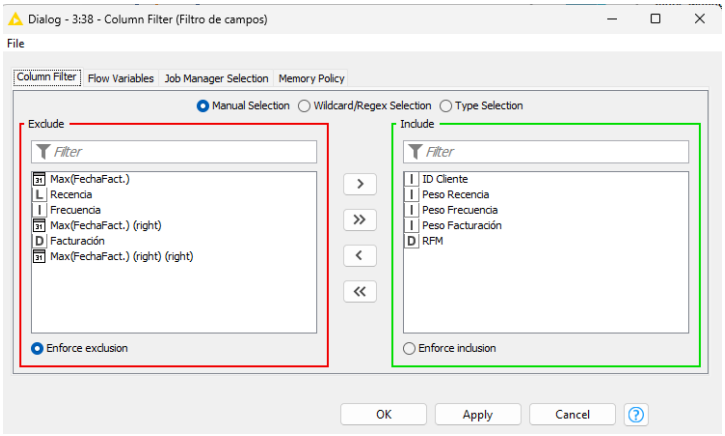


Figura 64. Configuración de nodo "Column Filter" - Segmentación RFM.
Fuente: Elaboración propia.

Así mismo, mediante el nodo “Rule Engine” se aplicarán las reglas antes mencionadas, en este caso, se ejecutó el siguiente código para la configuración de este nodo:

$Peso Recencia$ IN (4,5) AND $Peso Frecuencia$ IN (4,5) AND $Peso Facturación$ IN (4,5) => "Campeones"

$Peso Recencia$ IN (2,3,4,5) AND $Peso Frecuencia$ IN (3,4,5) AND $Peso Facturación$ IN (3,4,5) => "Clientes Fieles"

$Peso Recencia$ IN (3,4,5) AND $Peso Frecuencia$ IN (1,2,3) AND $Peso Facturación$ IN (1,2,3) => "Potencialmente Fieles"

$Peso Recencia$ IN (4,5) AND $Peso Frecuencia$ IN (1,2) AND $Peso Facturación$ IN (1,2,3,4,5) => "Nuevos clientes"

$Peso Recencia$ IN (3,4,5) AND $Peso Frecuencia$ IN (3,4,5) AND $Peso Facturación$ IN (1,2,3) => "Prometedores"

$Peso Recencia$ IN (2,3) AND $Peso Frecuencia$ IN (2,3) AND $Peso Facturación$ IN (2,3) => "Necesitan atención"

$Peso Recencia$ IN (2,3) AND $Peso Frecuencia$ <= 2 AND $Peso Facturación$ >= 4 => "A punto de riesgo"

$Peso Recencia$ <= 2 AND $Peso Frecuencia$ IN (1,2,3,4,5) AND $Peso Facturación$ IN (1,2,3,4,5) => "En riesgo"

$Peso Recencia$ <= 1 AND $Peso Frecuencia$ IN (4,5) AND $Peso Facturación$ IN (4,5) => "No puede perderlos"

$Peso Recencia$ <= 2 AND $Peso Frecuencia$ <= 2 AND $Peso Facturación$ <= 2 => "Perdidos"

Obteniendo el siguiente resultado de segmentación:

Row ID	Indicadores	Count*(ID Cliente)	Mean(Peso Recencia)	Mean(Peso Frecuencia)	Mean(Peso Facturación)
Row2	Clientes Fieles	5726	3.214	3.73	3.67
Row3	En riesgo	5396	1.223	1.478	1.608
Row1	Campeones	4205	4.551	4.677	4.618
Row6	Potencialmente Fieles	2197	3.738	2.078	1.954
Row4	Necesitan atención	1261	2	2.316	2.175
Row0	A punto de riesgo	167	2.323	1.844	4.317
Row7	Prometedores	161	3.72	4.075	1.963
Row5	Nuevos clientes	126	4.492	1.929	4.246

Figura 65. Resultado de Segmentación por Indicadores.
Fuente: Propia

Paso 4: Aplicación de Algoritmo K-Means

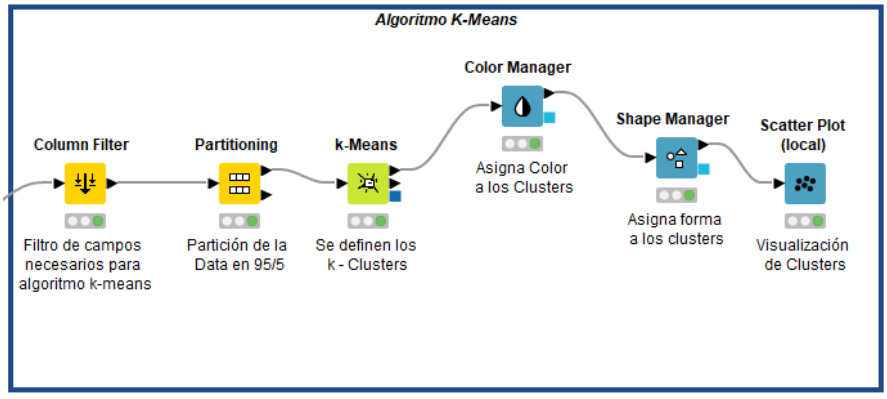


Figura 66. Algoritmo K-Means.
Fuente: Elaboración propia.

Como se observa en la figura 48, el primer nodo que se utiliza para la aplicación del algoritmo K-Means es el nodo “Column Filter”, ya que se necesita filtrar los campos necesarios para la aplicación de este algoritmo, pues es importante mencionar que el algoritmo K-Means solo trabaja con variables numéricas y no acepta variables categóricas, por lo tanto, estas serán excluidas.

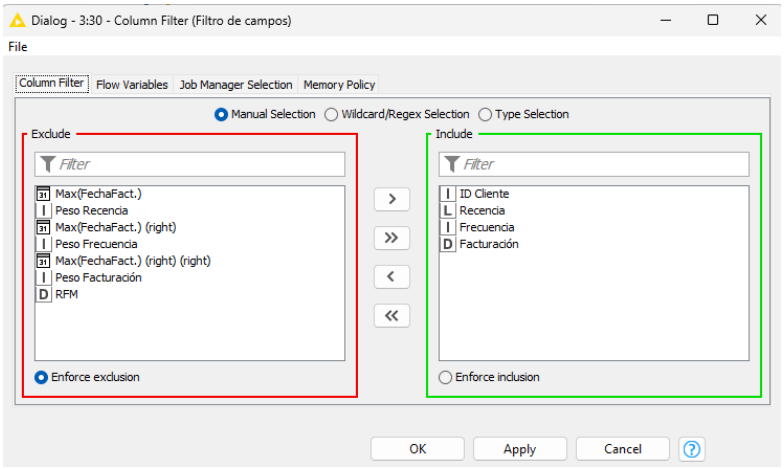


Figura 67. Configuración del nodo "Column Filter" - Algoritmo K-Means.
Fuente: Elaboración propia.

Una vez que se ejecutó el nodo configurado previamente, se procedió con la configuración del nodo “Partitioning”, con el objetivo de particionar la información con una relatividad del 95%, ya que este porcentaje será el número de filas de la tabla de entrada que están en la primera partición mediante un muestreo aleatorio de todas las filas.

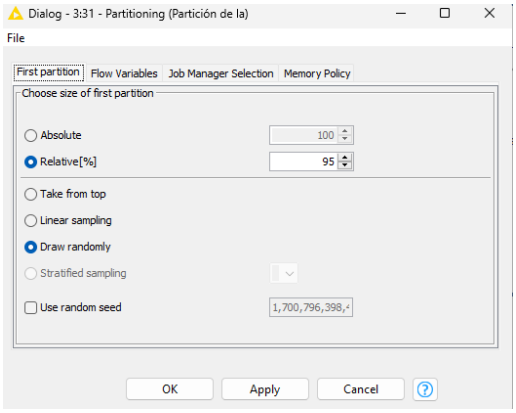


Figura 68. Configuración del nodo "Partitioning" - Algoritmo K-Means.
Fuente: Elaboración propia.

Cuando ya se ha particionado la data se procede con la asignación de numero predefinido de K Clusters, puesto que K-Means realiza una agrupación nítida que asigna un vector de datos a exactamente un clúster, este algoritmo de agrupamiento utiliza la distancia euclidiana en los atributos seleccionados inicializando los centroides con filas aleatorias de las tablas de entrada. En este proceso también se definió un número máximo de interacciones de 99, ya que este será el numero después de las cuales el algoritmo termina si no ha encontrado una solución estable antes.

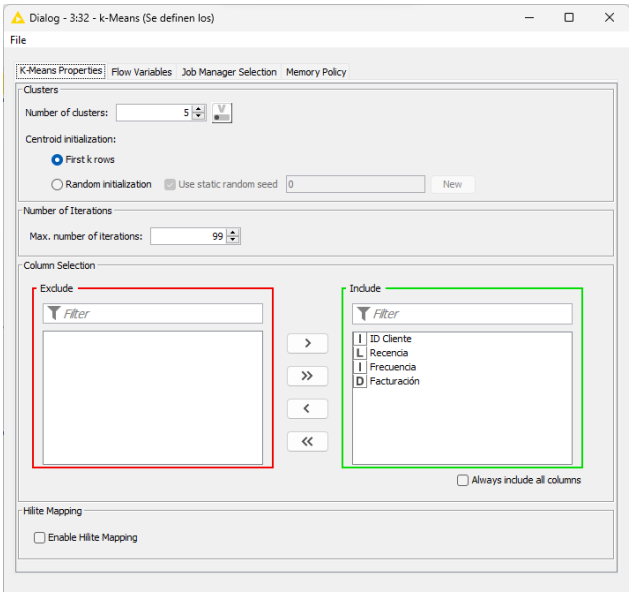


Figura 69. Configuración nodo "K-Means" - Algoritmo K-Means.
Fuente: Elaboración propia.

Los resultados obtenidos de la configuración aplicada en el nodo anterior reflejan que el algoritmo está evaluando los comportamientos de monto, recencia y frecuencia muy detalladamente pues los clústeres han hecho la segmentación de clientes de forma precisa mediante cada una de las iteraciones, tal como se observa en la figura 70.

Row ID	S Cluster	I Count*(ID Cliente)	D Percent*(Recencia)	D Percent*(Frecuencia)	D Percent*(Facturación)
Row0	cluster_0	4	0.022	0.022	0.022
Row1	cluster_1	7	0.038	0.038	0.038
Row2	cluster_2	31	0.17	0.17	0.17
Row3	cluster_3	158	0.864	0.864	0.864
Row4	cluster_4	18088	98.906	98.906	98.906

Figura 70. Resultado de Clusterización - Algoritmo K-Means.
Fuente: Elaboración propia.

Mediante los nodos “Color Manager” y “Shape Manager” se asignó color y forma a los clústeres para poder visualizarlos de manera gráfica y así comprender mejor la teoría anterior, por lo tanto:

-En el nodo “Color Manager” se asignó el set número 2 en la paleta de colores.

-En el nodo “Shape Manager” se adicionaron las siguientes figuras:

Shape Mapping

Values of Cluster	Shapes
cluster_0	Rectangle
cluster_1	Circle
cluster_2	Triangle
cluster_3	Reverse Triangle
cluster_4	Diamond

Figura 71. Configuración del nodo "Shape Manager" - Algoritmo K-Means.
Fuente: Elaboración propia.

Para la visualización de los resultados se utilizará el nodo “Scatter Plot (Local)”, para la visualización de la clusterización de forma gráfica

6.5 MEDIDAS DE CONTROL

6.5.1 INDICADORES Y MEDICIÓN

Para la técnica FRM se definieron los siguientes parámetros de medición de los indicadores de recencia, frecuencia y monto:

Tabla 11. Indicadores RFM.

Indicadores	Recencia	Frecuencia	Monto
Campeones	4, 5	4,5	4,5

Clientes Fieles	>=2	>=3	>=3
Potencialmente Fieles	>=3	<=3	<=3
Nuevos Clientes	4,5	1,2	2,3
Prometedores	>=3	>=3	<=3
Necesitan Atención	2,3	2,3	2,3
A punto de riesgo	2,3	<=2	>=4
En Riesgo	<=2	>=1	>=1
No puede tenerlos	<=1	4,5	4,5
Perdidos	<=2	<=2	<=2

Fuente: Elaboración propia

Para la Clusterización mediante el algoritmo K-Means se definieron los siguientes parámetros de medición:

Tabla 12. Escalas de lealtad K-Means.

Escala de Lealtad	Recencia (días)	Frecuencia (Transacciones)	Monto Promedio (Mensual)
Muy Alto	1 - 1	26 – 28	$724,322 - $891,231
Alto	1 - 1	21 – 26	$236,179 - $402,480
Medio	1 - 4	9 – 26	$91,202 - $215,510
Bajo	1 - 43	5 – 26	$10,447 - $84,928
Muy Bajo	1 - 730	0 – 26	($178) - $17,059

Fuente: Elaboración propia.

6.6 CRONOGRAMA DE IMPLEMENTACIÓN Y PRESUPUESTO

6.6.1 CRONOGRAMA

CRONOGRAMA DE IMPLEMENTACIÓN DEL MODELO DE CLUSTERIZACIÓN

	CADA SPRINT EQUIVALE A UNA SEMANA												Total	
Actividades	1	2	3	4	5	6	7	8	9	10	11	12	12 Semanas	Responsable
1. Compresión del Negocio														
1.1 Evaluación de la situación actual														Analista de Negocios/Negocio/Gerente de Proyectos
1.1 Definir Objetivos														Analista de Negocios/Negocio/Gerente de Proyectos
1.2 Establecer plan del proyecto														Analista de Negocios
2.Comprensión de los Datos														
2.1 Recopilación de las Fuentes de Datos														Analista de Negocios
2.2 Descripción de los Datos														Analista de Negocios
2.3 Analisis de Datos (EDA)														Analista de Negocios
2.4 Verificación de la calidad de los datos														Analista de Negocios
3. Preparación de los Datos														
3.1 Extracción de información														Analista de Negocios
3.2 Transformación de Datos														Analista de Negocios
3.3 Limpieza de los Datos														Analista de Negocios
3.4 Integración de los Datos														Analista de Negocios
3.5 Formato de los Datos														Analista de Negocios
4. Creación y Entrenamiento de Modelo														
4.1 Diseñar un Modelo de Machine Learning														Analista de Negocios
4.2 Contrucción del Modelo														Analista de Negocios
4.3 Entrenamiento de Modelo														Analista de Negocios
4.5 Analisis de Resultados														Analista de Negocios
5. Evaluación de Modelo														
5.1 Evaluar el modelo														Analista de Negocios
5.2 Evaluar los resultados obtenidos														Analista de Negocios
6. Implementacio/Despliegue														
5.1 Definición de Infraestructura y sus componentes para la seguridad														Analista de Negocios/Analista de Sistemas/Gerente de privacidad de datos
5.2 Implementación de Modelo														Analista de Negocios/Analista de Sistemas
5.3 Generación de Informes ejecutivos														Analista de Negocios
5.4 Cierre de Proyecto.														Gerente de Proyectos

Figura 72. Diagrama de Gantt.
Fuente: Elaboración Propia

6.6.2 PRESUPUESTO

Si la empresa cuenta con un departamento de inteligencia comercial y no dispone de los recursos para contratar una nueva persona entonces se recomienda que capacitar su personal, ya contratado, sin embargo, si el objetivo es implementar un área de Analítica para desarrollar esta y futuras propuestas, se recomienda contratar un Analista de negocios con un perfil técnico y conociendo del negocio. A continuación, se detalla el presupuesto que involucra el desarrollo de este proyecto de investigación.

Talento Humano:

El personal que se considera talento humano en la participación de este proyecto es:

Tabla 13. Presupuesto Recurso Humano.

Recurso Humano			
Rol	**Horas**	**Remuneración (L)**	**Valor Total (L)**
Analista de Negocios	528	170.00	89,760.00
		Total (L)	**89,760.00**

Fuente: Elaboración Propia.

Recursos de Hardware:

Los recursos materiales que se utilizaron fueron: Computadoras portátiles, Ratón, Monitores externos.

Tabla 14. Presupuesto Recursos Hardware.

Recursos Hardware			
Unidad	**Cantidad**	**Valor Unitario (L)**	**Valor Total (L)**
Computadores P.	1	25,000	25,000.00
Ratón	1	600	600.00
Monitores externos	1	700.00	700.00
		Total (L)	**26,300.00**

Fuente: Elaboración Propia.

Recursos de Servicios:

Los recursos de servicios que se tomaron en cuenta son: Luz, teléfono e internet:

Tabla 15. Presupuesto Recursos de Servicios

Recursos de Servicios			
Unidad	**Cantidad Hrs**	**Valor Total (L)**	
Internet	200		4,000.00
Luz	400		2,000.00
Teléfono	1		500.00
		Total (L)	**6,500.00**

Fuente: Elaboración propia.

Recursos de Software:

Para la ejecución del proyecto se utilizaron los siguientes recursos: Paquete de Office (Teams, Word, Excel, Power Point, Power Query), Knime, Power BI.

Tabla 16. Presupuesto Recursos Software.

Recurso de Software			
Unidad		**Cantidad**	**Valor Total (L)**
Paquete de Office 365		1	2,425.50
Licencia Power BI PRO (3 Meses)		1	751.92
		Total (L)	**3,177.42**

Fuente: Elaboración Propia.

Recurso de Capacitación:

Debido a que solo se va a contratar una persona, esta deberá tener las habilidades de procesamiento de datos en Knime y conocimiento análisis y visualización de datos en Power BI.

Tabla 17 Recurso de Capacitación.

Recurso de Capacitación	
Unidad	**Valor Total (L)**
Capacitation Udemy: Microsoft Power BI Data Analyst	2,082.22
Knime L1: Basic Proficiency in KNIME Analytics Platform	0.00
Knime L2: Advanced Proficiency in KNIME Analytics Platform	1,225.00
Knime L3: Proficiency in KNIME Software for Collaboration and Productionization	2,450.00
Total (L)	**5,757.22**

Fuente: Elaboración Propia

Recursos de Datos:

Los datos utilizados fueron proporcionados por la empresa Corporación Multi Inversiones, su costo se describe a continuación:

Tabla 18. Recursos de Datos.

Recurso de Datos	
Unidad	**Valor Total (L)**
Datos del Cliente	0.00
Total (L)	0.00

Fuente: Elaboración Propia.

Presupuesto Total:

El presupuesto total corresponde a la suma de los gastos que se han descrito en las tablas anteriores:

Tabla 19. Presupuesto Total.

	Presupuesto Total	
	Valor Total (L)	**Valor Total ($)**
Recurso Humano	89,760.00	3,664.00
Recurso de Hardware	26,300.00	1,074.00
Recurso de Software	3,177.42	130.00
Capacitaciones	5,757.22	235.00
Servicios	6,500.00	265.30
Recurso de Datos	0.00	0.00
Total (L)	**131,494.64**	**5,367.12**

Fuente: Elaboración Propia.

6.7 CONCORDANCIA DE LOS SEGMENTOS DE LA INVESTIGACIÓN

Tabla 20 Matriz de Concordancia

Título de Investigación: Implementación de un modelo de clusterización mediante la segmentación de perfil de clientes para Corporación Multi Inversiones.

<table>
<tr><th>Problema General</th><th>Objetivo General</th><th>Hipótesis</th><th>Teorías/Metodologías de Sustento</th><th colspan="3">Variable Independiente</th></tr>
<tr><td rowspan="2">¿Puede un modelo de clusterización ayudar a CMI a identificar y segmentar eficientemente los perfiles de sus clientes?</td><td rowspan="2">Implementar un modelo de clusterización que identifique la segmentación de perfil de cliente en el área comercial.</td><td rowspan="6">H1: Un modelo de clusterización puede segmentar eficientemente los perfiles de clientes en la empresa CMI, permitiendo identificar grupos de clientes claramente diferenciados y patrones significativos en sus comportamientos y preferencias.

H0: Un modelo de clusterización no es capaz de segmentar eficientemente los perfiles de clientes en la empresa CMI.</td><td>Teorías</td><td colspan="3">"Segmentación del Perfil de Cliente"</td></tr>
<tr><td rowspan="2">1. Teoría de Afinidad.
2. Teoría del Consumidor.
3. Teoría del Comportamiento del Consumidor.
4. Teoría de la Probabilidad.</td><td colspan="3">Variable Dependiente</td></tr>
<tr><td>Problemas Específicos</td><td>Objetivos Específicos</td><td>Metodologías</td><td colspan="3">"Modelo de Clusterización"</td></tr>
<tr><td>¿Se pueden crear clústeres de clientes en base a su comportamiento de compra?

¿Cuál es el nivel de lealtad identificado en los clientes de CMI?

En base al modelo de clusterización ¿Se puede crear un sistema de recomendación para mejorar la experiencia de los clientes actuales?</td><td>1. Crear grupos de clientes finales en base a su comportamiento de compra.

2. Identificar el nivel de lealtad de los clientes de CMI.

3. Proponer un sistema de recomendación en base al perfil de cliente para mejorar la experiencia de los clientes actuales.</td><td>1. Técnica RFM
2. Algoritmo K-Means</td><td>Dimensiones

Indicadores de Cliente

Indicadores de Algoritmo</td><td>Indicadores

Quintil de Recencia
Quintil Frecuencia
Quintil Monto

Número de Clúster
Cantidad de Iteraciones</td><td>Unidad Medida

Peso Recencia
Peso Frecuencia
Peso Monto

Número</td></tr>
<tr><td>Tipo y Diseño</td><td>Población y Muestra</td><td>Técnicas y Procedimientos</td><td colspan="3">Estadística Para Utilizar</td></tr>
<tr><td rowspan="2">Estudio: Longitudinal
Diseño: No Experimental
Alcance: Descriptivo
Enfoque: Cuantitativo</td><td>Población: 1,302,114 registros de transacciones de clientes, en un periodo de 2021-2023</td><td>Técnica: Observación</td><td colspan="3">Descriptivo</td></tr>
<tr><td>Muestra: 1,277,858 de registros de transacciones de clientes con la exclusión de clientes industriales y valores erróneos.</td><td>Procedimientos: 1. Obtención del dataset que fue brindado por parte de Corporación Multi Inversiones.
2. Limpieza, transformación y normalización de los datos.
3. Análisis EDA (Análisis Exploratorio de Datos).
4. Aplicación de la técnica RFM.</td><td colspan="3">Se usan tablas y figuras, datos emitidos por el modelo, lo cual ayuda a observar de forma visual y estructurada para la fácil comprensión de los datos numéricos.

Cuantitativo

Las Variables continuas presentes en el dataset facilitan la medición la técnica RFM a través de asignación de Peso a Recencia, Frecuencia y Monto.</td></tr>
</table>

Fuente: Elaboración propia.

BIBLIOGRAFÍA

Agudelo Viana, L. G., & Aigneren Aburto, J. M. (2008). *Diseños de investigación experimental y no-experimental.*

Agudelo, M., Alveiro, C., Saavedra, B., & Ramiro, M. (2013). El CRM como herramienta para el servicio al cliente en la organización. *Visión de futuro, 17*. https://doi.org/ISSN 1668-8708

Analytats. (21 de 03 de 2019). *Segmentación de clientes: un análisis RFM en Knime*. Analytats: https://blog.analystats.com

Aranda, D. F. (2017). Estudio de la precipitación máxima diaria anual en la Región Hidrologica, con Base en Distancias Euclidianas. *Investigaciones Geográficas, Boletín del Instituto de Geografía, UNAM*(65, 2008), 56-67. https://doi.org/ISSN 0188-4611

Arboleda, S. G. (2023). *Desarrollo de una herramienta para procesos ETL* . Universidad de Antioquia.

Arrelucea Zapata, G. (2020). *"IMPLEMENTACIÓN DE UN MODELO DE CLUSTERIZACIÓN PARA LA SEGMENTACIÓN DEL PERFIL DEL CLIENTE EN EL ÁREA COMERCIAL DE SUPERMERCADOS.* LIMA-PERU: UNIVERSIDAD PRIVADA DEL NORTE.

Barriga, N. R. (2023). *Modelo de Segmentación para SARLAFT en R4G.* Bogota, DC, Colombia.

Beltrán Martínez, M. (2001). *MINERÍA DE DATOS.* Puebla: Benemérita Universidad Autónoma.

Bocanegra, O. R., & Quispe, C. A. (2012). *Algoritmo de clustering utilizando k-means.* Lima: Facultad de Ingeniería en Sistemas.

Camargo Vega, J. J., Camargo Ortega, J. F., & Joyanes Aguilar, L. (2014). Conociendo Big Data. *Revista Facultad de Ingeniería*.

Casal, J., & Mateu, E. (2003). Tipos de Muestreo. *Rev. Epidem. Med. Prev.*(1), 3-7.

Chamba Jiménez, S. F. (2015). *Minería de Datos para segmentación de clientes en la empresa tecnológica Master PC.*

Cheng, C.-H., & Chen, Y.-S. (2009). Clasificación de la segmentación del valor del cliente mediante el modelo RFM y la teoría RS. *ElSevier*, 4176-4184.

Copyright IBM Corporation. (2021). *Conceptos básicos de ayuda de CRISP-DM*.

Corporación Multi Inversiones. (2019). *Política de Ciberseguridad.*

Corporación Multi Inversiones. (2023). *CMI*. CMI: https://somoscmi.com/es/cmi-alimentos/

Corrales, J. A. (2020). Segmentación de clientes: guía para categorizar a tus consumidores y orientar las decisiones de negocio. *Rockcontent Bog*. R.

Cuadros López, Á. J., Gonzales Caicedo, C., & Jiménez Oviedo, P. C. (2017). Análisis multivariado para segmentación de clientes basada en RFM. *Tecnura, 21*(54), 41-51. https://doi.org/10.14483/22487638.12957

da Silva, D. (16 de Octubre de 2020). Cuál es el significado de cluster y cuáles son las ventajas de implementarlo en tu empresa. *Blog de Zendesk*, 1.

da Silva, D. (05 de Agosto de 2022). *Como es el comporatmiento del consumidor*. zentesk.com.mx

Dell Technologies Forum. (12 de Diciembre de 2012). *https://www.dell.com/*. https://www.dell.com/: https://www.dell.com/es-es/dt/corporate/newsroom/announcements/2012/12/20121212-01.htm

El Heraldo. (2023). ¿Como definir el cluster de audiencia ideal según el objetivo de su marca? *Diario el Heraldo*, 1.

Fúnez, K. (10 de 06 de 2023). Clústeres. *Maestria en Analítica de Negocios*. Tegucigalpa, Francisco Morazán, Honduras: Unitec.

Gómez Munar, J. A., Páez Casallas, L. S., & Estrada Rodríguez, P. A. (2023). *Teoría del comportamiento del consumidor de Kotler en el uso de medio de pagos en Colombia.* Universidad EAN.

Gonzalez Marcos, A. (2006). *Desarrollo de técnicas de minería de datos.* Logroño: Universidad de La Rioja.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la investigación* (6a. ed ed.). México D.F: McGraw-Hill.

IBM Corp. (2019). IBM SPSS Statistics for Windows, Version 26.0. . *Software*. Armonk, NY: IBM Corp.

IBM Corp. (2023). *Análisis de datos exploratorio*. www.ibm.com.

KNIME. (2023). *Knime.com*.

La Prensa. (2022). ¿Que significan los clusters de audiencia para la comprensión de su marca? *Diario La Prensa*, 1.

Lafosse, N. F. (2016). *Extracción de patrones semanticamente distintos a partir de los datos almacenados en la plataforma Paideia.* San Miguel: Pontifica Universidad Católica del Perú.

LATAM, S. (31 de Mayo de 2023). *Salesforce LATAM Blog*. Salesforce LATAM Blog: http://salesforce.com

Lazzari, L. (2018). *La Segmentación de mercados mediante la teoría de la afinidad.*

López, A. J., Caicedo, C. G., & Oviedo, P. C. (2017). Análisis Multivariado Para Segmentación de Clientes Basada en FRM. *Tecnura*, 54.

Málaga, J. A. (2003). *Euroinnova International Online Education*. Euroinnova International Online Education: www.euroinnova.pe

Mamaní Rodríguez, Z., Del Pino Rodríguez, L., & Cortez Vasquez, A. (2017). Minería de datos distribuida usando clustering k-means en. *Revista Industrial Data*, 123-129.

Maradiaga Fernández, C. J., Lao León, Y. O., Curra Sosa, D. A., & Martín, R. L. (2022). Empleo de algoritmos KNN en metodología multicriterio para la clasificación de clientes, como sustento de la planeación agregada. *Retos de la Dirección, 16*(1), 178-198.

Meyer, Sandro. (04 de Agosto de 2020). *growthbay*. Personalisation according to Netflix: https://www.growthbay.ch/blog/personalisation-netflix

Microsoft. (2023). *support.microsoft.com*. https://support.microsoft.com/

Moreno, J. (07 de noviembre de 2022). *Cómo crear un perfil de cliente paso a paso (con ejemplos)*. GESTIÓN DE CLIENTES : blog.hubspot.es

Mulford, R., Callejas, D. E., & Mendoza, A. A. (2022). *Segmentación de clientes de una empresa utilizando recencia, frecuencia, monto (rfm) y métodos de clustering.* Bogota Colombia: DSpace software.

Munar, Perez. (04 de Mayo de 2023). *¿Qué es el clustering? Cómo segmentar a tu audiencia en marketing digital*. Cyberclick: https://www.cyberclick.es/numerical-blog/que-es-el-clustering-como-segmentar-a-tu-audiencia-en-marketing-digital#:~:text=A%20diferencia%20de%20la%20segmentaci%C3%B3n,m%C3%A1s%20%C3%BAtil%20para%20la%20marca.

Noreña, F. C. (2015). *Segmentación de clientes automatizada a partir de tecnicas de minería de datos (k-means Clustering).* Antioquía: Escuela de Ingeniería de Antioquía.

Pérez, J. A. (2007). LAS VARIABLES EN EL MÉTODO CIENTÍFICO. 171-177.

Porter, M. (1989). *La ventaja competitiva de las naciones.* Vergara.

Rivera, S. I. (2015). Big Data Marketing: una aproximación. *Revista Perspectivas*, 147-158. https://doi.org/ISSN 1994-3733

Rodríguez Suárez, Y., & Díaz Amador, A. (2009). Herramientas de Minería de Datos. *Revista Cubana de Ciencias Informáticas*, 73-80.

Rodríguez, A. R., & Gallardo, J. C. (2020). *Uso del algoritmo Adaboost y la regresión logística para la predicción de fuga de clientes en una empresa de telefonía móvil.* Lima, Perú: Universidad Nacional Agraria la Molina.

Rodríguez, H. (21 de 02 de 2022). *CREHANA*. crehana.com: https://www.crehana.com/blog/negocios/clusterizacion/

Rojas, E. M. (2020). Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo. *Revista Ibérica de Sistemas e Tecnologias de Informacion*, 586–599.

Universidad Internacional de Valencia. (29 de 08 de 2022). *Claustering: ¿Qué es y qué aplicaciones tiene?* https://www.universidadviu.com/es/actualidad/nuestros-expertos/claustering-que-es-y-que-aplicaciones-tiene

Vallejo Ballesteros, H., Guevara Iñiguez, E., & Medina Velasco, S. (2018). Minería de Datos. *Revista Científica Mundo de la Investigación y el Conocimiento*, 339-349.

Zelaya, O. (23 de Abril de 2021). *HONDURAS – La protección de datos en Honduras*. HONDURAS – La protección de datos en Honduras: https://central-law.com

ANEXOS

ANEXO 1: CARTA DE AUTORIZACIÓN DE LA EMPRESA O INSTITUCIÓN

Tegucigalpa MDC, Francisco Morazán, 27 / 10 / 2023
(Ciudad), (Departamento) (Día, mes y año)

Allan Geovany Ventura Diaz
(Nombre y apellidos del Director o Gerente)

Gerente de Operaciones
(Puesto Laboral)

Compañia Avícola de Centroamérica S.A de C.V. CADECA
(Empresa o Institución)

Blvd. FFAA, Contiguo a Plaza Millenium, entrada Col Tiloarque
(Dirección principal de la empresa o institución)

Estimado Señor(a): Allan Geovany Ventura Díaz

Reciba un cordial y atento saludo. Por medio de la presente deseamos solicitar su apoyo, dado que somos alumnos de UNITEC y nos encontramos desarrollando el Trabajo Final de Graduación previo a obtener nuestro título de maestría en Analítica de Negocio., Hemos seleccionado como tema Implementación de modelo de clusterización para segmentación de clientes, por lo que estaríamos muy agradecidos de contar con el apoyo de la empresa que usted representa para poder desarrollar nuestra investigación. En particular, dicha solicitud se circunscribe a peticionar que se nos autorice a realizar: Trabajo final de graduación.

(encuestas, sondeos, etc).

A la espera de su aprobación, me suscribo de

Usted. Atentamente,

Ana Carolina Carrillo
Firma, nombre y apellidos
No. de cuenta: 12213222

Emili Gisselle Flores
Firma, nombre y apellidos
No. de cuenta: 12213126

Por este medio, Compañia Avícola de Centroamérica S.A de C.V. CADECA
(empresa / institución),

Autoriza la realización dentro de sus instalaciones el proyecto de investigación de Postgrado antes mencionado.

Allan Geovany Ventura Díaz
Nombre y sello del Director / Gerente

Vo.Bo.

ANEXO 2: MODELO UTILIZADO EN KNIME