****

**UNIVERSIDAD MAYOR DE SAN SIMÓN**

FACULTAD DE CIENCIAS Y TECNOLOGÍA

DIRECCIÓN DE POSGRADO

**DIPLOMADO CIENCIA DE DATOS**

**SEGUNDA VERSIÓN**

**MODELO DE SEGMENTACIÓN PARA ENTENDER EL COMPORTAMIENTO DE COMPRAS DE CLIENTES EN UNA EMPRESA COMERCIALIZADORA DE PRODUCTOS ELECTRÓNICOS**

**PROYECTO PRESENTADO PARA OBTENER EL GRADO DE LICENCIATURA EN INGENIERÍA DE SISTEMAS**

**MODALIDAD DOBLE TITULACIÓN**

**POSTULANTE : ANDRES MOSCOSO MENA**

**TUTOR : M.SC. ING. DANNY LUIS HUANCA SEVILLA**

**Cochabamba – Bolivia**

**2025**

MODELO DE SEGMENTACIÓN PARA ENTENDER EL COMPORTAMIENTO DE COMPRAS DE CLIENTES EN UNA EMPRESA COMERCIALIZADORA DE PRODUCTOS ELECTRÓNICOS

Por

Andres Moscoso Mena

El presente documento, Trabajo de Grado es presentado a la Dirección de Posgrado de la Facultad de Ciencias y Tecnología en cumplimiento parcial de los requisitos para la obtención del grado académico de Licenciatura en Ingeniería de Sistemas, modalidad Doble Titulación, habiendo cursado el Diplomado “Ciencia de Datos” propuesta por el Centro de Estadística Aplicada (CESA) en su segunda versión.

ASESOR/TUTOR

M.Sc. Ing. Danny Luis Huanca Sevilla

COMITÉ DE EVALUACIÓN

Ing. M.Sc. Ronald Edgar Patiño Tito. (Presidente)

Ing. M.Sc. Guillen Salvador Roxana. (Coordinador)

Ing. M.Sc. Henrry Frank Villarroel Tapia. (Tribunal)



**DIRECCIÓN DE POSGRADO, FACULTAD DE CIENCIAS Y TECNOLOGIA**

Cochabamba, Bolivia

**Aclaración**

**Este documento describe el trabajo realizado como parte del programa de estudios de Diplomado “Ciencia de Datos” en el Centro de Estadística Aplicada CESA y la Dirección de Posgrado de la Facultad de Ciencias y Tecnología. Todos los puntos de vista y opiniones expresadas en el mismo son responsabilidad exclusiva del autor y no representan necesariamente las de la institución.**

**Resumen**

El siguiente proyecto presenta el desarrollo de un modelo de segmentación de aprendizaje no supervisado para lograr entender el comportamiento de compras de los clientes de una empresa comercializadora de productos electrónicos. Hoy en día, muchas de las empresas recurren a la tecnología para lograr anticiparse a futuras necesidades, especialmente tras el crecimiento de esta durante y después de la pandemia. Una de las herramientas más destacadas es la Inteligencia Artificial, que cada vez se integra más en la vida cotidiana, simplificando las tareas. Una de sus ramas, el Machine Learning, permite a las máquinas aprender de datos y experiencias, siendo bastante útil en marketing, ya que ayuda en la identificación de patrones y segmentación del público objetivo. Esta división permite optimizar campañas, prevenir pérdidas y personalizar la experiencia del consumidor, además de automatizar procesos, ahorrando recursos y mejorando el análisis de los mismos.

Para la ejecución de dicho modelo, se procedió a realizar una preparación de los datos obtenidos de la empresa Drustvo S.R.L., la cual abarcaba registros de ventas del 2021 al 2023. A lo largo del proyecto se hizo uso del análisis RFM, para lograr identificar medidas como la recencia, frecuencia, y valor monetario de las compras de los clientes. Valores que también fueron implementados en tres modelos de aprendizaje no supervisado, K-Means, Clustering Jerárquico y K-Medoids, siendo este último el que presento mejores resultados, al obtener una división de tres clústeres visiblemente definida, con la cuál pudo realizarse una evaluación más a fondo sobre el comportamiento de compra de los clientes.

Una vez realizado el análisis pertinente, tanto de los datos provistos como de los resultados obtenidos por el modelo K-Medoids. Se pudo determinar que la empresa gozó de un incremento exponencial en sus ganancias a lo largo de los tres años que abarcaba el dataset, siendo el año 2023 el que registró más ingresos. Obteniendo una segmentación en 3 grupos, identificados como 126 Clientes Nuevos (52,07%), 106 Clientes Perdidos (43,8%) y 10 Clientes con Potencial (4,13%) de los clientes respectivamente. A pesar de esto, la mayoría de los clientes, se encuentran entre grupos de baja frecuencia y valor monetario, significando que la base de sus ganancias se debe a clientes esporádicos, con un bajo porcentaje de clientes que realizan compras frecuentes y recientes.

**Palabras clave**

Aprendizaje No Supervisado, Análisis RFM, Segmentación de Clientes, K-Medoids, Entendimiento de compra.

*Dedicado a mis padres, quienes siempre depositaron su confianza y apoyo permanente, impulsándome a alcanzar mis metas. A mi madre, por su amor incondicional y dedicación abnegada a lo largo de toda mi vida, todo lo que soy lleva tu fuerza como base. A mi padre, por sus desvelos compartidos en el estudio, aunque reservado, tus consejos sirvieron de orientación en mis decisiones.*

*A mis abuelos, por ser siempre una guía, ejemplo y fuentes de inspiración en valores, espiritualidad y moral. A ellos les debo todo lo que tengo, y la persona que aspiro a ser.*

**Agradecimientos**

*Al M.Sc. Ing. Danny Luis Huanca Sevilla, por su orientación y ayuda en la redacción de este proyecto, así como el tiempo brindado en la colaboración en el papel de tutor.*

*Al plantel docente del Diplomado de Ciencia de Datos 2da Versión, por su compromiso con la enseñanza y el conocimiento otorgado para la formación de nuevos profesionales.*

*A la empresa Drustvo S.R.L. por la información brindada para la elaboración de este proyecto.*

**Tabla de contenidos**

[1 Introducción 1](#_Toc198645827)

[1.1 Antecedentes 2](#_Toc198645828)

[1.2 Justificación 3](#_Toc198645829)

[1.3 Planteamiento del problema 3](#_Toc198645830)

[1.4 Objetivo general 4](#_Toc198645831)

[1.4.1 Objetivos específicos 4](#_Toc198645832)

[2 Marco teórico 5](#_Toc198645833)

[2.1 Inteligencia Artificial 5](#_Toc198645834)

[2.2 Machine Learning 6](#_Toc198645835)

[2.2.1 Etapas de Machine Learning 6](#_Toc198645836)

[2.3 Tipos de Aprendizaje de Machine Learning 8](#_Toc198645837)

[2.4 Aprendizaje No Supervisado 8](#_Toc198645838)

[2.4.1 K-Means 9](#_Toc198645839)

[2.4.2 Clustering Jerárquico 10](#_Toc198645840)

[2.4.3 K-Medoids 10](#_Toc198645841)

[2.5 Análisis RFM 11](#_Toc198645842)

[3 Marco metodológico 13](#_Toc198645843)

[3.1 Área de estudio 13](#_Toc198645844)

[3.2 Flujograma metodológico 13](#_Toc198645845)

[3.3 Fuentes de información 15](#_Toc198645846)

[3.4 Herramientas usadas 16](#_Toc198645847)

[3.5 Preparación de los datos 16](#_Toc198645848)

[3.5.1 Información de las variables 16](#_Toc198645849)

[3.5.2 Limpieza de datos 17](#_Toc198645850)

[3.5.3 Tabla de Clientes 18](#_Toc198645851)

[3.5.4 Tabla de Productos 20](#_Toc198645852)

[3.5.5 Tabla de Ventas 21](#_Toc198645853)

[3.5.6 Modelo Relacional de Tablas 22](#_Toc198645854)

[3.6 Análisis RFM 23](#_Toc198645855)

[3.6.1 Tabla de valores RFM 23](#_Toc198645856)

[3.6.2 Cálculo de valores RFM 25](#_Toc198645857)

[3.7 Elaboración de los modelos de aprendizaje no supervisado 28](#_Toc198645858)

[3.7.1 Algoritmo K-Means 31](#_Toc198645859)

[3.7.2 Algoritmo de Clúster Jerárquico 32](#_Toc198645860)

[3.7.3 Algoritmo K-Medoids 33](#_Toc198645861)

[3.8 Validación del entendimiento de compra con la segmentación obtenida 34](#_Toc198645862)

[4 Análisis de Resultados y Discusión 36](#_Toc198645863)

[4.1 Resultados de la preparación de datos 36](#_Toc198645864)

[4.2 Resultados del análisis RFM 37](#_Toc198645865)

[4.3 Resultados de la elaboración de los modelos de aprendizaje no supervisado 39](#_Toc198645866)

[4.4 Resultados de la validación del entendimiento de compras de los clientes 41](#_Toc198645867)

[4.5 Discusión de resultados 43](#_Toc198645868)

[5 Conclusiones 46](#_Toc198645869)

[6 Recomendaciones 48](#_Toc198645870)

[7 Bibliografía 49](#_Toc198645871)

[Anexos 51](#_Toc198645872)

[Anexo 1. Archivo Excel con los registros de ventas de la empresa 51](#_Toc198645873)

[Anexo 2. Tabla de las variables cuantitativas 52](#_Toc198645874)

[Anexo 3. Tabla de las variables cualitativas 52](#_Toc198645875)

[Anexo 4. Tabla de Clientes obtenida de la preparación de datos 53](#_Toc198645876)

[Anexo 5. Tabla de Productos obtenida de la preparación de datos 54](#_Toc198645877)

[Anexo 6. Tabla de Ventas obtenida de la preparación de datos 55](#_Toc198645878)

[Anexo 7. Modelo Relacional de las tablas Clientes, Productos y Ventas 56](#_Toc198645879)

[Anexo 8. Tabla del cálculo de los valores RFM 56](#_Toc198645880)

[Anexo 9. Tabla de los intervalos y puntajes RFM 57](#_Toc198645881)

[Anexo 10. Tabla de los tipos de clientes basados en los puntajes RFM 57](#_Toc198645882)

[Anexo 11. Tabla de Entrenamiento para la aplicación de los modelos 58](#_Toc198645883)

[Anexo 12. Tabla minable para el entrenamiento de los modelos 59](#_Toc198645884)

[Anexo 13. Gráfica de matriz de correlación 60](#_Toc198645885)

[Anexo 14. Gráfica de histogramas de recency, frequency y monetary 60](#_Toc198645886)

[Anexo 15. Código de la eliminación de outliers y re-escalado de datos 61](#_Toc198645887)

[Anexo 16. Código de cálculo de clústeres para K-Means 62](#_Toc198645888)

[Anexo 17. Valores obtenidos del Método del Codo y Silhoutte Score para K-Means 63](#_Toc198645889)

[Anexo 18. Modelado de K-Means 63](#_Toc198645890)

[Anexo 19. Resultados del modelo K-Means 64](#_Toc198645891)

[Anexo 20. Código de cálculo de clústeres para Clustering Jerárquico 64](#_Toc198645892)

[Anexo 21. Valores obtenidos del Método del Codo y Silhoutte Score para Clustering Jerárquico 65](#_Toc198645893)

[Anexo 22. Modelado de Clustering Jerárquico 65](#_Toc198645894)

[Anexo 23. Resultados del modelo Clustering Jerárquico 66](#_Toc198645895)

[Anexo 24. Código de cálculo de clústeres para K-Medoids 67](#_Toc198645896)

[Anexo 25. Valores obtenidos del Método del Codo y Silhoutte Score para K-Medoids 68](#_Toc198645897)

[Anexo 26. Modelado de Clustering K-Medoids 68](#_Toc198645898)

[Anexo 27. Resultados del modelo K-Medoids 69](#_Toc198645899)

[Anexo 28. Tabla obtenida con los clústeres de K-Medoids 69](#_Toc198645900)

[Anexo 29. Tabla de unión con los resultados obtenidos del modelo K-medoids 70](#_Toc198645901)

[Anexo 30. Gráficas obtenidas de la preparación de datos (Hoja 1) 71](#_Toc198645902)

[Anexo 31. Gráficas obtenidas de la preparación de datos (Hoja 2) 72](#_Toc198645903)

[Anexo 32. Gráficas obtenidas del análisis RFM 73](#_Toc198645904)

[Anexo 33. Gráficas obtenidas de la validación del entendimiento de compra 74](#_Toc198645905)

[Anexo 34. Gráficas obtenidas de la discusión de resultados 75](#_Toc198645906)

[Anexo 35. Carta de aprobación del tutor Carta de aprobación del tutor 76](#_Toc198645907)

[Anexo 36. Código QR de enlace al repositorio del proyecto 77](#_Toc198645908)

[Anexo PRINCIPAL: CD 77](#_Toc198645909)

**Lista de figuras**

[Figura 1‑1: Árbol de problema 4](#_Toc198646020)

[Figura 2‑1: Relaciones sinérgicas entre Ciencia de Datos, Probabilidad Estadística, y Machine Learning 6](#_Toc198646021)

[Figura 2‑2: Etapas de Machine Learning 7](#_Toc198646022)

[Figura 2‑3: División de puntos de datos en clústeres similares (izquierda). Grupos de clústeres con demarcaciones naturales (derecha) 9](#_Toc198646023)

[Figura 3‑1: Imagen de la ubicación y fachada de la empresa Drustvo S.R.L. 13](#_Toc198646024)

[Figura 3‑2: Flujograma 14](#_Toc198646025)

[Figura 3‑3: Vista de tabla de registros del archivo Excel 15](#_Toc198646026)

[Figura 3‑4: Vista de la tabla en Power BI 16](#_Toc198646027)

[Figura 3‑5: Columnas "categoría" y "marca", ambas con un solo valor 18](#_Toc198646028)

[Figura 3‑6: Vista de tabla una vez terminada la limpieza de datos 18](#_Toc198646029)

[Figura 3‑7: Columnas filtradas para la tabla Clientes 19](#_Toc198646030)

[Figura 3‑8: Casos especiales con mismo "nit", pero nombres diferentes 19](#_Toc198646031)

[Figura 3‑9: Comandos usados para la corrección de los registros 20](#_Toc198646032)

[Figura 3‑10: Métricas de las columnas posterior a la corrección 20](#_Toc198646033)

[Figura 3‑11: Vista final de la tabla Clientes 20](#_Toc198646034)

[Figura 3‑12: Métricas de la tabla Productos luego de la eliminación de duplicados 21](#_Toc198646035)

[Figura 3‑13: Vista final de la tabla Productos 21](#_Toc198646036)

[Figura 3‑14: Vista inicial de tabla Ventas 22](#_Toc198646037)

[Figura 3‑15: Tabla Ventas con las columnas “id\_Cliente”, "precio" y "total" añadidas 22](#_Toc198646038)

[Figura 3‑16: Relaciones entre las tablas Cliente, Productos, Ventas y Calendario 23](#_Toc198646039)

[Figura 3‑17: Vista del modelo relacional 23](#_Toc198646040)

[Figura 3‑18: Columnas con los datos necesarios para la elaboración de la tabla RFM 24](#_Toc198646041)

[Figura 3‑19: Creación de tabla RFM 24](#_Toc198646042)

[Figura 3‑20: Vista final de tabla RFM 25](#_Toc198646043)

[Figura 3‑21: Creación de columna "recencyPoints" 26](#_Toc198646044)

[Figura 3‑22: Creación de columna "frecuencyPoints" 26](#_Toc198646045)

[Figura 3‑23: Creación de columna "monetaryPoints" 27](#_Toc198646046)

[Figura 3‑24: Vista final de la tabla “dim\_RFM” 27](#_Toc198646047)

[Figura 3‑25: Vista de la tabla Entrenamiento desde DAX Studio 28](#_Toc198646048)

[Figura 3‑26: Vista de librería importadas para el modelado de los algoritmos 29](#_Toc198646049)

[Figura 3‑27: Métricas de la tabla de entrenamiento 29](#_Toc198646050)

[Figura 3‑28: Matriz de correlación de las variables "recency", "frequency" y "monetary" 30](#_Toc198646051)

[Figura 3‑29: Gráfica de histogramas de las 3 variables 31](#_Toc198646052)

[Figura 3‑30: Método del Codo (izquierda) y Silhouette Score (derecha) 31](#_Toc198646053)

[Figura 3‑31: Clústeres mediante K-Means 32](#_Toc198646054)

[Figura 3‑32: Grafico de Dendrograma (izquierda) y Silhouette Score (derecha) 32](#_Toc198646055)

[Figura 3‑33: Resultados de Clúster Jerárquico 33](#_Toc198646056)

[Figura 3‑34: Gráficos de Método del Codo y Silhouette Score 33](#_Toc198646057)

[Figura 3‑35: Resultado de K-Medoids 34](#_Toc198646058)

[Figura 3‑36: Obtención de clústeres y conteo de clientes por clúster 34](#_Toc198646059)

[Figura 3‑37: Tabla Merge, para el análisis de los tipos de cliente 35](#_Toc198646060)

[Figura 4‑1: Cantidad de Ventas por Año (izquierda), Ganancias Totales por Año (derecha) 36](#_Toc198646061)

[Figura 4‑2: Ganancias Totales por Trimestre de cada Año 37](#_Toc198646062)

[Figura 4‑3: Top 5 de los Productos Más Vendidos y con Mayor Cantidad de Unidades Vendidas 37](#_Toc198646063)

[Figura 4‑4: Representacion percentil de la cantidad de cada tipo de cliente 38](#_Toc198646064)

[Figura 4‑5: Promedio de Puntajes RFM para Clientes “En Riesgo” y “Perdidos” 38](#_Toc198646065)

[Figura 4‑6: Promedio de Puntajes RFM para Clientes "Alto Valor", "Ideal", "Leales" y "Nuevos" 39](#_Toc198646066)

[Figura 4‑7: Segmentación mediante K-Means y Clustering Jerárquico con K=2 40](#_Toc198646067)

[Figura 4‑8: Clustering mediante K-Medoids, K=3 40](#_Toc198646068)

[Figura 4‑9: Cantidad de Clientes Pertenecientes a Cada Cluster obtenido por K-Medoids 41](#_Toc198646069)

[Figura 4‑10: Porcentaje Atribuido a cada Tipo de Cliente 42](#_Toc198646070)

[Figura 4‑11: Porcentajes de clientes por clúster 43](#_Toc198646071)

[Figura 4‑12: Relación de Promedios del Proyecto Comparativo 44](#_Toc198646072)

[Figura 4‑13: Relación de Promedios de los valores RFM 45](#_Toc198646073)

**Lista de tablas**

[Tabla 3‑1: Descripción de las variables halladas en el dataset 15](#_Toc198646074)

[Tabla 3‑2: Tabla de Variables Cuantitativas 17](#_Toc198646075)

[Tabla 3‑3: Tabla de Variables Cualitativas 17](#_Toc198646076)

[Tabla 3‑4: Cálculo de rango, intervalo y amplitud para RFM 25](#_Toc198646077)

[Tabla 3‑5: Tabla de puntajes para las variables RFM 26](#_Toc198646078)

[Tabla 3‑6: Combinaciones y descripción de posibles tipos de clientes según el análisis RFM 28](#_Toc198646079)

[Tabla 4‑1: Valores Máximos y Mínimos de RFM 42](#_Toc198646080)

[Tabla 4‑2: Valores obtenidos en el presente proyecto 43](#_Toc198646081)

[Tabla 4‑3: Valores del Proyecto Comparativo 44](#_Toc198646082)

# Introducción

En la actualidad, las empresas de diferentes ámbitos suelen buscar un apoyo en la tecnología, debido al gran crecimiento de la misma durante y después de la pandemia, con la intención de permitirles anticiparse a las necesidades que puedan surgir en un futuro (Vaidya, 2022). Entre estas tecnologías, una de las que presenta mayor crecimiento en estos últimos años es la Inteligencia Artificial, la cual no solo ayuda minimizando la complejidad de diferentes tareas, también se vuelve una herramienta de uso cada vez más cotidiano en la vida de las personas, ya que la misma puede ser aplicada en diferentes aspectos (Hernández, 2022).

Es así como llegamos a una de sus ramas, la cual es el Machine Learning, cuyo principal objetivo es permitir a las computadoras lograr emular la manera en que los humanos aprendemos, realizamos tareas y mejoramos en base a las experiencias vividas, cuyo equivalente en este caso, se vería reflejado en bases de datos procesados (Oracle, s.f.). Esta herramienta ha demostrado su valía, sobre todo en áreas como el marketing y publicidad dentro de las empresas, puesto que puede reconocer patrones entre los datos cargados, que puedan permitirle pronosticar comportamientos los cuales podrían llegar a ser usados para la segmentación del público objetivo (Jiménez, 2023).

La importancia de la segmentación del público objetivo de la empresa o bien clientes, yace en la posibilidad de personalizar la experiencia de cada grupo hallado, para poder satisfacer y llegar a cumplir las necesidades que puedan tener o solicitar de los productos o servicios que brinda la empresa (Smolic, 2024). De esta manera se puede separar en diferentes grupos a los clientes, en base a uno o más criterios, lo cual puede llegar a servir a la empresa, para realizar diferentes tipos de campañas dirigidas de marketing, encontrar características especiales dentro las compras de los clientes, analizar su comportamiento e identificar patrones que permitan prevenir perdidas o abandonos de servicio, o incluso toma de decisiones en cuanto al reabastecimiento de productos (AnalytixLabs, 2024).

Otros beneficios que aporta el uso de machine learning en la segmentación de clientes, es que siendo una tarea que consume bastante tiempo realizarla de forma manual, mediante el entrenamiento de un modelo destinado a realizar esta tarea, se puede ahorrar bastantes recursos y dirigir este esfuerzo humano a otro tipo de problemas de más importancia, además este modelo se puede ir adaptando y actualizando a medida que sea necesario para que siga presentando resultados eficientes, así como una mejor precisión en su categorización (Kumar, 2023).

## Antecedentes

Como se ha mencionado en el punto anterior, la importancia de la segmentación, en este caso de los clientes en una empresa, tiene una gran relevancia para un análisis para la toma de decisiones, y de esta forma crear planes y estrategias que puedan beneficiar en la venta y comercio de los productos ofrecidos (Vaidya, 2022).

Bajo este precepto es que existen diferentes proyectos dirigidos a la implementación y uso de machine learning y modelos de aprendizaje entrenados para hacer la respectiva segmentación. Por ejemplo, el proyecto titulado “Segmentación de Clientes de una Empresa Comercializadora de Productos de Consumo Masivo en la Ciudad de Popayán Soportado en Machine Learning y Análisis RFM (Recency, Frecuency y Money)” desarrollado por Fabian Palacios y Nelson Pastor, busca lograr segmentar a los clientes de una empresa de productos lácteos para generar estrategias de marketing que beneficien a la misma, mediante el uso del modelo RFM y del algoritmo de machine learning denominado K-Means.

En este caso se hizo una comparativa entre la cantidad de segmentación obtenida usando ambos métodos, obteniendo 5 segmentaciones por RFM y 7 por K-Means (Palacios Abadía & Pastor Patiño, 2020). En este proyecto, definieron usar el algoritmo K-Means, ya que es el más usado al momento de identificar segmentos, como es mencionado en una monografía a la que hacen referencia, en la que enfatizan y demuestran, usando 6 datasets diferentes, la adaptabilidad del algoritmo para obtener resultados beneficiosos con el uso de la información proporcionada (Wagstaff, Cardie, Rogers, & Schroedl, 2001). Concluyendo que los resultados obtenidos por el algoritmo permiten resultados más eficaces y rápidos en comparación con el modelo RFM, pues el algoritmo es totalmente autónomo en cuanto a sus cálculos.

En la monografía “Segmentación de Clientes Mediante Análisis de Patrones de Compra para la optimización de Estrategias Comerciales” escrito por José Berrio y Orlando Olea, abordan la necesidad de crear estrategias de basadas en las compras realizadas por diferentes clientes de una empresa comercializadora para la optimización del tiempo de los empleados, también haciendo uso del algoritmo K-Means para la formación de clústeres (Berrio Lasprilla & Olea Gómez, 2024). En este caso el algoritmo define 3 tipos de clústeres, bajo los que los autores determinan diferentes enfoques para su análisis, concluyendo que el algoritmo que encontraron más adecuado para una distribución más equilibrada es el K-Means RBF.

Otro trabajo interesante es el de Ana Carrillo y Emili flores, titulado “Implementación de un Modelo de Clusterización Mediante la Segmentación de Perfil de Clientes para Corporación Multi Inversiones”, el cual también hace uso del modelo RFM para la clasificación de sus clientes en base a variables como la recencia, frecuencia y valor monetario. También hace uso de los modelos K-Means y K- Meloid, de entre los cuales el primero muestra resultados más eficientes, puesto que sus resultados presentan una mejor precisión al realizar la agrupación de los clientes (Carrillo García & Flores Velásquez, 2024).

## Justificación

En los últimos años existe un mayor interés no solo en el exterior, también dentro de nuestro país, Bolivia, por el uso e implementación de la inteligencia artificial, la cual ha sido considerada como un aspecto fundamental en el progreso hacia el futuro para las empresas. No solo hay un interés, también existen conferencias y charlas que motivan e impulsan estas ideas, ya que el uso de estas novedosas tecnologías es sinónimo de mejora en la eficiencia y la productividad dentro de las instituciones empresariales (Ecofinanzas & El Deber, 2023).

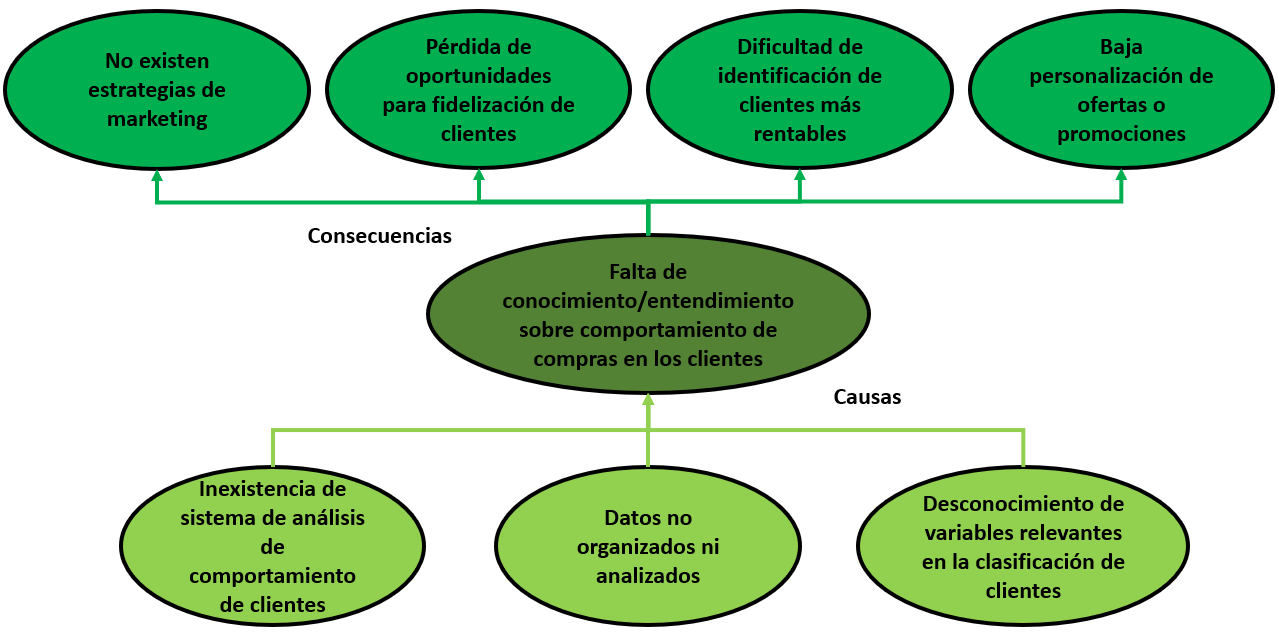
Por tanto, la intención que motiva realizar este proyecto es, poder apoyar a la empresa comercializadora, de cuyos datos se hará uso para el entrenamiento de los modelos, para poder en este caso hacer un análisis de las ventas de sus diferentes productos a sus distintos clientes, de esta forma se desea poder obtener un modelo de machine learning que pueda segmentar y diferenciar a sus clientes basándose en diferentes patrones que puedan ser observados o interpretados en los registros de ventas. Así con esta información, el sector encargado de marketing de la empresa pueda tomar las decisiones correspondientes, con la intención de beneficiar a la entidad comercializadora, y de esta manera presentar un mejor servicio e implementación de posibles descuentos a sus diferentes clientes.

## Planteamiento del problema

Actualmente la empresa comercializadora con la que se trabaja, no cuenta con una división evidente entre sus distintos clientes, ni una condición especifica que permita determinar descuentos que pueden llegar a ser aplicados en sus distintas compras, más allá de una afinidad por la lealtad que algunos clientes muestran hacia con la empresa, o en otros casos, debido a compras de grandes montos de uno o varios productos.

Esto puede llegar a influir en toma de decisiones comerciales, dificultad en futuras campañas enfocadas en marketing, premiar de mejor manera la fidelidad de los clientes o crear una experiencia personalizada que motive a los clientes a permanecer afines a la empresa y el servicio que brinda.

Ante esta situación, se plantea una necesidad por la segmentación de clientes mediante el uso de machine learning, en específico de un modelo de aprendizaje no supervisado, el cual permita transformar los datos de los registros obtenidos, en conocimiento y una herramienta capaz de contribuir en las diferentes estrategias de la empresa, lo que permitirá a la misma subsanar desventajas que pueda tener por falta de uso de tecnologías más recientes para un estudio de sus datos. En base a esto se determinó un árbol de problemas con causas y consecuencias en relación a la problemática principal planteada como la “Falta de conocimiento/entendimiento sobre comportamiento de compras en los clientes”, como puede verse en la Figura 1-1.



**Figura 1‑1: Árbol de problema**

**Fuente: (Elaboración propia, 2025)**

## Objetivo general

Desarrollar un modelo de segmentación para entender el comportamiento de compras de clientes en una empresa comercializadora de productos electrónicos.

### Objetivos específicos

* Realizar la preparación de los datos provistos por la empresa comercializadora, respecto a sus ventas dentro de los años 2021-2023.
* Realizar el análisis RFM en los datos para obtener puntuaciones con las cuales se aplique algoritmos de modelos de aprendizaje no supervisado de machine learning, para lograr obtener la segmentación de los clientes.
* Elaborar modelos de aprendizaje no supervisado para poder identificar y englobar a los clientes en base a los valores RFM obtenidos, para seleccionar el más adecuado para su análisis.
* Validar que los resultados del modelo de segmentación elegido, lograron mejorar el entendimiento de compra de clientes, para poder realizar recomendaciones necesarias para el uso de la empresa en futuras decisiones o planes de marketing.

# Marco teórico

## Inteligencia Artificial

La inteligencia Artificial no tiene en sí una definición estándar, pero si puede ser expresada como la aptitud de una máquina de poder desenvolverse de la misma manera en la que lo hace el pensamiento humano para poder ejecutar los procesos que se necesitan para el aprendizaje y reconocimiento, esto debido a la comparativa entre el intelecto de una máquina y la de los humanos (Sosa Sierra, 2007).

Siendo parte de una de las ramas de la computación, la inteligencia artificial representa la simulación de los procesos llevados a cabo para el razonamiento humano, que conlleva al aprendizaje y experiencia obtenida por diferentes acciones (Ponce Gallegos, y otros, 2014).

La inteligencia artificial inicialmente se fue construyendo con base en entendimientos e hipótesis de existencia previa en otras áreas como ser la matemática, la psicología, las ciencias de la computación, la lingüística, la filosofía, la economía y la neurociencia. De esta forma, estas diferentes áreas fueron aportando mediante sus herramientas y competencias en el desarrollo de esta nueva área de aprendizaje.

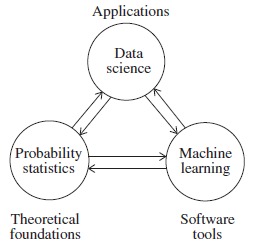
Podemos mencionar los diferentes aportes que cada ciencia mencionada ofrecieron al área de estudio de la inteligencia artificial, por ejemplo:

* **La filosofía** planteaba los cimientos en los que se basa para su funcionamiento, siendo que el pensamiento era la herramienta que definía cuál sería la conducta correcta para poder iniciar.
* **La matemática**, por otro lado, ofrece las herramientas necesarias para la veracidad lógica dentro de las probabilidades dadas, así como el razonamiento de los algoritmos.
* **La psicología** en su momento, respaldó la idea de que los humanos y los animales pueden ser vistos como sistemas que procesan la información, lo cual puede verse reflejado en estudios como el cognitivismo y la psicología cognitiva, ambas centradas en el estudio de los procesos internos del pensamiento, la toma de decisiones, etc.
* **Las ciencias de la computación** por su parte, implementan teorías de IA, mediante el uso de modelos cognitivos y sus artefactos, los cuales no podrían ser viables si no fuera por los avances en velocidad y memoria que provienen de la tecnología computacional.
* **La lingüística**, esta área es clave en la representación del conocimiento, ya que gracias a ella se forma la lingüística computacional o procesamiento del lenguaje natural, lo que permite que las computadoras comprendan e interpreten el lenguaje humano.
* **La economía** influye en la toma de decisiones enfocadas en las ganancias o pérdidas. Entre las teorías más conocidas están la Teoría de la Decisión, de Juegos, y los Procesos de Decisión de Markov.
* **Y la neurociencia** aportando en los conocimientos reunidos sobre el funcionamiento del cerebro para el procesamiento de la información, el cual la IA trata de imitar mediante diferentes algoritmos (Redes Neuronales).

## Machine Learning

Podemos definir Machine Learning (Aprendizaje Automático en español) como una rama de la inteligencia artificial, la cual consiste en la creación de programas (software) que, basados en el análisis de datos, generan predicciones mediante el uso de fórmulas y métodos creados por expertos en las áreas de la tecnología y la matemática. En otras palabras, forma parte de la programación que se está dedicada a la construcción de herramientas para el análisis de datos que ayudan a cumplir metas dentro del área de la ciencia de datos (Saleh, Majzoub, & Saleh, 2025).

Cabe mencionar la conexión que esta rama tiene con las disciplinas de la ciencia de datos y la probabilidad y estadística la cual se puede ver en la Figura 2-1,ya que dependen unas de otras ayudando a crear modelos para la resolución de problemas más específicos.



**Figura 2‑1: Relaciones sinérgicas entre Ciencia de Datos, Probabilidad Estadística, y Machine Learning**

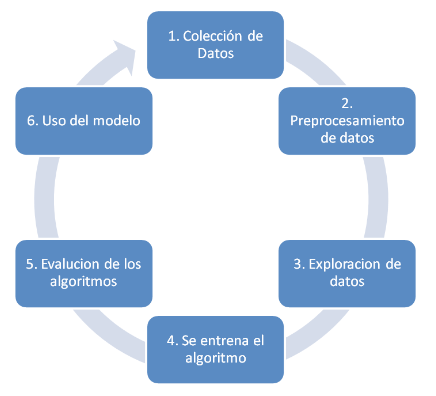
**Fuente: (Saleh, Majzoub, & Saleh, 2025)**

Uno de sus objetivos específicos es que a través del análisis de datos que realiza, mediante técnicas como: la regresión linear, árboles de decisión, redes neuronales o redes bayesianas, entre otras, permitirán el reconocer patrones, obtener conocimiento, información y poder realizar predicciones (Rojas, 2020).

### Etapas de Machine Learning

Existen diferentes etapas importantes dentro del aprendizaje automático que transforman los datos en modelos útiles y funcionales. Cada uno de ellos tiene su propio propósito fundamental para lograr que los modelos obtenidos sean precisos y eficaces (PureStorage, 2024). Estos, como pueden verse en la Figura 2-2, son:

* **Recolección de datos:** Consiste en obtener la información de distintas fuentes, siendo que los datos deben ser de buena calidad, esto es preciso para que el modelo pueda desempeñarse de manera ideal.
* **Preparación de datos:** Antes de ser usados, los datos deben ser limpiados y organizados. Esto significa que se deben corregir errores, eliminar valores vacíos o inconsistencias y la normalización de los datos para poder ser procesados correctamente.
* **Creación de características:** Es el proceso para elegir o crear las variables que tengan mayor relevancia, las que ayudarán al modelo a aprender. Las mejores características permitirán identificar los patrones y mejorar la precisión del modelo.
* **Entrenamiento del modelo:** En esta etapa se escoge el algoritmo a ser usado, y se lo entrena con los datos previamente preparados para que pueda reconocer las relaciones y realice las predicciones.
* **Evaluación del modelo:** Finalmente se mide la funcionalidad del modelo mediante el uso de métricas como la precisión o exactitud, permitiendo detectar errores y realizar ajustes al modelo en caso de que fuese necesario.



**Figura 2‑2: Etapas de Machine Learning**

**Fuente: (Rojas, 2020)**

## Tipos de Aprendizaje de Machine Learning

Existen cuatro enfoques en cuanto a los tipos de aprendizaje que existen dentro del área del machine learning, estos son: Aprendizaje supervisado, no supervisado, semi supervisado, y por refuerzo. Cada uno tiene sus propios usos y ventajas específicos. Conocer cómo funciona cada uno de estos métodos ayuda a entender el grado de autonomía que puede alcanzar un sistema, además de permitir identificar en que áreas llegarían a ser más efectivas (Mining, 2019).

* **Aprendizaje supervisado:** Consiste en enseñar al algoritmo como realizara su tarea, utilizando un conjunto de datos previamente etiquetados bajo una interpretación especifica, con el fin de identificar patrones útiles en el análisis (Rojas, 2020).
* **Aprendizaje no supervisado:** Se diferencia del supervisado debido a que trabaja con datos que no estan clasificados ni etiquetados, siendo su objetivo principal el identificar patrones o similitudes para poder generar estas agrupaciones en base a características comunes (Rojas, 2020).
* **Aprendizaje semi supervisado:** Este método de aprendizaje se sitúa entre el aprendizaje supervisado y no supervisado, ya que los datos con los que trabaja poseen etiquetas limitadas o poco precisas, y en algunos casos ausentes. Su objetivo es lograr modelos más eficientes y complejos sin dejar el aprendizaje completamente sin supervisión (Mining, 2019).
* **Aprendizaje reforzado:** Este no necesita de datos previamente etiquetados, puesto que aprende a tomar decisiones por sí mismo a través de experiencia, mejorando su desempeño a partir de recompensas o resultados obtenidos, mejorando a medida que encuentra la mejor manera de solucionar un problema (Rojas, 2020).

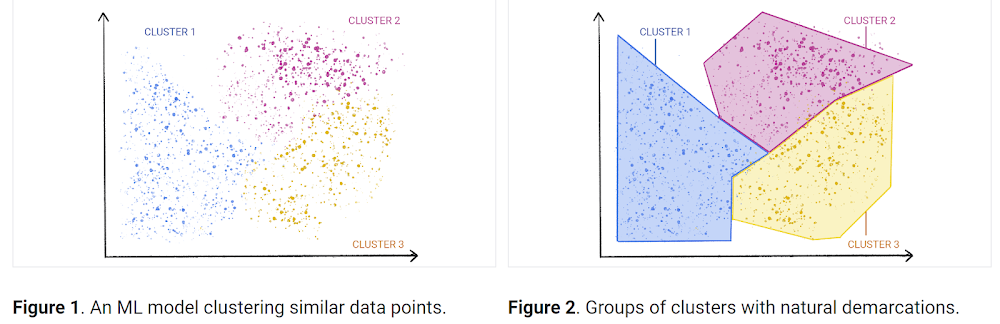
Para este proyecto se hará mayor énfasis en el segundo tipo, el aprendizaje no supervisado, ya que será el utilizado para el entrenamiento de los modelos elegidos a trabajar con los datos obtenidos.

## Aprendizaje No Supervisado

Como habíamos mencionado, el aprendizaje no supervisado trabaja de forma exclusiva con datos de entrada sin etiquetar, permitiéndole al modelo descubrir por cuenta propia los patrones y agrupaciones a partir de la información que este recibe. Estos algoritmos no se hallan dirigidos hacia una forma directa, por el contrario, se diseñan para que puedan identificar similitudes en los datos proporcionados de forma autónoma, sin la intervención de la mano humana para la clasificación. De esta manera, una de sus aplicaciones más relevantes es la estimación de densidad estadística, la cual ayuda a los analistas a comprender de mejor manera los aspectos complejos de grandes volúmenes de datos, como podria ser tendencias económicas o demográficas (Mining, 2019).

Otras aplicaciones para este tipo de aprendizaje se hallan en la utilidad que predispone para tareas como el análisis exploratorio de datos, segmentación de clientes, estrategias de venta cruzada y reconocimiento de imágenes. Siendo dentro de este enfoque que se puede llegar a distinguir tres tareas fundamentales las cuales son: El **agrupamiento**, el cual permite encontrar subconjuntos homogéneos dentro de un conjunto mayor, la **asociación** enfocada en la detección de relaciones significativas entre las variables, y la **reducción** **de dimensionalidad**, que simplifica los conjuntos de datos manteniendo una estructura relevante para un análisis más fácil (Google Cloud, s.f.).

En la Figura 2-3, se puede apreciar una visualización de las segmentaciones de los datos en 3 clústeres diferentes, mediante la división de los puntos, y demarcaciones naturales.



**Figura 2‑3: División de puntos de datos en clústeres similares (izquierda). Grupos de clústeres con demarcaciones naturales (derecha)**

**Fuente: (Google Cloud, s.f.)**

### K-Means

Uno de los modelos que pertenecen al aprendizaje no supervisado es K-Means, el cual es un algoritmo que usa técnicas más sencillas y populares, haciendo de su diseño una herramienta accesible para implementar soluciones de aprendizaje automático (machine learning), sin una gran complejidad. Este método utiliza la agrupación de datos similares, permitiéndole así descubrir patrones ocultos en grandes volúmenes de datos.

Su objetivo principal es el de dividir un conjunto de datos en “k” clústeres o grupos. Para lo cual, el modelo parte de una cantidad determinada de clústeres y va identificando puntos centrales denominados centroides, los cuales representaran el centro de cada grupo (Mining, 2019).

Este proceso da paso a 4 etapas:

* **Inicialización:** Donde se seleccionan al azar k centroides iniciales dentro del conjunto de datos.
* **Asignación:** Donde cada punto de datos es asignado al centroide más cercano, dando forma a los clústeres iniciales.
* **Reubicación:** Se vuelve a realizar el cálculo de los centroides como el promedio de los puntos asignados a cada grupo.
* **Repetición:** Finalmente, el proceso se repite, ajustando los centroides hasta llegar a cumplir con un criterio de parada.

Los criterios de parada pueden manifestarse cuando existe una estabilización de los centroides, siendo que estos dejan de mostrar un cambio significativo entre cada iteración, lo que indicaría que se a alcanzado una solución estable. Otra razón para que el modelo se detenga es alcanzar un límite de iteraciones, previamente definido por el programador, algo bastante útil en condiciones en las que no se puede llegar a determinar un punto óptimo claro. Pese a que una mayor cantidad de iteraciones daría como resultado una mayor precisión, esto también incrementaría el tiempo de procesamiento.

### Clustering Jerárquico

El clustering jerárquico es una técnica de aprendizaje automático no supervisado que es usualmente implementado para organizar diferentes datos en grupo según el parecido que puedan llegar a tener. Este método puede funcionar de 2 maneras:

* **Aglomerativa:** Este tipo de agrupación sigue un enfoque de abajo hacia arriba. Comienza tratando cada uno de los puntos de los datos como un grupo individual, para luego ir uniéndolos con otros puntos más parecidos, hasta llegar a formar un solo grupo grande (IBM, s.f.).

Existen diferentes métodos que se pueden usar para la decisión sobre qué puntos unir:

* + **Método de Ward:** Este método mide cuanto llega a aumentar el error al juntar 2 grupos.
  + **Enlace promedio:** Este usa la distancia promedio entre todos los puntos de 2 grupos distintos.
  + **Enlace completo:** Usa la mayor de las distancias entre 2 puntos, pertenecientes cada uno a un grupo distinto.
  + **Enlace simple:** Este último método, usa la menor de las distancias entre 2 puntos, pertenecientes cada uno a un grupo distinto.

Para realizar la medición de las distancias entre dos puntos, se usa la distancia euclidiana o la distancia de Manhattan, siendo la primera la habitual para el cálculo de la distancia en una línea recta entre dos puntos.

* **Divisiva:** Esta segunda agrupación es lo opuesto de la aglomerativa, ya que su enfoque es de arriba abajo, comenzando por los datos juntos en un solo grupo. A partir de aquí, los va dividiendo en grupos más pequeños según las diferencias que tengan. Pese a ser el método usado con menos frecuencia, también forma parte del enfoque jerárquico (IBM, s.f.).

### K-Medoids

El algoritmo K.Means presenta como una de sus principales desventajas, la sensibilidad a valores atípicos dentro de un dataset, pudiendo llegara distorsionar los resultados de agrupamiento. Es en estos casos que el algoritmo K-Medoids resulta como una mejor opción, ya que emplea objetos reales, en reemplazo de promedios, para obtener puntos de referencia, volviéndolo más robusto cuando los datos presentan ruido o valores extremo (Kaur, Kaur, & Singh, 2014).

Una de las diferencias fundamentales entre K-Means y K-Medoids, radica en que el segundo, reduce casi al minimo el criterio de error absoluto, mientras que el primero trata de minimizar la suma de los errores cuadrados (SSE). También cabe mencionar, que K.-Medoids es iterativo, ya que en cada ciclo busca que cada punto representativo de un clúster (denominado como medoid) sea el más adecuado posible para el grupo en cuestión. Es durante estas iteraciones que se considera el reemplazo de un medoid por otro arbitrario dentro del conjunto, y se evalúa como afectaría a las asignaciones de los puntos de clúster original. De esta manera se calcula un costo de intercambio, el cual pertenece al criterio del error absoluto, y se acumula dentro del costo global del modelo, eligiendo al final el que menor costo tenga (Aggarwal & Reddy, 2014).

A pesar de esto, entre las desventajas de este algoritmo, podemos mencionar que los resultados pueden variar en cada ejecución, debido a su aleatoriedad inicial, así también como su costo computacional.

## Análisis RFM

El análisis RFM, es la abreviación de tres indicadores esenciales con los que trabaja, estos son: Recencia, Frecuencia y Valor Monetario. Este análisis es una estrategia moderna usada en el mercadeo para poder identificar a los clientes más valiosos de una empresa, tomando en cuenta su historial de compras, puesto que al realizar una evaluación con la frecuencia que compran, cuanto gastan y cuan reciente fue su última compra, se puede llegar a estimar la probabilidad de respuesta a futuras acciones comerciales por parte de los clientes.

Esta técnica se basa en el principio de que los consumidores que realizan compras recientes y frecuentes, o que hacen una mayor inversión de montos, suelen mostrar una tendencia a ser mucho más receptivos a mensajes y promociones que la empresa pueda ofrecer. También se puede observar que los clientes más recientes tienen una tendencia a reaccionar de forma más positiva a campañas de marketing que aquellos clientes que llevan bastante tiempo sin realizar una compra en el negocio (Morelo Tapias, 2014).

El enfoque de RFM toma como base la Ley de Pareto, también conocida como la regla del 80/20, que plantea lo siguiente: Una pequeña proporción de los clientes, usualmente alrededor del 20%, es responsable de producir la mayor parte de los ingresos, alrededor del 80% de ellos. A pesar de que puede sonar a una generalización, esta distribución ha demostrado tener validez en diversos sectores de índole económica e incluso en actividades deportivas.

Debido a la evolución tecnológica y el apogeo de los sistemas CRM (Customer Relationship Management, o Gestión de Relación con los Clientes en español), el análisis RFM ha sido integrado considerablemente en plataformas digitales de gestión de clientes, convirtiéndose de esta manera en una herramienta fundamental para orientar la toma de decisiones comerciales.

Ya en la práctica, cada cliente es asignado con una puntuación de entre 1 a 5 para cada una de las tres dimensiones del modelo RFM. La combinación de estas puntuaciones es representada como una “celda” RFM. Los clientes que obtengan calificaciones de 5 en las tres dimensiones son considerados como los más valiosos, ya que esto se interpretaría como los que realizan compras con más frecuencia, hacen compras de grandes sumas de dinero, y sus compras son las más recientes. Por tanto, las empresas que aplican esta metodología suelen ofrecer una atención prioritaria y mejor enfocada a este tipo de clientes, reconociéndolos como una porción importante dentro de sus beneficios.

A pesar de esto, se debe tomar en cuenta la aplicación prudente del análisis RFM, ya que el presionar enormemente a los clientes más rentables puede dar resultados contraproducentes, dejando de lado a aquellos clientes con puntuaciones bajas, pudiendo llegar a perder oportunidades de mejoría. En su lugar, es recomendable trabajar enérgicamente en estos últimos para incentivar en su evolución a consumidores leales, promoviendo una mayor frecuencia y volumen en sus compras.

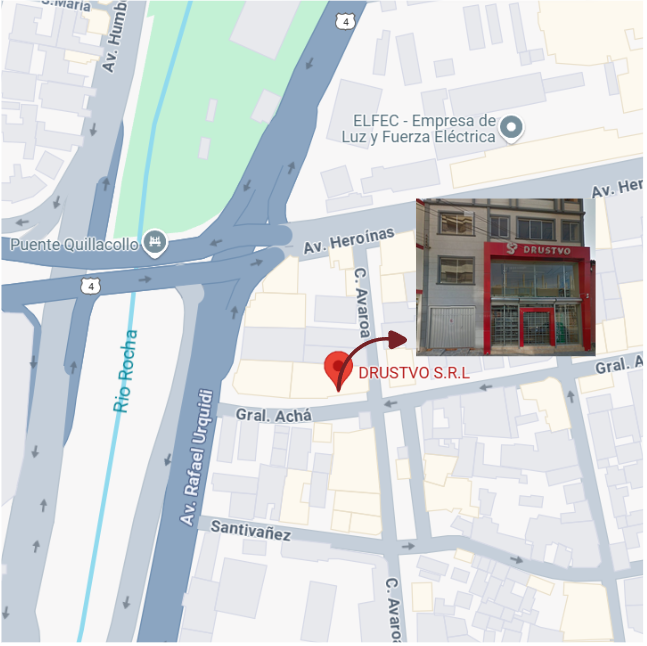
Cabe mencionar que esta técnica es bastante usada para la segmentación de clientes, basados en su comportamiento de compra como ya se había mencionado, siendo esta la razon de su uso en este proyecto.

# Marco metodológico

## Área de estudio

El área de estudio se centra en el municipio de Cercado, en la ciudad de Cochabamba.

La empresa comercializadora de productos eléctricos Drustvo S.R.L. se encuentra en el departamento de Cochabamba, ubicada en la zona central del municipio de Cercado, entre las calles General Achá y Avaroa como puede verse, junto con su fachada, en la captura de Google Maps en la Figura 3-1, y está operativa desde el año 2014.



**Figura 3‑1: Imagen de la ubicación y fachada de la empresa Drustvo S.R.L.**

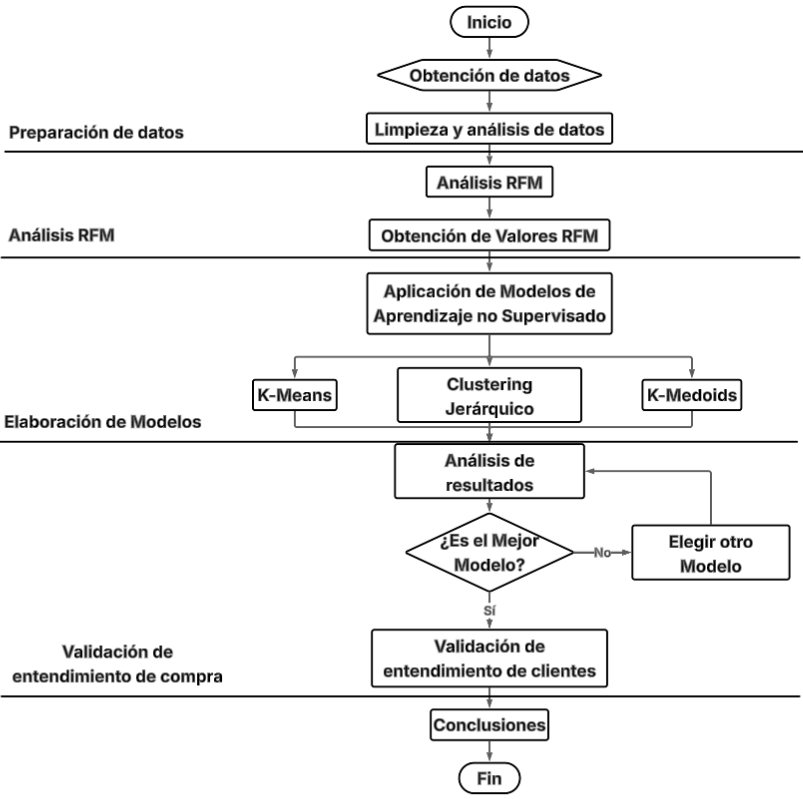
**Fuente: (Google Maps, 2025)**

## Flujograma metodológico

Inicialmente se realizó la recolección de los datos, logrando obtener estos gentilmente, por la empresa comercializadora Drustvo S.R.L., la cual pudo ofrecer un registro de ventas realizados desde el año 2021 al 2023. Posterior a esto, se tiene pensado realizar un análisis de los datos obtenidos, así como su limpieza y preparación para su uso en los modelos de aprendizaje no supervisado, mediante la herramienta Power BI.

Luego de obtener los resultados de cada modelo (K-Means, K-Medoids, Clúster Jerárquico), se procederá a cargar los datos preparados con la ayuda de un Notebook en Colab de Google, donde mediante código se obtendrá los clústeres apropiados para cada modelo y su posterior implementación, obteniendo visuales que puedan ayudar con la evaluación de los resultados obtenidos. Finalmente se realizará una comparativa con los resultados, para definir cuál es el modelo con mejor segmentación, para posteriormente analizar y validar que estos ayudarán con el entendimiento de compra de los clientes.

Todo este proceso se encuentra resumido en el flujograma de la Figura 3-2.



**Figura 3‑2: Flujograma**

**Fuente: (Elaboración propia, 2025)**

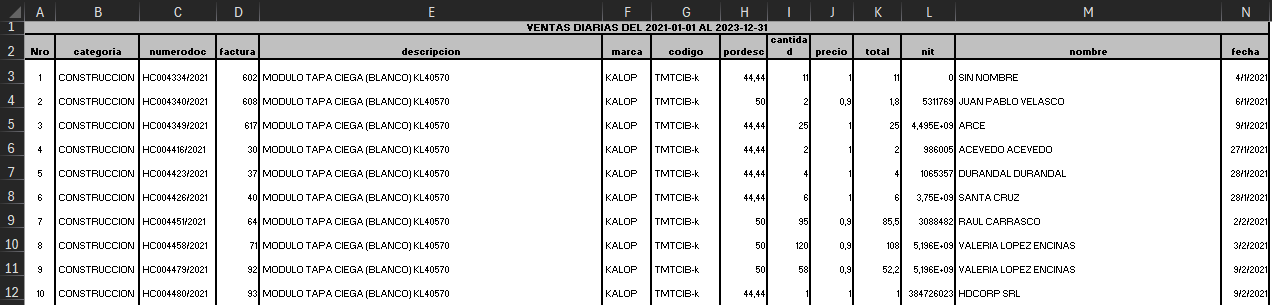
## Fuentes de información

Los datos obtenidos, como ya se mencionó en puntos previos, provienen de la empresa comercializadora de productos eléctricos Drustvo S.R.L. Estos consisten en una porción de los registros de ventas perteneciente a la categoría de “Construcción”, entre los años 2021 al 2023, los cuales llegan a un total de 2530, siendo adquiridos en un documento de Excel como se ve en la Figura 3-3, cuyas variables estan descritas en la Tabla 3-1.

|  |  |
| --- | --- |
| Variables | Descripción |
| Nro | Es un índex que lleva cuenta de todos los registros listados en la tabla |
| categoria | Se refiere a la categoría dentro de la que esta asignado el producto |
| numerodoc | Es un código asignado a cada factura de ventas realizadas |
| factura | Es el número de factura emitida, reiniciándose cada inicio de año |
| descripcion | Nombre con breve descripción del producto |
| marca | Marca del producto |
| codigo | Código único del producto |
| pordesc | Porcentaje de descuento aplicado al precio |
| cantidad | Cantidad vendida del producto |
| precio | Precio individual de cada producto |
| total | Monto total del costo de la venta (Cantidad x Precio) |
| nit | NIT del cliente |
| nombre | Nombre del cliente |
| fecha | Fecha en la que la venta fue realizada |

**Tabla 3‑1: Descripción de las variables halladas en el dataset**

**Fuente: (Elaboración propia, 2025)**



**Figura 3‑3: Vista de tabla de registros del archivo Excel**

**Fuente: (Anexo 1, 2025)**

## Herramientas usadas

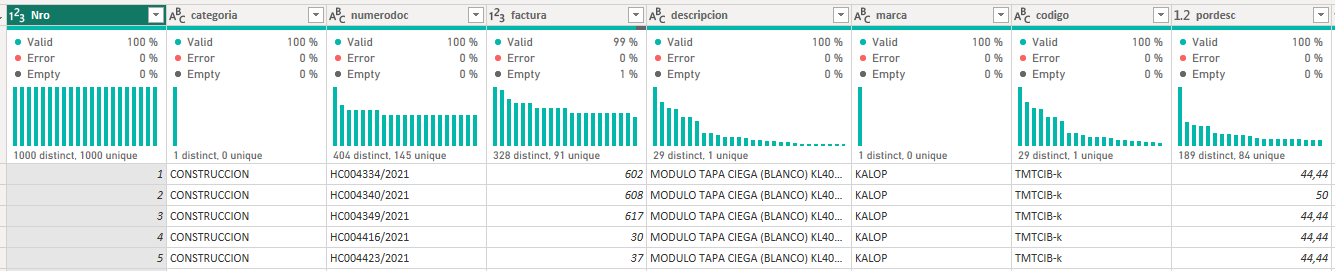
Entre las herramientas seleccionadas para el desarrollo del proyecto tenemos:

* **Power BI:** Herramienta para el análisis y visualización de datos, usada en este caso para la limpieza de datos, y para la creación y modelado de tablas de relación basado en los registros obtenidos. Así como para la visualización de gráficos para el entendimiento de los datos obtenidos.
* **DAX Studio:** Una herramienta complementaria de Power BI para consultas, en este caso se hizo uso de la misma para poder rescatar la tabla minable a ser usada para la segmentación de clientes con los algoritmos de aprendizaje no supervisado.
* **Notebook de Colab:** Un entorno de desarrollo, en el cual se escribió y preparo el código para poder ejecutar los 3 algoritmos, así como la visualización de las segmentaciones generadas por los mismo.

Dentro de los subsiguientes puntos se ira detallando el uso de estas herramientas en cada sección del desarrollo del proyecto.

## Preparación de los datos

Para empezar, se realizó la importación del archivo Excel dentro de una nueva instancia en blanco en Power BI. Para esto selecciono el archivo con los datos provistos por la empresa, y se escogió la opcion de tablas sugeridas ya que presentaba una mejor visualización de la tabla de registros. Luego se eligió la opcion “Transform Data” para finalizar con el cargado de los datos, obteniendo una visualización previa como se observa en la Figura 3-4.



**Figura 3‑4: Vista de la tabla en Power BI**

**Fuente: (Elaboración propia, 2025)**

### Información de las variables

Una vez obtenidos y cargados los datos en una tabla de Power BI, se hizo una revisión de cada variable, ya que esta herramienta permite ver datos puntuales de las columnas seleccionadas, obteniendo así métricas importantes de las variables cuantitativas en la Tabla 3-2, y las variables cualitativas en la Tabla 3-3:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables Cuantitativas | Conteo | Error | Vacío | Min | Max | Promedio |
| NRO | 2530 | 0 | 0 | 1 | 2530 | 1265,5 |
| FACTURA | 2530 | 0 | 27 | 1 | 1211 | 373,35 |
| PORDESC | 2530 | 0 | 0 | -17,52 | 60,67 | 30,56 |
| CANTIDAD | 2530 | 0 | 0 | 1 | 572 | 13,2 |
| PRECIO | 2530 | 0 | 0 | 0,8 | 300 | 9,68 |
| TOTAL | 2530 | 0 | 0 | 0,9 | 3068 | 71,21 |
| NIT | 2530 | 0 | 0 | 0 | 9239669011 | - |
| FECHA | 2530 | 0 | 0 | 4/1/2021 | 30/12/2023 | 13/6/2022 |

**Tabla 3‑2: Tabla de Variables Cuantitativas**

**Fuente: (Elaboración propia, 2025)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables Cualitativas | Conteo | Error | Vacío | Distinto |
| CATEGORIA | 2530 | 0 | 0 | 1 |
| NUMERODOC | 2530 | 0 | 0 | 680 |
| DESCRIPCION | 2530 | 0 | 0 | 123 |
| MARCA | 2530 | 0 | 0 | 1 |
| CODIGO | 2530 | 0 | 0 | 123 |
| NOMBRE | 2530 | 0 | 0 | 232 |

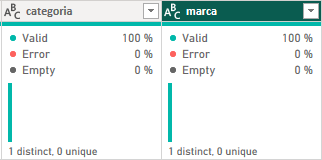
**Tabla 3‑3: Tabla de Variables Cualitativas**

**Fuente: (Elaboración propia, 2025)**

Con esta vista previa de las variables, ya podemos notar que existen datos nulos en la columna de “factura”, así como un posible error de registro dentro los datos de “porcdesc” los cuales analizaremos a más profundidad durante la limpieza de los datos.

### Limpieza de datos

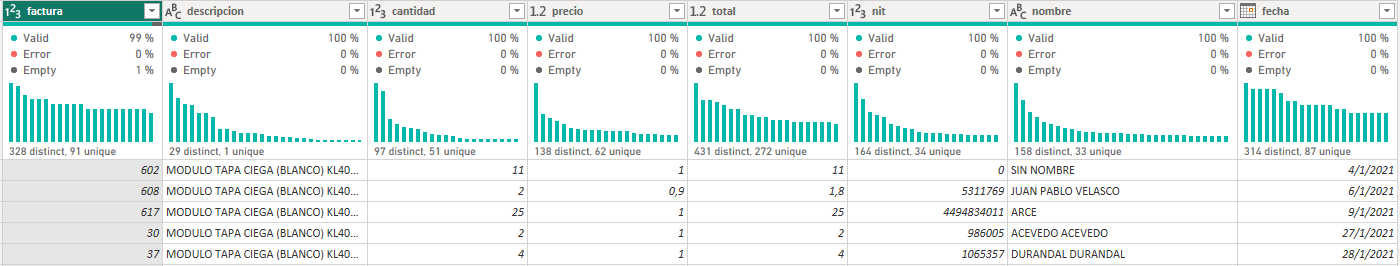
Iniciando con la limpieza de los datos, se procedió con la eliminación de las columnas: nro, categoría, factura, y marca. Se tomo esta decisión debido a que la columna “nro” solo funciona como un índex de los registros, por tanto, no es útil en el análisis posterior que realizaremos. Las columnas “categoría” y “marca” también serán eliminadas, ya que solo constan de un valor que se repite en todos los registros como se ve en la Figura 3-5.



**Figura 3‑5: Columnas "categoría" y "marca", ambas con un solo valor**

**Fuente: (Elaboración propia, 2025)**

Luego nos encontramos con la columna de “factura”, de la cual excluiremos aquellos registros que tengan valores nulos, ya que estos representarían facturas de ventas anuladas. Se observo también que la columna “numerodoc” cumplía la misma función que “factura” en cuanto al registro del número de venta, y a diferencia de la columna “factura”, esta no repetía valores anualmente, ya que es un código que se genera por el sistema para cada venta. Por este motivo se decidió finalmente la eliminación de “factura”, obteniendo la tabla visible en la Figura 3-6, una vez concluida la limpieza de los datos.



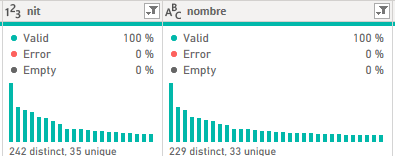
**Figura 3‑6: Vista de tabla una vez terminada la limpieza de datos**

**Fuente: (Elaboración propia, 2025)**

Ya con estas columnas procederemos a realizar la creación de tablas individuales para un modelo de relaciones con los valores de: Ventas, Cliente y Productos.

### Tabla de Clientes

Para la tabla con informacion de los clientes, se filtró solo las columnas de “nit” y “nombre”, ya que eran las únicas con datos relevantes relacionados con los clientes, pudiendo observar que todos los datos son válidos en ambas columnas como se ve en la Figura 3-7.

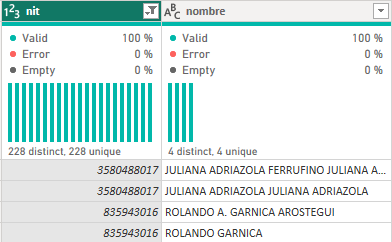


**Figura 3‑7: Columnas filtradas para la tabla Clientes**

**Fuente: (Elaboración propia, 2025)**

Luego se procedió a realizar una revisión entre los valores de la columna “nombre”, dentro la cual se pudo notar que existían registros de “SIN NOMBRE” entre los nombres de los clientes, los cuales también carecían de un NIT, teniendo valor de 0 dentro su columna. Al ser datos que no aportaran en la segmentación, fueron eliminados.

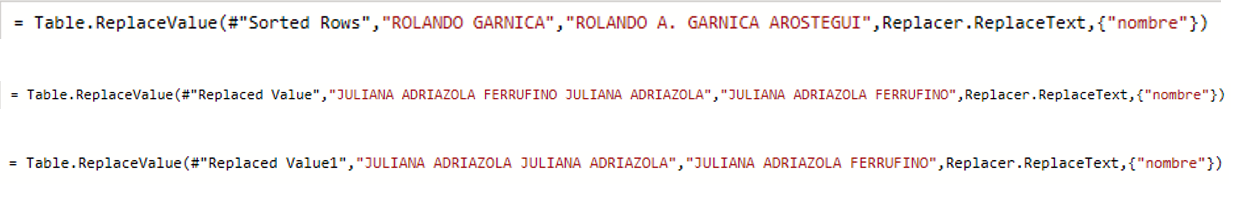
Continuando con la limpieza, se decidió eliminar los duplicados desde la columna “nit”, puesto que podían existir registros de ventas que tuviesen el mismo nombre, pero con diferente NIT. Durante este proceso se pudo notar que existían dos casos especiales en los que un mismo NIT pertenecía a dos nombres diferentes. En el primer caso el nombre y el apellido se repetían, mientras que, en el segundo caso uno de los registros tenía el nombre completo del cliente, mientras que en otro registro el nombre estaba abreviado, como puede verse en la Figura 3-8.



**Figura 3‑8: Casos especiales con mismo "nit", pero nombres diferentes**

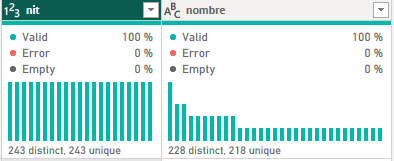
**Fuente: (Elaboración propia, 2025)**

Este error fue subsanado eliminando la duplicación del nombre en el primer caso, y modificando estos dos casos, para que ambos mantengan los valores en los que los nombres estaban más completos, como se puede ver en los comandos de la Figura 3-9. Luego se procedió con la eliminación de duplicados desde los valores de “nit”, logrando comprobar que todos los valores de esta columna eran únicos como se ve en la Figura 3-10.



**Figura 3‑9: Comandos usados para la corrección de los registros**

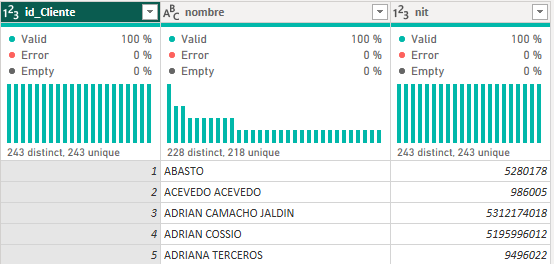
**Fuente: (Elaboración propia, 2025)**



**Figura 3‑10: Métricas de las columnas posterior a la corrección**

**Fuente: (Elaboración propia, 2025)**

Ya concluida la limpieza de los datos respecto a los clientes, se reordeno la columna de nombres de los clientes de forma alfabética, y se añadió un Índex para darle el nombre de “id\_Cliente”, dando como vista final de la tabla Clientes la Figura 3-11.



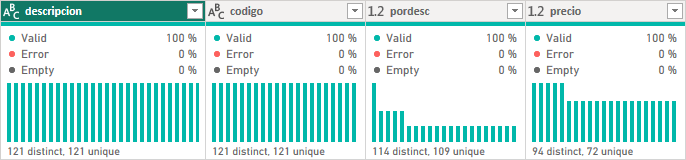
**Figura 3‑11: Vista final de la tabla Clientes**

**Fuente: (Elaboración propia, 2025)**

### Tabla de Productos

Continuando con la tabla de productos, se determinó hacer uso de las columnas “descripcion”, “codigo”, “pordesc” y “precio”. Habiendo mantenido una aclaración con el encargado de poder facilitar el registro de las ventas de la empresa, se determinó que los valores de la columna “pordesc” no eran correctos, ya que estos iban variando en base a un valor que se asignaba al precio de los productos, dependiendo de factores externos, por tanto, para poder lograr una mejor estimación de los precios, se decidió filtrar los valores de “pordesc” de forma ascendente.

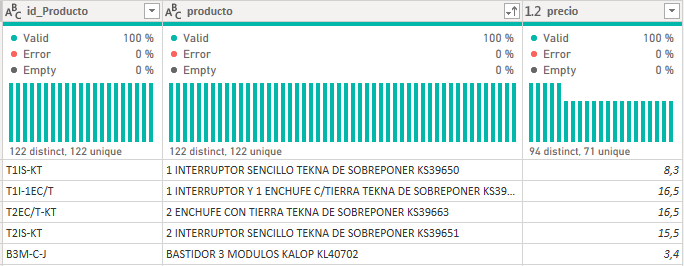
De esta manera al eliminar los duplicados, se mantendría el precio más elevado registrado para cada producto, hallando un total de 121 productos vendidos con sus respectivos códigos a lo largo de 3 años, como se ve en la Figura 3-12.



**Figura 3‑12: Métricas de la tabla Productos luego de la eliminación de duplicados**

**Fuente: (Elaboración propia, 2025)**

Luego se cambió el nombre de la columna “codigo” por “id\_Producto” y se reordeno las columnas para una vista más prolija dando como resultado la vista de la Figura 3-13.

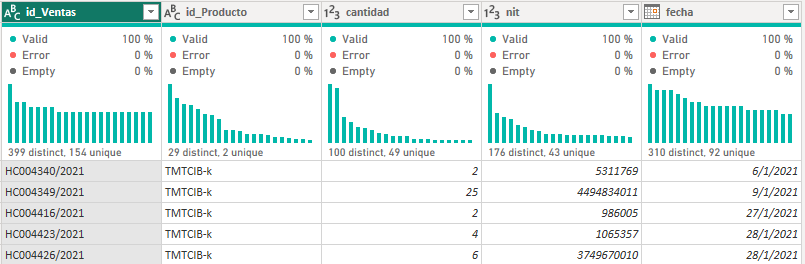


**Figura 3‑13: Vista final de la tabla Productos**

**Fuente: (Elaboración propia, 2025)**

### Tabla de Ventas

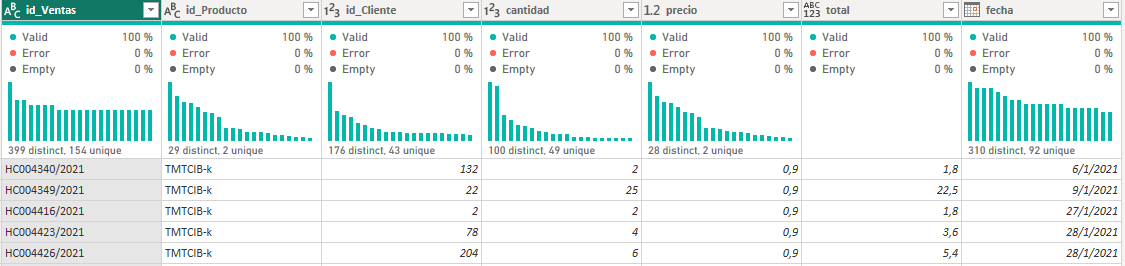
Una vez obtenidos las tablas de Clientes y Productos, solo restaba elaborar la de ventas. Primero se eliminó la mayoría de las columnas, dejando solo “numerodoc”, “codigo”, “cantidad”, “nit” y “fecha”, cambiando los nombres a los 2 primeros por “id\_Venta” y “id\_Producto”, tal como se ve en la Figura 3-14.



**Figura 3‑14: Vista inicial de tabla Ventas**

**Fuente: (Elaboración propia, 2025)**

A continuación, gracias al modelo relacional (explicado en el siguiente punto), se pudo obtener la columna “id\_Cliente” que reemplazaría a la columna “nit”. Además, se obtuvo los precios ya filtrados de la tabla de Productos, de esta manera se hizo el cálculo del total de cada venta (cantidad \* precio) dando como resultado la Figura 3-15.



**Figura 3‑15: Tabla Ventas con las columnas “id\_Cliente”, "precio" y "total" añadidas**

**Fuente: (Elaboración propia, 2025)**

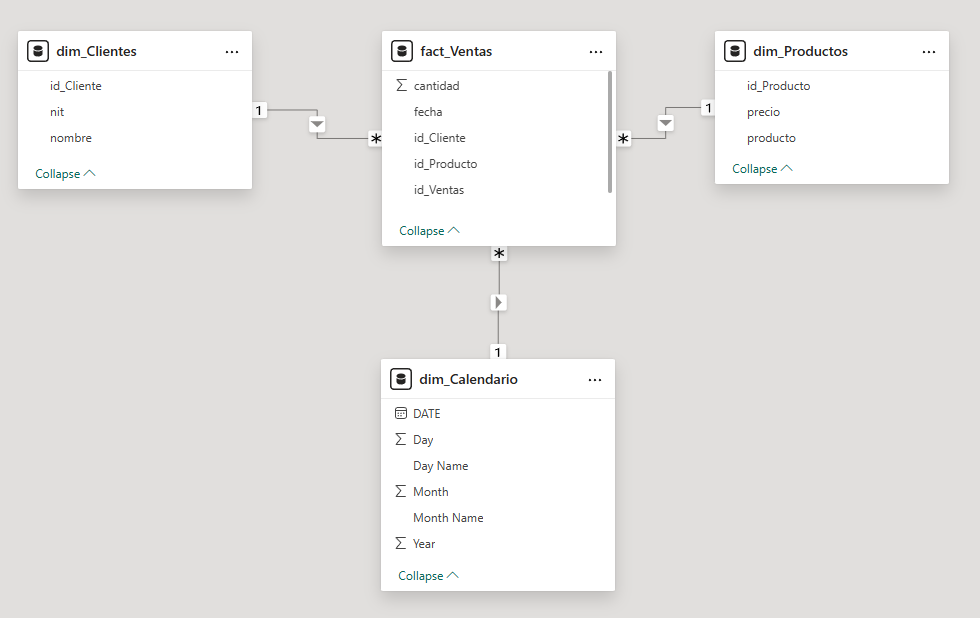
### Modelo Relacional de Tablas

Con la elaboración de las tablas de Clientes, Productos y Ventas, solo basto realizar la asignación de relaciones entre ellas. Las relaciones se basaron en los IDs de las tablas Cliente y Productos, siendo una relacion de uno a muchos, de estas dos a la tabla de ventas como se aprecia en la Figura 3-16. Adicionalmente, se creó una tabla de Calendario, la cual tendría todas las fechas existentes entre los años 2021 al 2023. Una vez realizadas las relaciones, se paso al apartado de “Model View” donde se podía ver el resultado final de las tablas con sus relaciones como en la Figura 3-17.



**Figura 3‑16: Relaciones entre las tablas Cliente, Productos, Ventas y Calendario**

**Fuente: (Elaboración propia, 2025)**



**Figura 3‑17: Vista del modelo relacional**

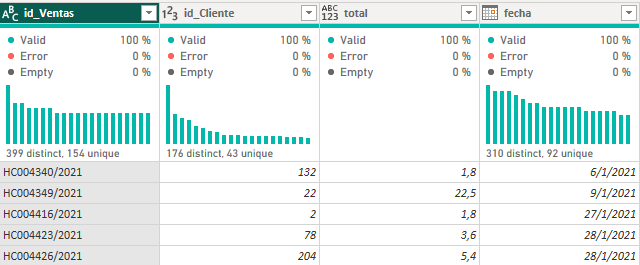
**Fuente: (Elaboración propia, 2025)**

Con este último paso se habría concluido con la preparación de las tablas y la limpieza de los datos, dando paso a la preparación del análisis RFM.

## Análisis RFM

### Tabla de valores RFM

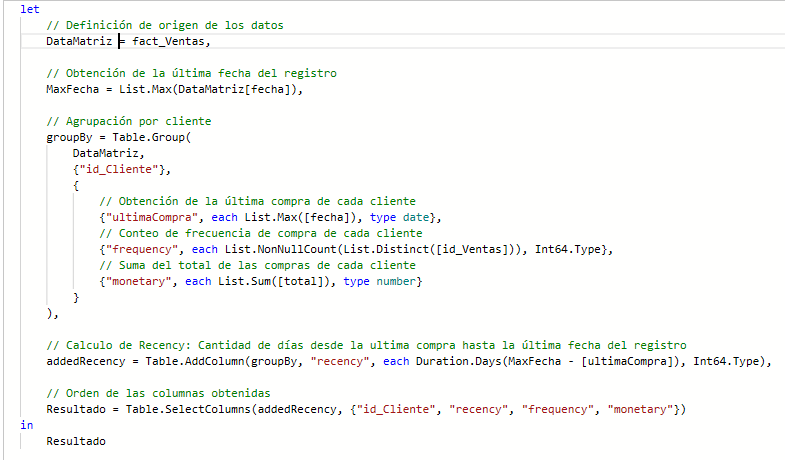
Una vez obtenido las tablas limpias separadas de los registros de ventas, se seleccionó las columnas “id\_Ventas”, “id\_Cliente”, “total” y “fecha”, ya que cuentan con los datos necesarios para el análisis RFM. Este análisis fue seleccionado porque puede ofrecernos nuevas variables respecto a cada cliente, basándonos en su frecuencia de compra, monto total gastado y la cantidad de días transcurridos desde su última compra. De esta manera se obtuvo el código de venta y el del cliente, el total de la compra, y la fecha de esta como se ve en la Figura 3-18.



**Figura 3‑18: Columnas con los datos necesarios para la elaboración de la tabla RFM**

**Fuente: (Elaboración propia, 2025)**

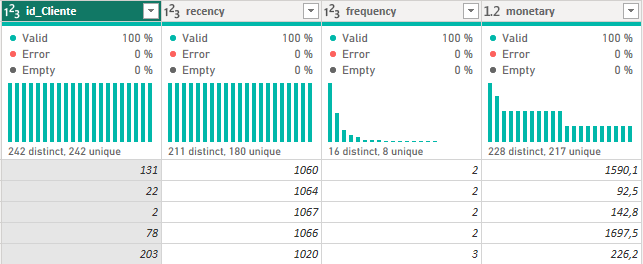
Con estos valores en mente, se creó una nueva tabla mediante el editor avanzado disponible, y mediante el código visible en la Figura 3-19, se calculó, la fecha de la última compra del cliente, un conteo de cada compra realizada por cliente y finalmente una suma de todas las compras realizadas por cliente.



**Figura 3‑19: Creación de tabla RFM**

**Fuente: (Elaboración propia, 2025)**

Una vez ingresado este código, se obtuvo como resultado la tabla “dim\_RFM”, con los valores del código de cliente, su recencia, frecuencia de compra y el valor total de compras realizadas, que puede verse en la Figura 3-20.



**Figura 3‑20: Vista final de tabla RFM**

**Fuente: (Elaboración propia, 2025)**

### Cálculo de valores RFM

Una vez generada la tabla a la cual denominada como “dim\_RFM”, se puede obtener las métricas necesarias para generar la matriz de valores de RFM que será útil para la puntuación de los mismos.

Iniciando con la obtención de los valores máximo y minimo de las columnas “recency”, “frequency” y “monetary”, se calcula el rango en base a la resta del valor máximo menos el valor mínimo, el intervalo será cinco, ya que es el definido por defecto para este análisis. Por último, se calcula la amplitud de cada intervalo, dividiendo el rango entre la cantidad de intervalos establecidos. Dando de esta manera los resultados visibles en la Tabla 3-4.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Recency | Frequency | Monetary |
| Max | 1090 | 57 | 36660,1 |
| Min | 0 | 1 | 3,4 |
| Rango | 1090 | 56 | 36656,7 |
| Intervalos | 5 | | |
| Amplitud | 218 | 11,2 | 7331,34 |

**Tabla 3‑4: Cálculo de rango, intervalo y amplitud para RFM**

**Fuente: (Elaboración propia, 2025)**

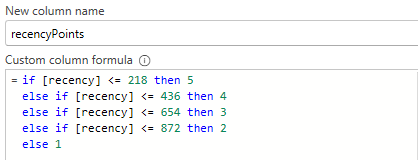
Con estos datos se pasó a realizar la asignación de puntaje a los diferentes intervalos obtenidos gracias a los valores de amplitud conseguidos para cada una de las tres variables. Se asignan los puntajes más altos a los valores más altos de las variables Frequency y Monetary, y en el caso de Recency, la asignación es al revés, ya que se otorga un mayor puntaje cuando la cantidad de días desde la última compra es la menor posible. Teniendo como resultado la Tabla 3-5, que muestra el puntaje asignado a cada intervalo obtenido.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Recency | | Frequency | | Monetary | |
| Intervalos | **Puntuación** | **Intervalos** | **Puntuación** | **Intervalos** | **Puntuación** |
| 0 - 218 | 5 | 1 - 12,2 | 1 | 3,4 - 7334,74 | 1 |
| 219 - 436 | 4 | 12,3 - 22,4 | 2 | 7335,75 - 14666,08 | 2 |
| 437 -654 | 3 | 22,5 - 33,6 | 3 | 14667,09 - 21997,42 | 3 |
| 655 - 872 | 2 | 33,7 - 44,8 | 4 | 27997,43 - 29328,76 | 4 |
| 873 - 1090 | 1 | 44,9 - 56 | 5 | 29328,77 - 36660,1 | 5 |

**Tabla 3‑5: Tabla de puntajes para las variables RFM**

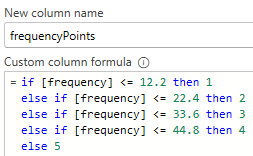
**Fuente: (Elaboración propia, 2025)**

Una vez definido los puntajes, estos serán agregados a la tabla “dim\_RFM” mediante los comandos que asignan el puntaje dentro de cada intervalo, para Recency en la Figura 3-21, para Frequency en la Figura 3-22 y para Monetary los de la Figura 3-23.



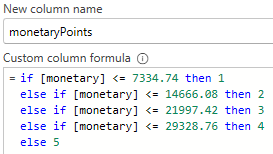
**Figura 3‑21: Creación de columna "recencyPoints"**

**Fuente: (Elaboración propia, 2025)**



**Figura 3‑22: Creación de columna "frecuencyPoints"**

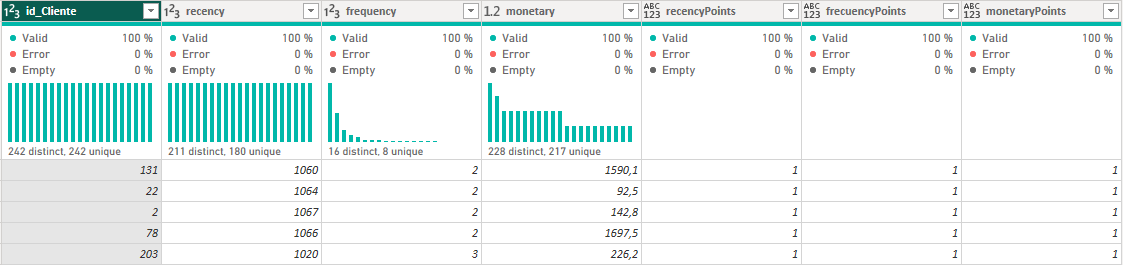
**Fuente: (Elaboración propia, 2025)**



**Figura 3‑23: Creación de columna "monetaryPoints"**

**Fuente: (Elaboración propia, 2025)**

Dando como resultado la vista previa de esta tabla la Figura 3-24, en la que se hallan tanto los valores de RFM, así como los puntajes de cada uno, con el código del cliente al que pertenecen.



**Figura 3‑24: Vista final de la tabla “dim\_RFM”**

**Fuente: (Elaboración propia, 2025)**

Posteriormente, se creó una nueva columna que será el puntaje combinado de las tres variables, para que se pueda aplicar un conteo a cada combinación registrada y para que sea más facil poder obtener una interpretación de la segmentación de clientes obtenidas por el análisis RFM. Ya con estos valores, se asignó una etiqueta de tipo de cliente, con su respectiva descripción, junto con un conteo de cuantos pertenecen a este grupo como se puede apreciar en la Tabla 3-6.

|  |  |  |  |
| --- | --- | --- | --- |
| Puntaje RFM | Cantidad | Tipo de Cliente | Descripción |
| 555 | 1 | Ideal | Clientes Recientes, frecuentes y de alto gasto |
| 531 | 1 | Alto Valor | Compras recientes y con frecuencia, no mucho gasto |
| 522 | 1 | Leales | Frecuentes y reciente con gasto medio |
| 521 | 3 | Leales | Frecuentes y recientes gasto bajo |
| 512 | 1 | Nuevos Prometedores | Reciente, poca frecuencia, gasto medio. Hay potencial |
| 511 | 56 | Nuevos | Recientes, baja frecuencia y gasto |
| 421 | 1 | Riesgo | No tan recientes, baja frecuencia y gasto |
| 412 | 1 | Riesgo | Similar al de arriba |
| 411 | 52 | Riesgo | Similar al de arriba |
| 311 | 35 | Dormidos | Tiempo que no compran, baja frecuencia y gasto |
| 212 | 1 | Casi Perdidos | Bajos valores, se están perdiendo |
| 211 | 27 | Casi Perdidos | Similar al de arriba |
| 111 | 62 | Perdidos | Compras lejanas, baja frecuencia y montos |

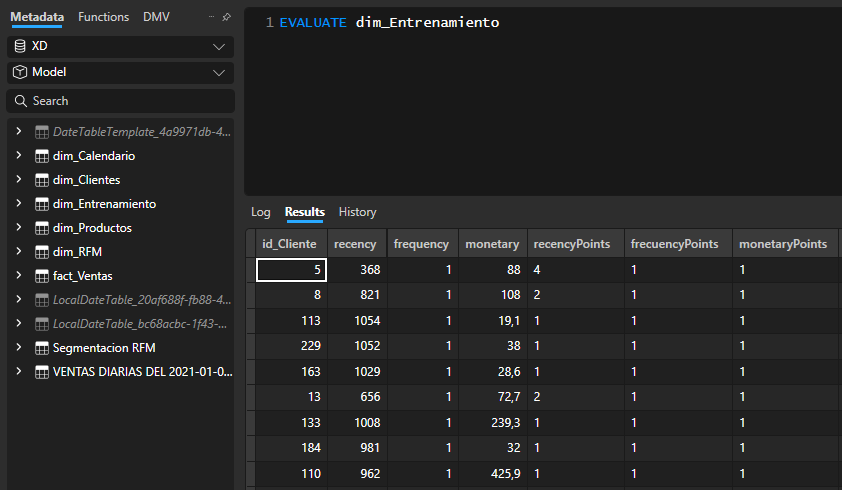
**Tabla 3‑6: Combinaciones y descripción de posibles tipos de clientes según el análisis RFM**

**Fuente: (Elaboración propia, 2025)**

Como se puede observar, existen 13 posibles combinaciones, entre las cuales se puede determinar 8 tipos distintos de grupos de clientes como ser: “Cliente Ideal”, de “Alto Valor”, “Leales”, “Nuevos”, “Riesgo”, “Dormidos”, “Casi Perdidos” y “Perdidos”, teniendo la mayor cantidad de clientes los grupos “Nuevos”, “Riesgo” y “Perdidos”, denotando la necesidad de la empresa por mantener la atención de sus clientes mediante posibles campañas de marketing.

## Elaboración de los modelos de aprendizaje no supervisado

Ya con los valores de “Recency”, “Frequency” y “Monetary” obtenidos, se creó una copia de la tabla “dim\_RFM” y se le asignó el nombre de “dim\_Entrenamiento”, dejando solo los valores del código de cliente, valores RFM y sus puntajes. Con la ayuda del complemento DAX Studio, que permite extraer una tabla de un proyecto de Power BI, mediante consultas, se logró obtener la tabla que será usada en los algoritmos de aprendizaje no supervisado. Como podemos ver en la Figura 3-25, basta con poner el comando EVALUATE, seguido por el nombre de la tabla deseada, eligiendo la opcion de “Static” en Output, para la obtención de un archivo con formato Excel.

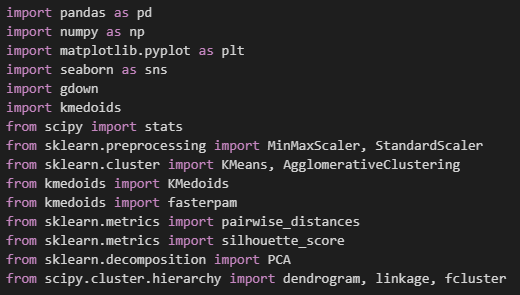


**Figura 3‑25: Vista de la tabla Entrenamiento desde DAX Studio**

**Fuente: (Elaboración propia, 2025)**

A continuación, se creó un Notebook de Colab nuevo, donde en primer lugar, se instaló las librerías de gdown, y kmedoids, las cuales permitirán descargar el archivo desde Google Drive, y la implementación del algoritmo de K-Medoids respectivamente.

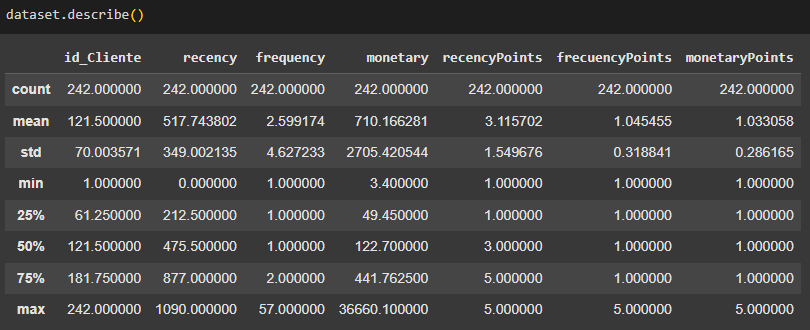
Luego se importó las librerías necesarias como pandas, numpy, matplot, seaborn, gdown, kmedoids, kmeans, agglomerative clustering, entre otros para la implementación de los algoritmos a usar, creación de gráficos, normalización y escalado de datos y técnicas para hallar la cantidad de clústeres óptimos para cada algoritmo, como se ve en la Figura 3-26.



**Figura 3‑26: Vista de librería importadas para el modelado de los algoritmos**

**Fuente: (Elaboración propia, 2025)**

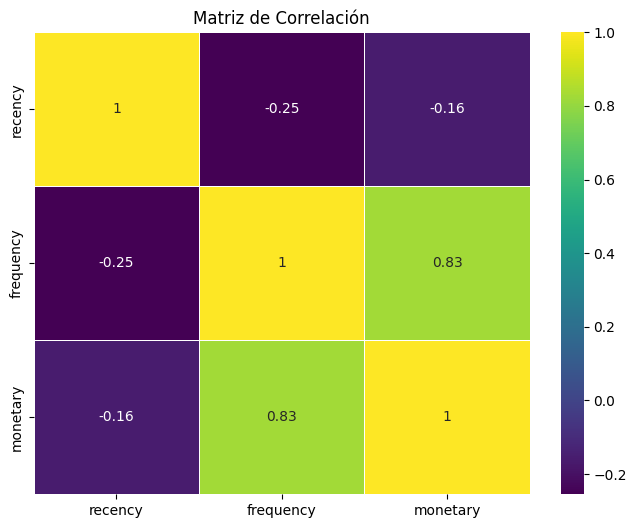
Luego se procedió a descargar el archivo de Excel que contiene la tabla de Entrenamiento mencionada en el punto anterior de este documento, el cual está ubicado en una carpeta de Google Drive. Y comprobamos que no tenga valores nulos y que es posible ver los datos contenidos, luego en la Figura 3-27, se pude observar la descripción de las métricas generales de la tabla.



**Figura 3‑27: Métricas de la tabla de entrenamiento**

**Fuente: (Elaboración propia, 2025)**

Se decidió usar solo las variables de “recency”, “frequency” y “monetary” para la aplicación en los algoritmos, por tanto, en la Figura 3-28 se obtuvo una matriz de correlación entre estos tres valores.

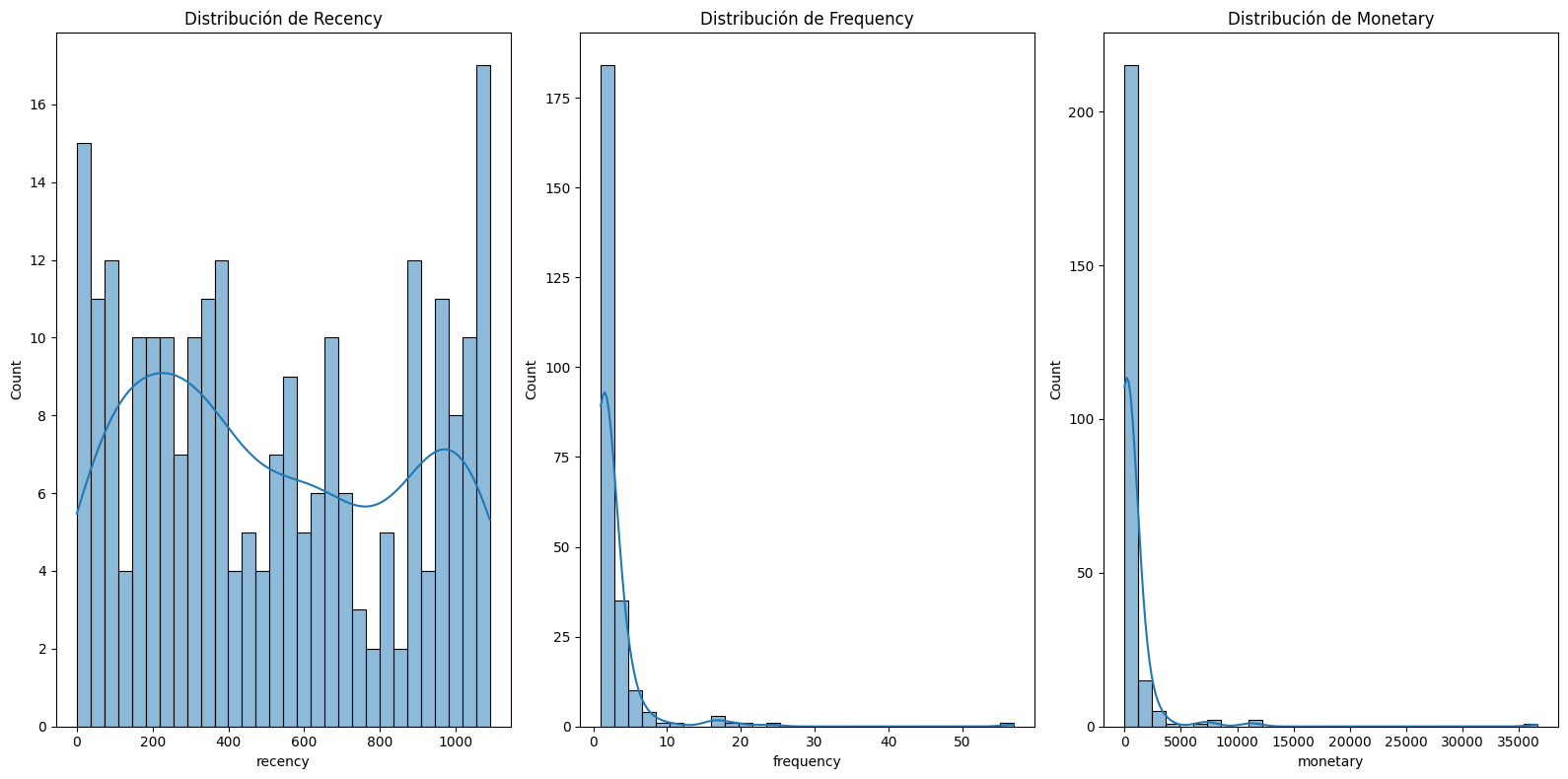


**Figura 3‑28: Matriz de correlación de las variables "recency", "frequency" y "monetary"**

**Fuente: (Elaboración propia, 2025)**

Aquí podemos notar que existe una buena correlación entre “frequency” y “monetary” ya que podemos intuir que los clientes que compran con más frecuencia también realizan compras de montos más grandes. Por otro lado, la variable “recency”, muestra una correlación baja con las otras 2 variables, denotando un comportamiento más independiente, pudiendo interpretarse como que las compras frecuentes fueron realizadas en las últimas fechas registradas, así como que el monto gastado no depende de la cercanía con la última compra realizada.

También en la Figura 3-29 se graficó histogramas de distribución de las variables, en donde podemos observar un gran sesgo en los valores de “frequency” y “monetary”, a diferencia de “recency” la cual tiene una distribución más uniforme entre sus valores. Debido a estos resultados, fue necesario aplicar un tratamiento de outliers con el método de winsorización y un re-escalado de datos, permitiendo obtener mejores resultados cuando se trate de encontrar el número optimo de clústeres, así como la segmentación por parte de los algoritmos. Una vez aplicado estas mejoras, se procederá a implementar los algoritmos.

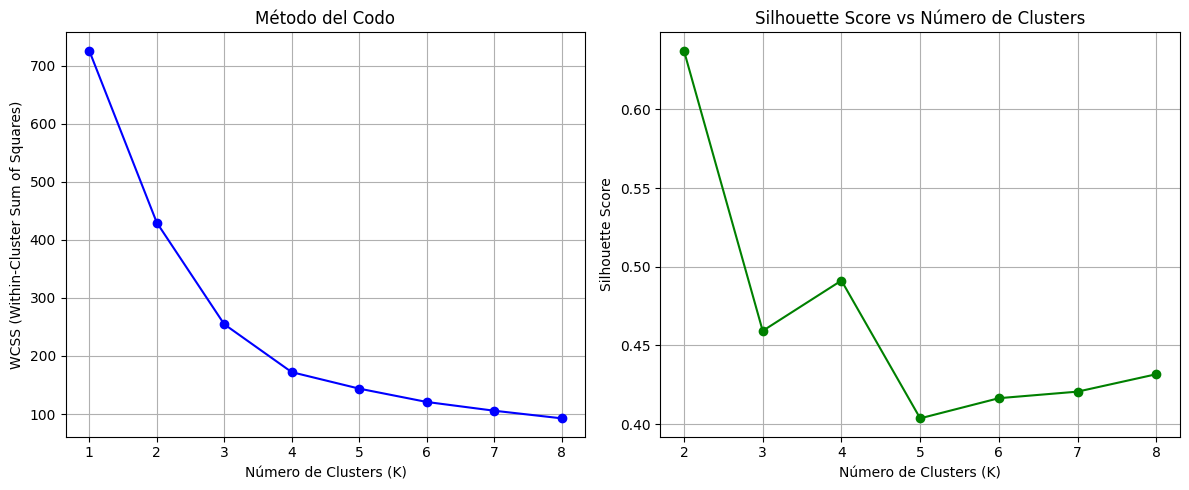


**Figura 3‑29: Gráfica de histogramas de las 3 variables**

**Fuente: (Elaboración propia, 2025)**

### Algoritmo K-Means

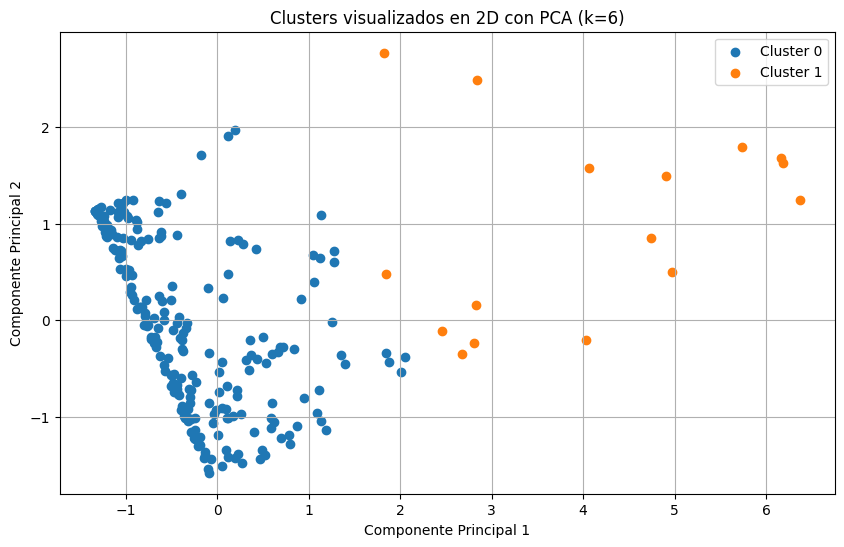
Hecho el re-escaldo de los datos, se continuo con la implementación del método del codo y de Silhouette de la Figura 3-30, para poder hallar el número optimo de clústeres a usar, siendo que en ambos casos es visible que el número recomendado es de dos.



**Figura 3‑30: Método del Codo (izquierda) y Silhouette Score (derecha)**

**Fuente: (Elaboración propia, 2025)**

Por tanto, se procedió a usar el algoritmo de K-Means con 2 clústeres, dando como resultado la gráfica de la Figura 3-31.

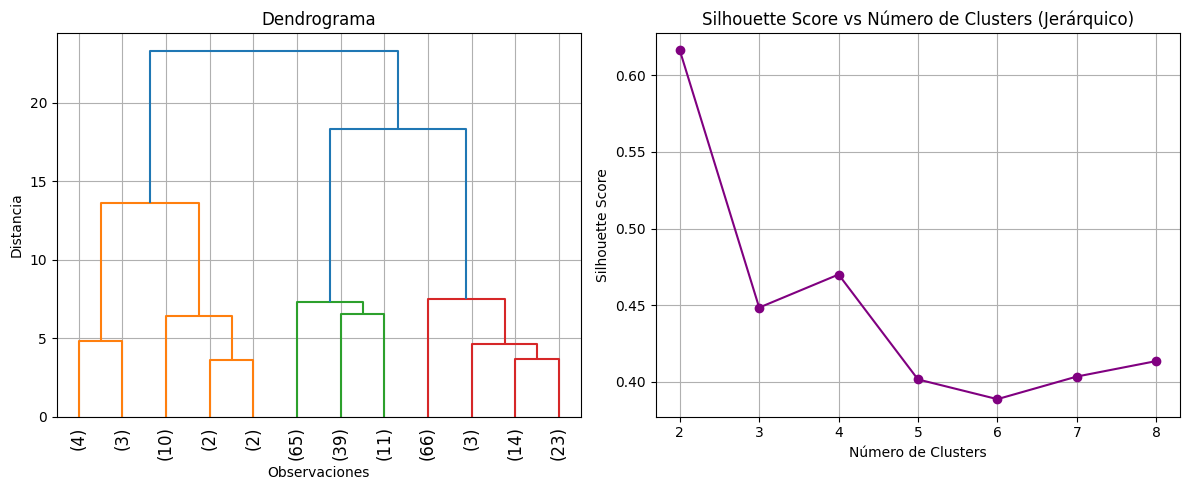


**Figura 3‑31: Clústeres mediante K-Means**

**Fuente: (Elaboración propia, 2025)**

### Algoritmo de Clúster Jerárquico

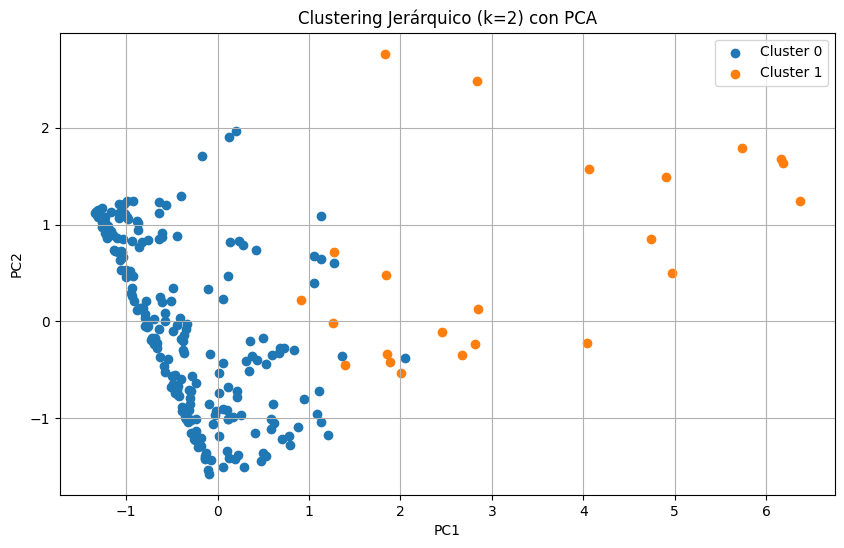
De igual manera, antes de realizar el clúster jerárquico, se usó 2 métodos de obtención de clústeres óptimos. En este caso se usó un gráfico de dendrograma, junto con el método de Silhouette como se ve en la Figura 3-32, obteniendo nuevamente que el número ideal de clústeres seria dos.



**Figura 3‑32: Grafico de Dendrograma (izquierda) y Silhouette Score (derecha)**

**Fuente: (Elaboración propia, 2025)**

Aplicando esta cantidad de clústeres en el algoritmo podemos observar los resultados en el gráfico de la Figura 3-33.

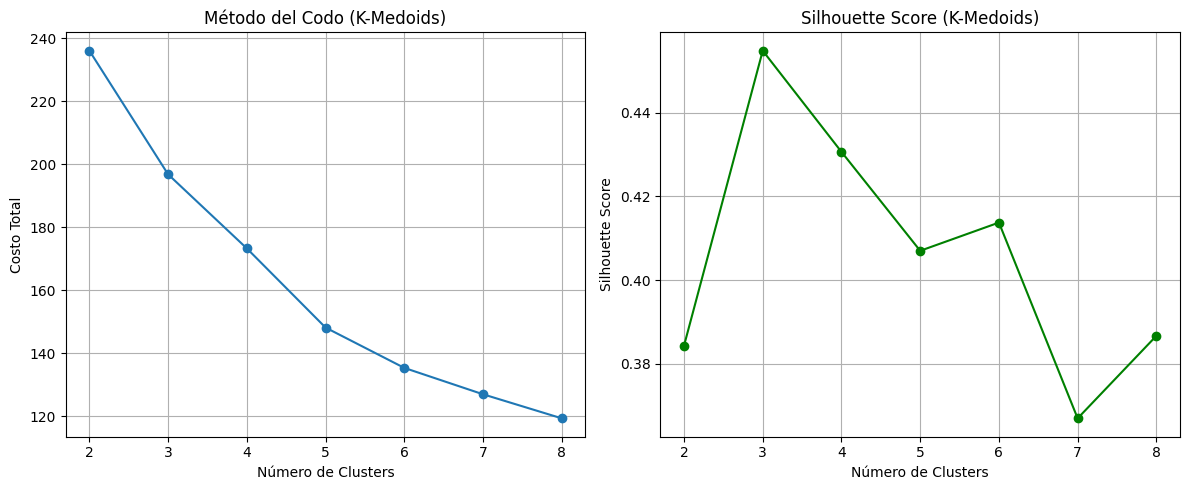


**Figura 3‑33: Resultados de Clúster Jerárquico**

**Fuente: (Elaboración propia, 2025)**

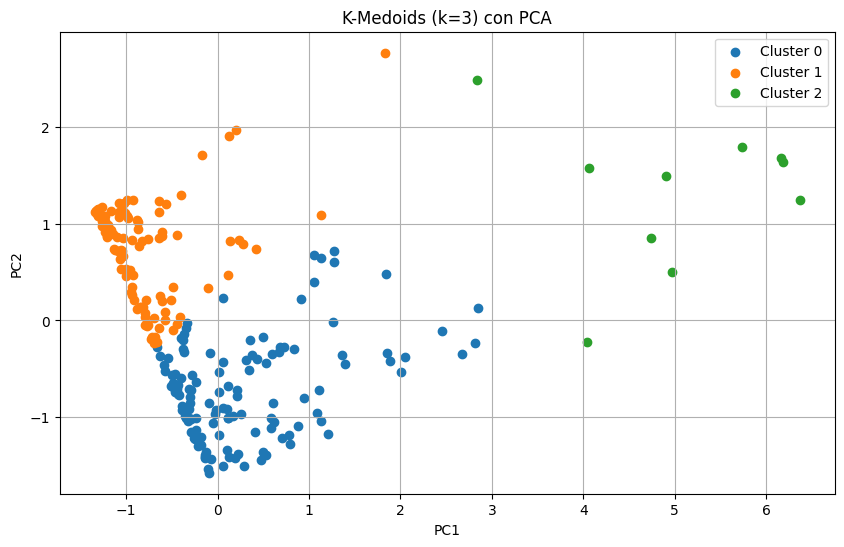
### Algoritmo K-Medoids

Para este último algoritmo, se hizo nuevamente uso de los métodos de codo y puntaje de Silhouette, destacando el valor de 3 en la gráfica del segundo en la Figura 3-34, por tanto, se procedió a configurar el algoritmo con esta cantidad, teniendo como resultado la segmentación de la Figura 3-35.



**Figura 3‑34: Gráficos de Método del Codo y Silhouette Score**

**Fuente: (Elaboración propia, 2025)**

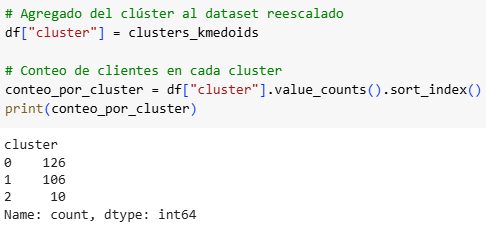


**Figura 3‑35: Resultado de K-Medoids**

**Fuente: (Elaboración propia, 2025)**

## Validación del entendimiento de compra con la segmentación obtenida

Una vez obtenido los resultados del modelo K-Medoids y habiendo seleccionado este como el mejor modelo de entre los tres, se procedió a la obtención de la lista de clústeres obtenidos y la unión de este con el DataFrame usado para la aplicación de los modelos, realizando también un conteo de clientes por clústeres como puede verse en la Figura 3-36.

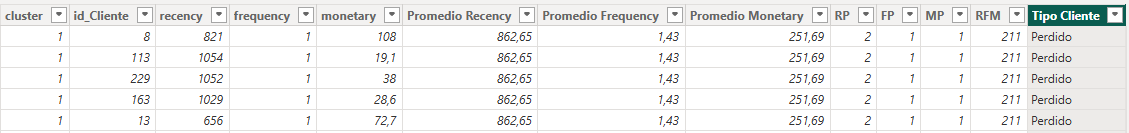


**Figura 3‑36: Obtención de clústeres y conteo de clientes por clúster**

**Fuente: (Elaboración propia, 2025)**

Luego, se descargó estos datos en un archivo llamado “winsorizacion.csv” para su posterior evaluación mediante el uso de la herramienta Power BI. Una vez cargados los datos en una nueva tabla “winsorizacion”, se procedió a realizar una nueva tabla combinada, con la opción “Merge Querys as New” usando la tabla obtenida del modelo, y la tabla “dim\_Entrenamiento” para poder obtener los valores de Recency, Frequency, y Monetary originales de cada cliente mediante su id, separando los mismos por los clústeres encontrados por el modelo de K-Medoids.

Se hizo el cálculo de los promedios de cada variable, así como de los puntajes individuales y RFM en conjunto, para poder asignar un tipo de cliente a cada clúster, obteniendo los tipos de cliente: “Nuevo”, “Perdido” y “Con Potencial” como puede verse en la Figura 3-37.



**Figura 3‑37: Tabla Merge, para el análisis de los tipos de cliente**

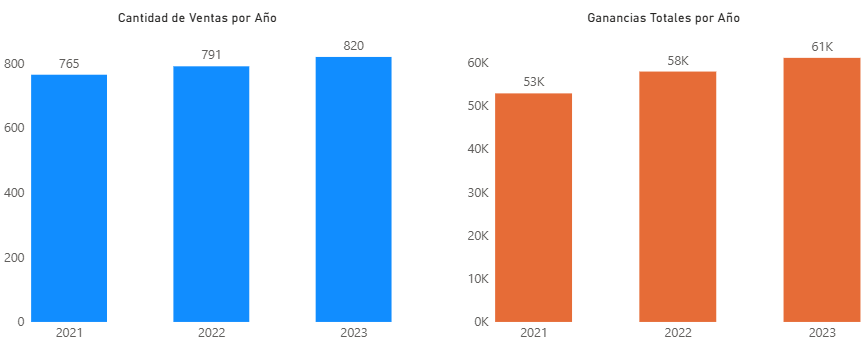
**Fuente: (Elaboración propia, 2025)**

Con estos datos a mano, se pudo continuar con la realización de las gráficas que permitirían visualizar datos más detallados de cada tipo de clúster, los cuales serán analizados en el siguiente capítulo.

# Análisis de Resultados y Discusión

## Resultados de la preparación de datos

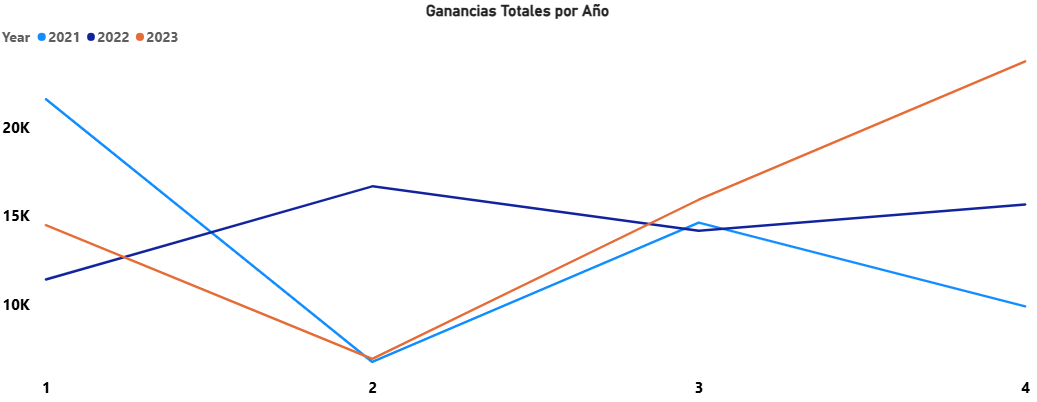
Una vez realizada la limpieza de los datos y la creación de las tablas con información de los clientes, los productos y las ventas, se pudo obtener observaciones mediante una exploración de los datos ya ordenados del dataset obtenido de la empresa comercializadora. De esta manera, se pudo observar que tanto la cantidad de ventas, así como las ganancias tuvieron un incremento durante los tres años que abarca el dataset, llegando a un ingreso de 61 mil Bs en el año 2023, como puede verse en la Figura 4-1.



**Figura 4‑1: Cantidad de Ventas por Año (izquierda), Ganancias Totales por Año (derecha)**

**Fuente: (Elaboración propia, 2025)**

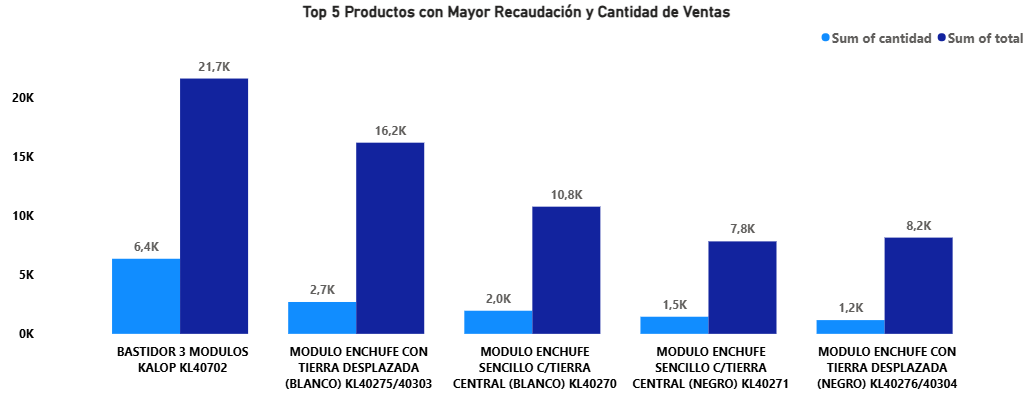
No solamente esto, en la Figura 4-2 obtenemos una vista por cuartiles durante los tres años, en la que podremos notar que existe una tendencia en la que se registran una gran cantidad de ingresos en el primer trimestre de los años 2021 y 2023. Solo en el año 2022 es que los ingresos mantuvieron un ritmo casi constante a lo largo de todo el año, teniendo un pico en el segundo trimestre del año. Cabe resaltar también que, en el año 2021, se registraron dos caídas en las ventas durante el segundo y el cuarto semestre del año. Mientras que, en el año 2023, después del segundo semestre, hubo un aumento exponencial en las ventas, por la cual es el año con mayor recaudación.



**Figura 4‑2: Ganancias Totales por Trimestre de cada Año**

**Fuente: (Elaboración propia, 2025)**

También se pudo obtener de este dataset, cuáles son los cincos mejores productos, por la cantidad de unidades vendidas, así como por las ganancias obtenidas durante los tres años. Estos productos son: Bastidor 3 Modulos Kalop con más de 6 mil unidades vendidas y una ganancia de más de 21 mil Bs, continuando con los Módulos de Enchufe con Tierra Desplazada Blanco y Negro, y los Módulos de Enchufe Sencillo c/Tierra Central Blanco y Negro, como pueden apreciarse en la Figura 4-3.

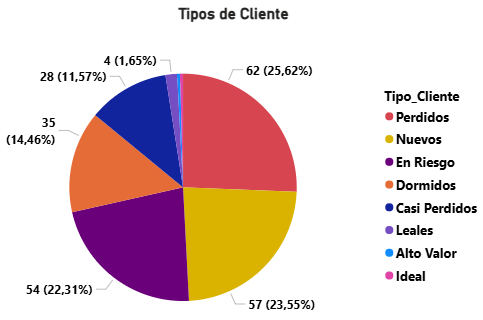


**Figura 4‑3: Top 5 de los Productos Más Vendidos y con Mayor Cantidad de Unidades Vendidas**

**Fuente: (Elaboración propia, 2025)**

## Resultados del análisis RFM

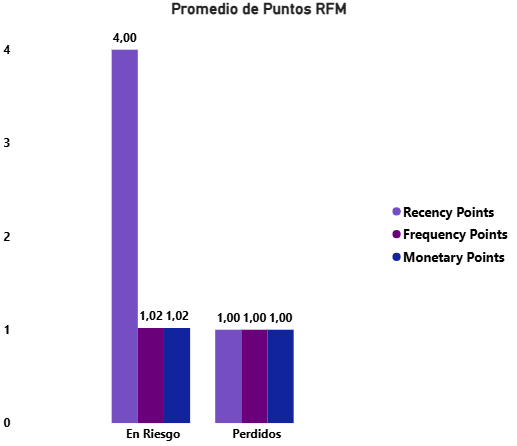
Posteriormente se aplicó el análisis RFM al dataset, pudiendo identificar ocho tipos diferentes de clientes según las métricas obtenidas de Recency (Cantidad de días desde la última compra), Frequency (Cantidad de compras hechas por un cliente) y Monetary (Monto total de las compras de un cliente). Entre estos valores podemos destacar la gran cantidad de clientes etiquetados como Perdidos y en Riesgo, debido a que representan casi la mitad de los clientes totales con un 47.93% como se puede observar en la Figura 4-4.



**Figura 4‑4: Representacion percentil de la cantidad de cada tipo de cliente**

**Fuente: (Elaboración propia, 2025)**

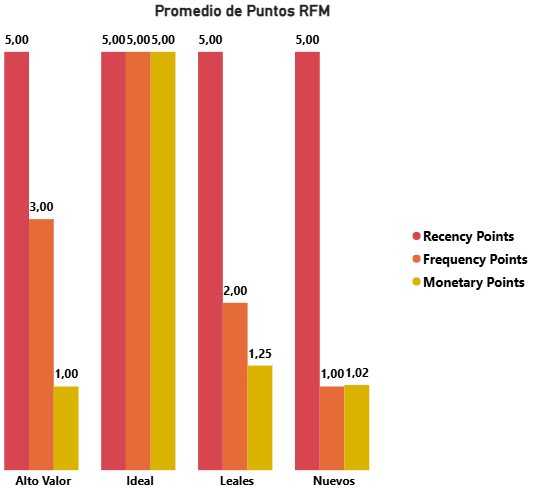
Esto indica que los clientes de la empresa, almenos aquellos que realizan compras de productos dirigidos a construcción, comprenden en su mayoría a personas que compraron muy pocas veces y un monto bajo. Pese a contarse entre ellos, a clientes catalogados como “En Riesgo”, los cuales tienen un valor de Recency alto (4), lo cual significa que sus compras fueron bastante recientes, como se aprecia en la Figura 4-5.



**Figura 4‑5: Promedio de Puntajes RFM para Clientes “En Riesgo” y “Perdidos”**

**Fuente: (Elaboración propia, 2025)**

Por otro lado, podemos ver que la diferencia con los puntajes de los clientes valorados como de Alto Valor, Ideal, Leales y Nuevos, radica en que destacan valores altos en el puntaje RFM, mayormente en la variable Recency con un puntaje de 5 como se ve en la Figura 4-6. Se debe resaltar que la diferencia entre los clientes Nuevos y En Riesgo es justamente debido a esta variable (Recency), que en este caso distan por un punto, pero es necesario hacer notar que, este punto representa una diferencia de casi un año desde la última compra, ya que los Nuevos clientes, estarían abarcando un lapso máximo de 7 meses desde su última compra.

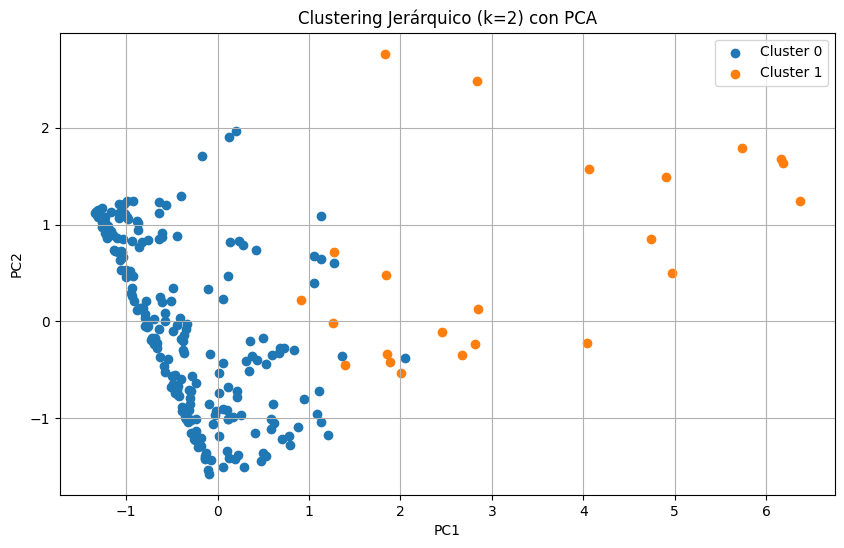
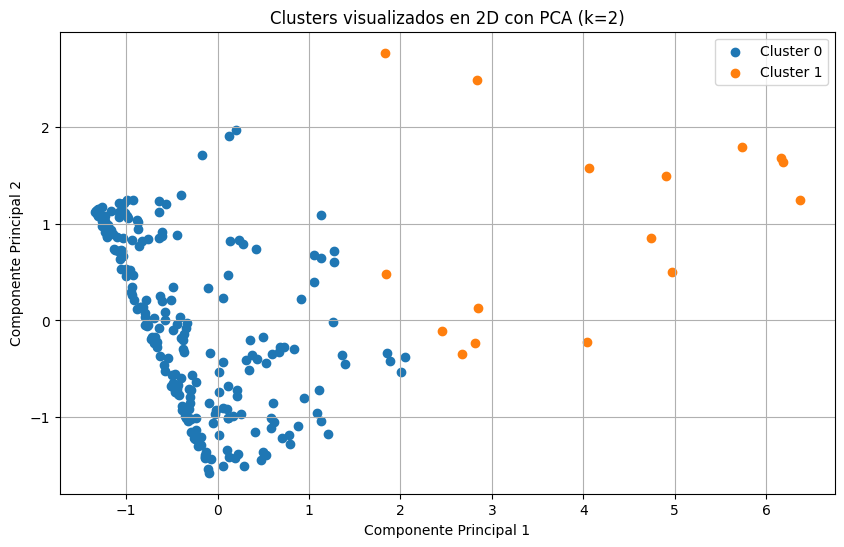


**Figura 4‑6: Promedio de Puntajes RFM para Clientes "Alto Valor", "Ideal", "Leales" y "Nuevos"**

**Fuente: (Elaboración propia, 2025)**

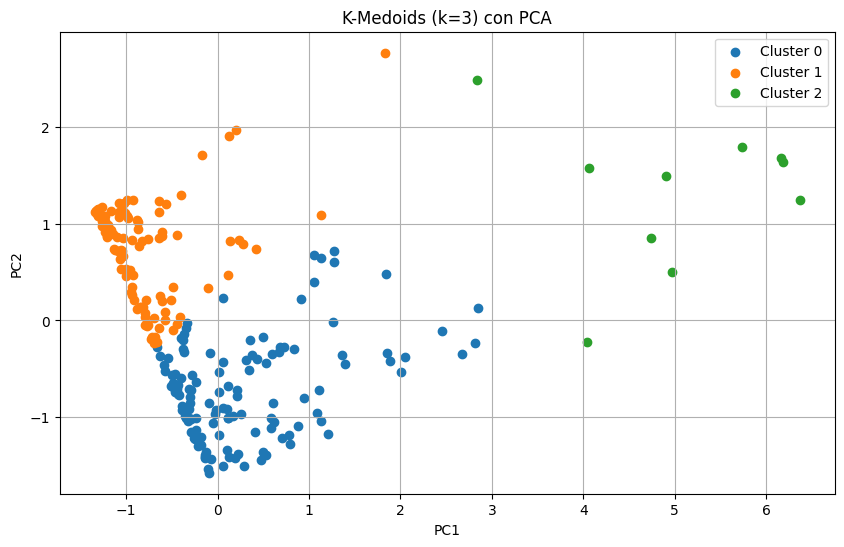
## Resultados de la elaboración de los modelos de aprendizaje no supervisado

Para los resultados de los modelos K-Means, Clustering Jerarquico y K-Medoids, se tomó en cuenta principalmente, los resultados de las gráficas de Silhouette Score, ya que muestran más claramente cuál es la cantidad de clústeres ideales para lograr una mayor calidad de agrupamiento en cada modelo. Por este motivo es que se eligió el valor de K=2 para K-Means (SS=0.64) y Clustering Jerárquico (SS=0.61) y K=3 para K-Medoids (SS=0.45), obteniendo así las Figuras 4-7 y 4-8, mediante visualización por PCA, en donde se puede ver la segmentación de cada modelo en base a los clústeres definidos.



**Figura 4‑7: Segmentación mediante K-Means y Clustering Jerárquico con K=2**

**Fuente: (Elaboración propia, 2025)**



**Figura 4‑8: Clustering mediante K-Medoids, K=3**

**Fuente: (Elaboración propia, 2025)**

Pese al tratamiento realizado para evitar tener valores con outliers mediante el uso de Winsorización, se puede apreciar que en la visualización de todos los modelos existen puntos bastantes distanciados, lo que evidenciaría la existencia de valores que aun podrían considerarse outliers. A pesar de esto, se mantuvo estos datos, ya que debido al uso del análisis RFM, estos podrían llegar a representar clientes importantes dentro de los rangos más extremos.

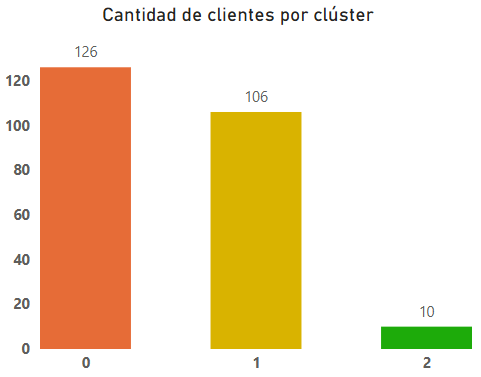
Para K-Means se puede ver una separación bastante clara de los dos grupos resultantes, mostrando la mayoría de los puntos dispersos dentro de uno de estos, siendo una división razonable de los datos por parte del modelo, posicionándolo como una opción válida para una segmentación simple de los clientes.

En el caso del Clustering Jerárquico, pese a que se asignó el valor de K=2, se pude notar que no existe una división clara entre los puntos, ya que en la parte central estos aparecen mezclados. Esto puede deberse a que el método de enlace no era el más adecuado, ya que no fue capaz de realizar una buena separación entre los valores recibidos.

Finalmente, el modelo K-Medoids, que es más robusto que el modelo K.Means frente a outliers, selecciona mejor sus medoides para realizar las agrupaciones, mostrando una mejor diferenciación entre los grupos, a pesar de tener el valor más bajo en cuanto a Silhouette Score (0.45) y del grupo determinado como el Clúster 2, el cual consta de pocos puntos y bastante dispersos entre sí, el cuál podría interpretarse como un grupo que abarca los clientes con valores más extremos dentro de los valores de RFM.

## Resultados de la validación del entendimiento de compras de los clientes

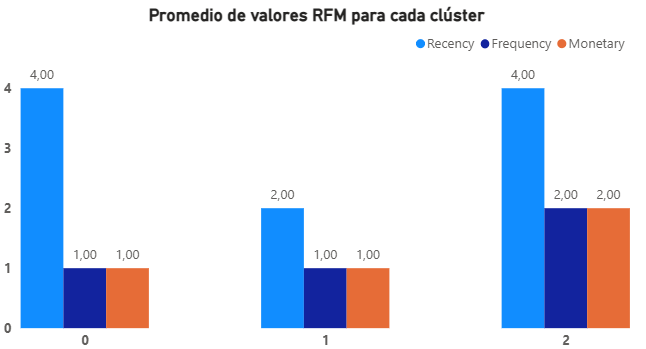
Tomando en cuenta los resultados obtenidos de los tres modelos mencionados en el punto anterior, se determinó usar el clustering del algoritmo K-Medoids con valor de K=3, debido a que mostraba una mayor cantidad de clústeres, así como una mejor delimitación en comparación a los otros dos modelos. De esta manera se procedió a revertir los valores re-escalados de la tabla de entrenamiento añadiendo los clústeres para poder visualizar la cantidad de clientes que pertenece a cada uno, notando como en la Figura 4-9, la distribución se ve más equilibrada entre dos de los tres segmentos, siendo que el Clúster 0 comprende 126 clientes, 106 clientes en el Clúster 1 y 10 clientes en el Clúster 2.



**Figura 4‑9: Cantidad de Clientes Pertenecientes a Cada Cluster obtenido por K-Medoids**

**Fuente: (Elaboración propia, 2025)**

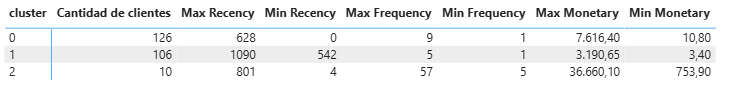
Para poder obtener más información respecto a cada clúster, se obtuvo el promedio de los valores RFM, para luego darles un puntaje del 1 al 5, como se había hecho previo a su tratamiento, notando en este caso, que los clústeres 0 y 2 son bastante similares en cuanto a puntaje RFM se trata, diferenciándose por un punto en las variables Frequency y Monetary, como se ve en la Figura 4-11.



**Figura 4‑10: Porcentaje Atribuido a cada Tipo de Cliente**

**Fuente: (Elaboración propia, 2025)**

Ahondando un poco más en los valores de recencia, frecuencia, y valor monetario, en la Tabla 4-1 podemos ver los parámetros máximos y mínimos entre los que se encuentra cada clúster



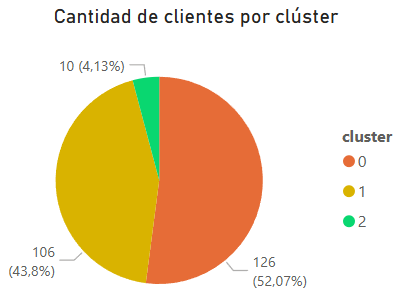
**Tabla 4‑1: Valores Máximos y Mínimos de RFM**

**Fuente: (Elaboración propia, 2025)**

Con estos valores presentes podemos resaltar que:

* **Clúster 0:** Presenta un alto valor de recencia, indicando que las compras fueron recientes, pero con baja frecuencia y bajos montos de compra, por tanto, podría considerarse como Clientes Nuevos.
* **Clúster 1:** Este grupo, por los bajos valores obtenidos en el puntaje RFM, podrían considerarse como Clientes Perdidos, ya que son antiguos e inactivos, debido a que ha transcurrido bastante tiempo desde su última compra, y tanto la frecuencia como el monto eran bajos.
* **Clúster 2:** Similar al Clúster 0, se diferencian por una mayor frecuencia de compra y gasto entre sus miembros, pudiendo interpretarse como Clientes con Potencial para la empresa, en vista de que, con un enfoque en una futura fidelización, se podría asegurar su frecuencia de compra, así como su inversión en los productos de la empresa.

Finalmente, podemos mencionar que tanto el Clúster 0 como el 1, representan la mayor parte de los clientes, en especial el primero que abarca más del 50% como se ve en la Figura 4-11, lo que indicaría que la mayor parte de los clientes de la empresa no son constantes y en su mayoría son clientes nuevos o esporádicos.



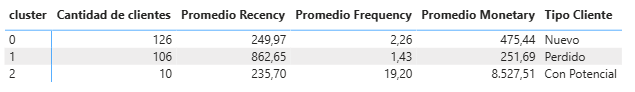
**Figura 4‑11: Porcentajes de clientes por clúster**

**Fuente: (Elaboración propia, 2025)**

## Discusión de resultados

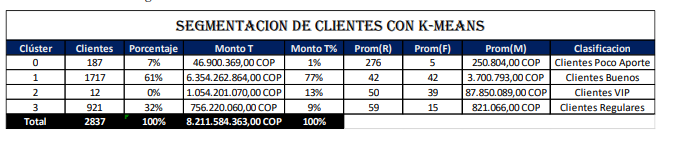
Los resultados obtenidos en el trabajo “Implementación de un modelo de clusterización mediante la segmentación de perfil de clientes para corporación multi inversores” de Ana Carrillo y Emili Flores fueron obtenidos aplicando también el análisis RFM en los resultados obtenidos mediante el modelo de K-Means a una población de 2837 clientes (Carrillo García & Flores Velásquez, 2024).

Realizando una comparativa entre la distribución de los clientes podemos notar que los resultados obtenidos en este proyecto, mediante el uso de K-Medoids, identificaron a tres grupos a partir de 242 clientes, con una distribución relativamente equilibrada entre 100 y 120 clientes a excepción de un grupo que cuenta con 10 como puede verse en la Tabla 4-2. En cambio, la Tabla 4-3, muestra los resultados del otro estudio, donde obtuvieron 4 clústeres con una distribución mucho más desigual, con más del 60% de sus clientes agrupados en el Clúster 1, mientras que el Clúster 2 contiene apenas 12 clientes. Esto podría señalar una segmentación más equilibrada y manejable entre los diferentes grupos debido al uso del modelo K-Medoids.



**Tabla 4‑2: Valores obtenidos en el presente proyecto**

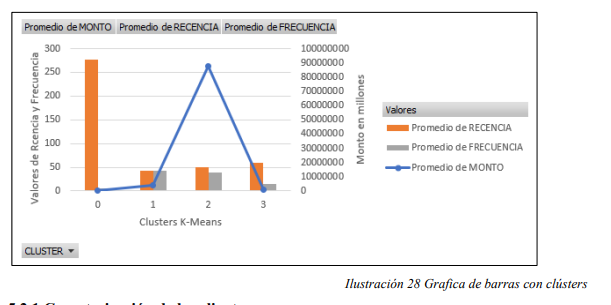
**Fuente: (Elaboración propia, 2025)**



**Tabla 4‑3: Valores del Proyecto Comparativo**

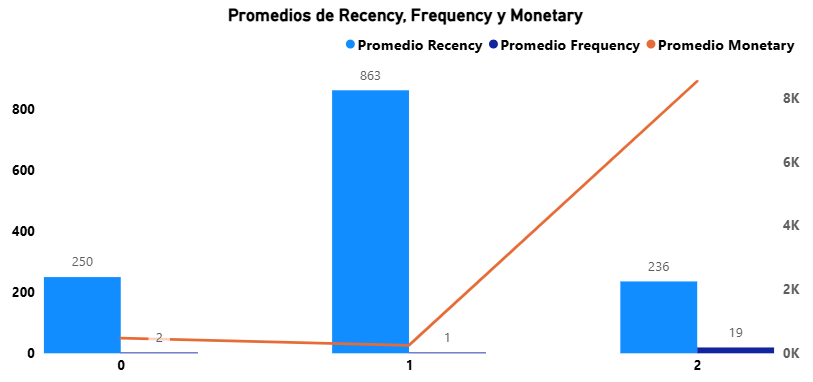
**Fuente: (Carrillo García & Flores Velásquez, 2024)**

En las Figuras 4-12 y 4-13 se puede observar la relación que existe entre los valores de RFM, en donde los Clientes con Potencial obtenidos y los Clientes VIP del otro estudio destacan una similitud, ya que en ambos casos los tamaños de los grupos son pequeños, lo que refleja la importancia de aplicar estrategias diferenciadas para estos clientes de mayor importancia. También es importante resaltar que, en ambos estudios, la mayor parte de los ingresos provienen de clientes recientes que a pesar de no ser muy frecuentes, sobre todo los analizados en este proyecto, realizan compras significativas dentro de la empresa.

****

**Figura 4‑12: Relación de Promedios del Proyecto Comparativo**

**Fuente: (Carrillo García & Flores Velásquez, 2024)**



**Figura 4‑13: Relación de Promedios de los valores RFM**

**Fuente: (Elaboración propia, 2025)**

En resumen, la segmentación de los grupos es más favorable en cuanto al uso del modelo K-Medoids, debiéndose a que el modelo trabaja mejor encontrando grupos más representativos y menos influenciados por valores atípicos en comparación al uso de K-Means. Esto puede ser más útil influenciando en una mejor interpretación de los tipos de clientes para las estrategias de marketing que pueda implementar las empresas objetivo. Otro punto importante a resaltar es la cantidad de clientes de cada estudio, siendo que la aplicación del modelo K-Medoids, podria ofrecer mejores resultados en poblaciones grandes como las manejadas en el estudio comparativo.

# Conclusiones

La finalidad de este proyecto yace en segmentar los clientes de la empresa Drustvo S.R.L., ayudándose de métricas del análisis RFM para poder identificar grupos diferenciados que permitiesen aportar una mayor visión en cuanto a sus comportamientos, y de esta manera implementar mejores estrategias comerciales y más efectivas. Con este fin es que se hizo uso de modelos de aprendizaje no supervisado usando variables que median que tan reciente había sido la última compra (Recency), la frecuencia con la que compraban los clientes (Frequency), y el valor monetario total que habrían invertido en sus compras (Monetary).

Para tal efecto, se realizó una preparación de los datos provistos de la empresa, los cuales abarcaban registros de ventas de productos, en su totalidad de la categoría “Construcción”, entre los años 2021 y 2023. De los cuales, una vez hecha una limpieza y una estructuración de tres tablas (Clientes, Productos, y Ventas), se pudo observar que sus ingresos fueron creciendo cada año, llegando a registrar más de 60 mil bolivianos en el año 2023, siendo el producto más vendido el “Bastidor 3 Modulos Kalop” con más de 6 mil unidades vendidas, y 242 clientes registrados a lo largo de los tres años.

Posterior a esto, se aplicó un análisis RFM al dataset, el cual, calcula que tan recientes fueron las compras, que tan frecuente son realizadas y que montos totales comprenden. Con estos valores se obtuvo un puntaje por el cual se pudo dividir a los clientes de la empresa, obteniendo 8 categorías, de las cuales, las etiquetadas como Clientes Perdidos o en Riesgo comprendían un 47.93%, siendo que estos representaban a clientes que habrían realizado compras hace bastante tiempo, muy pocas veces y de poco valor. Por otro lado, los clientes considerados Ideales, de Alto Valor o Leales, comprendían un porcentaje mínimo (2.49%), debido a que juntos representaban un total de 6 de los 242 clientes dentro de los registros.

Una vez obtenidos los valores de Recency, Frequency y Monetary, junto con sus puntajes, se procedió a hacer uso de tres métodos de clusterización: K-Means, Clustering Jerárquico y K-Medoids, para poder comparar los resultados obtenidos de cada uno y así elegir cual podría ofrecer un mejor enfoque, tanto en la cantidad de agrupaciones, como en la claridad de la división de los segmentos. Aplicándose el método Silhouette Score para evaluar la segmentación entre grupos de clústeres, y gráficas del análisis PCA para visualizar las divisiones de los mismos, observando que:

* El modelo K-Means mostro una separación bastante clara en sus 2 clústeres, comprobando que el puntaje obtenido por Silhouette Score (0.64) era bastante acertado. Pero para este caso, lo que se buscaba era un poco más de variedad en cuanto a la cantidad de grupos hallados, siendo este el motivo para descartar este modelo como el mejor.
* El modelo de Clustering Jerárquico mostraba relaciones bastante interesantes en su gráfica de dendrogramas, y pese a obtener un puntaje de 0.61 en Silhouette Score, estas no pudieron ser reflejadas en los resultados finales, ya que no mostraba una separación adecuada entre los dos grupos, al no existir una delimitación clara entre ambos en la gráfica obtenida por PCA, siendo la razón por la que este modelo también fuera descartado como la mejor opción.
* Finalmente, el modelo K-Medoids, a diferencia de sus contrapartes, mostró una inclinación por el uso de tres clústeres, pese a tener un menor valor en cuanto al Silhouette Score (0.45) en comparación a los otros modelos. Sin embargo, se pudo observar una mejor delimitación entre estos grupos en la gráfica obtenida por PCA. Siendo esta la razón por la que se decidió elegir este modelo como el mejor.

Para finalizar, se hizo un análisis de los clústeres obtenidos por el modelo K-Medoids, donde se pudo observar que 126 clientes pertenecían al Clúster 0, 106 clientes al Clúster 1 y 10 al Clúster 2.

Aplicando los puntajes y valores de RFM a estos clientes, se les pudo agregar las etiquetas de Clientes Nuevos, Perdidos y con Potencial respectivamente. Destacando que:

* Los Clientes Nuevos tienen un promedio de 249,97 en recencia, 2,26 de frecuencia y 475,44 en valor monetario, obteniendo un puntaje RFM de 411.
* Los Clientes Perdidos contaban con un promedio de 862,65 en recencia, 1,43 en frecuencia y 251,69 en valor monetario, con un puntaje RFM de 211.
* Y los Clientes con Potencial cuentan con un promedio de 235,7 en recencia, 19,20 en frecuencia y 8527,51 en valor monetario, con un puntaje RFM final de 422.

Esto significaría que, si bien la empresa cuenta con una buena cantidad de ingresos y ventas de sus productos, estos pertenecen en su mayoría a clientes esporádicos, que en pocas ocasiones hacen compras bastante significativas. Gracias a que estas agrupaciones son más equilibradas, en comparación con las obtenidas mediante el análisis RFM, se puede determinar mejores acciones a tomar para cada segmento, demostrando así la utilidad de la aplicación de modelos de aprendizaje no supervisado y un mejor entendimiento en cuanto al comportamiento de cada grupo de clientes.

# Recomendaciones

A partir de los resultados obtenidos y el análisis realizado en el capítulo 4, se obtuvieron tres grupos/tipos de clientes dentro de la empresa. Para los Clientes Perdidos, se recomienda hacer uso de campañas de reenganche, que puedan volver a atraer su interés, mediante promociones de artículos que solían comprar más.

En el caso de los Clientes Nuevos**, s**e recomendaría aplicar un programa de bienvenida, cupones o descuentos incentivando a futuras compras, así como un monitoreo en cuanto a su evolución, pudiendo ascenderlos a clientes leales, en el mejor de los casos.

En cuanto a los Clientes con Potencial**,** el objetivo principal seria lograr una fidelización de este grupo de clientes, pudiendo ofrecer programas de recompensas o descuentos preferenciales, ya que este grupo comprende aquellos clientes que generan mayores ingresos, para la empresa, siendo el motivo principal para tratar de convertirlos en clientes recurrentes.

Se debe mencionar que el dataset abarcaba el sector de productos de construcción que ofrece la empresa, y el análisis y las recomendaciones dadas pertenecen al mismo. En caso de desear un análisis a mayor escala, se necesitaría de un dataset que abarque todos los sectores para poder tener una visión total de los clientes de la empresa, y obtener también mejores resultados pudiendo obtener un análisis diferenciado para cada sector.

También es necesario mencionar que, se usó el método de winsorización y no otro método de eliminación de outliers debido a que estos llegan a representar clientes importantes que deben ser tomados en cuenta por los modelos para realizar una mejor segmentación, ya que al usarse los valores obtenidos del análisis RFM, los valores extremos notados en las gráficas de PCA, demuestran que pertenecen a grupos con valores altos, como en este caso, que agrupaban a clientes considerados con Potencial.

Finalmente, el uso de K-Medoids fue bastante útil en este proyecto debido a la cantidad reducida de clientes, ya que, si este fuera más grande también tendría un mayor coste computacional, debido a los procesos que realiza el modelo para hallar los centros de cada segmentación. Un punto negativo siendo que es el más robusto y de mejores resultados ante datos con outliers.

# Bibliografía

AnalytixLabs. (19 de Junio de 2024). *Customer Segmentation with Machine Learning: Targeting the Righ Audience*. Obtenido de Medium: https://medium.com/%40byanalytixlabs/customer-segmentation-with-machine-learning-targeting-the-right-audience-656f5d2ce8f8

Berrio Lasprilla, J. A., & Olea Gómez, O. J. (2024). *Universidad de Antioquia.* Obtenido de Segmentación de clientes mediante análisis de patrones de compra para la optimización de estrategias comerciales: https://bibliotecadigital.udea.edu.co/entities/publication/290eb42c-21e3-454b-b033-c26e56f9d3ef

Carrillo García, A. C., & Flores Velásquez, E. G. (Febrero de 2024). Obtenido de UNITEC Centro de Recursos para el Aprendizaje y la investigación: https://repositorio.unitec.edu/items/37ea4ac0-fca9-489a-9940-e8347203e5c4

Ecofinanzas, & El Deber. (25 de Agosto de 2023). *La Inteligencia Artificial 4.0 en la actividad industrial*. Obtenido de Money: https://www.money.com.bo/ecofinanzas/la-inteligencia-artificial-4-0-en-la-actividad-industrial/

Google Cloud. (s.f.). *¿Qué es el aprendizaje no supervisado?* Obtenido de Google Cloud: https://cloud.google.com/discover/what-is-unsupervised-learning?hl=es-419

Hernández, J. (22 de Septiembre de 2022). *Factor Trabajo*. Obtenido de Inteligencia artificial: qué aporta y qué cambia en el mundo del trabajo: https://blogs.iadb.org/trabajo/es/inteligencia-artificial-que-aporta-y-que-cambia-en-el-mundo-del-trabajo/

IBM. (s.f.). *¿Qué es el aprendizaje no supervisado?* Obtenido de IBM: https://www.ibm.com/mx-es/topics/unsupervised-learning?mhsrc=ibmsearch\_a&mhq=Qu%26eacute%3B%20es%20el%20aprendizaje%20no%20supervisado

Jiménez, B. (6 de Octubre de 2023). *Programaticaly*. Obtenido de ¿Cómo está transformando el machine learning la segmentación de anuncios?: https://www.programaticaly.com/portada/como-machine-learning-transformando-segmentacion-anuncios

Kumar, D. (19 de Diciembre de 2023). *Implementing Customer Segmentation Using Machine Learning [Beginners Guide]*. Obtenido de Neptune.AI: https://neptune.ai/blog/customer-segmentation-using-machine-learning

Mining, E. (2019). *Machine Learning for Beginners.*

Morelo Tapias, K. A. (2014). *Sistema para Caracterización de Perfiles de Clientes de la Empresa Zona T.* Obtenido de Universidad de Cartagena: https://repositorio.unicartagena.edu.co/entities/publication/772465a5-e837-472f-9fcc-3e9a47ca6732

*Oracle*. (s.f.). Obtenido de ¿Qué es el machine learning?: https://www.oracle.com/co/artificial-intelligence/machine-learning/what-is-machine-learning/

Palacios Abadía, F. A., & Pastor Patiño, N. A. (Abril de 2020). Obtenido de Fundación Universitaria de Popayán - Repositorio: https://fupvirtual.edu.co/repositorio/files/original/8c2c34d9830a14dba03a7b38c9b408a10b966abf.pdf

Ponce Gallegos, J. C., Torres Soto, A., Quezada Aguilera, F. S., Silva Sprock, A., Martínez Flor, E. U., Casali, A., . . . Pedreño, O. (2014). *Inteligencia Artificial.* Iniciativa Lationamericana de Libors de Texto abiertos (LATIn).

PureStorage. (2024). *¿Qué es un proceso de aprendizaje automático?* Obtenido de PureStorage: https://www.purestorage.com/la/knowledge/what-is-machine-learning-pipeline.html

Rojas, E. M. (Abril de 2020). *Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo.* Obtenido de ProQuest: https://www.proquest.com/docview/2388304894/9603C69F7DE046F8PQ/1?sourcetype=Scholarly%20Journals

Saleh, R., Majzoub, S., & Saleh, A. M. (2025). *Fundamental of Robust Machine Learning.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Smolic, H. (4 de Marzo de 2024). *How to Use Machine Learning for Customer Segmentation*. Obtenido de Medium: https://hrvoje-smolic.medium.com/how-to-use-machine-learning-for-customer-segmentation-49612667301d

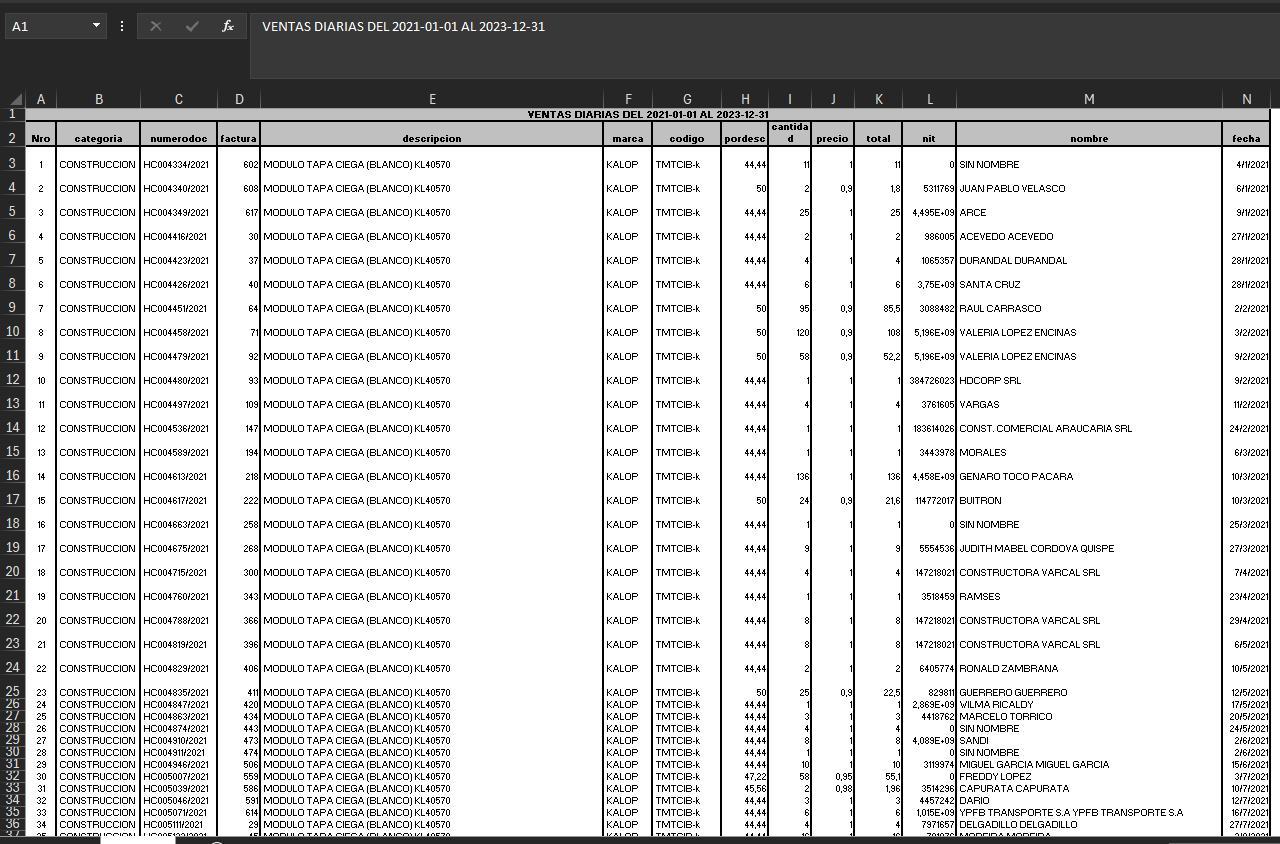
Sosa Sierra, M. (2007). *Inteligencia artidicial en la gestión financiera empresarial.* Barranquilla, Colombia: Pensamiento & Gestión.

Vaidya, V. (16 de Mayo de 2022). *Nisum*. Obtenido de Segmentar a los clientes utilizando el apredizaje automático en 2020 y después: https://www.nisum.com/es/nisum-knows/segment-customers-by-using-machine-learning-in-2020-and-beyond

Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). *BibSonomy.* Obtenido de Constrined K-means Clustering with Background Knowledge: https://www.bibsonomy.org/bibtex/27b3384929286f957c3152f7825206073/hotho

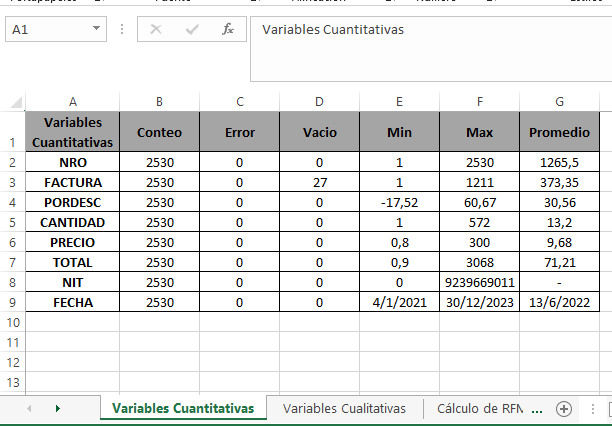
# Anexos

## Archivo Excel con los registros de ventas de la empresa



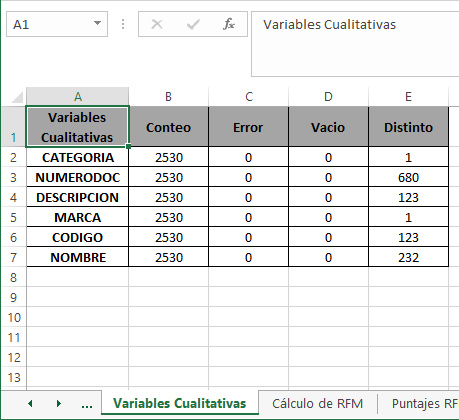
Ubicación: CD: DATA Y VISUALIZACIÓN/Datos\_empresa.xlsx

## Tabla de las variables cuantitativas



Ubicación: CD: TABLAS/Tablas de Proyecto.xlsx

## Tabla de las variables cualitativas



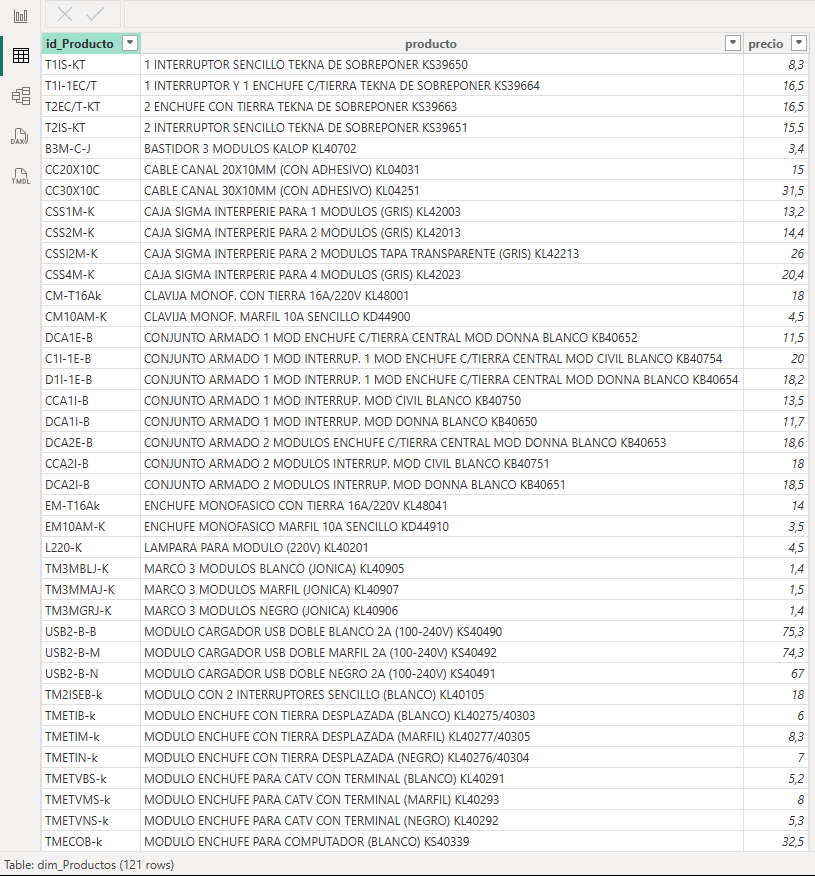
Ubicación: CD: TABLAS/Tablas del Proyecto.xlsx

## Tabla de Clientes obtenida de la preparación de datos



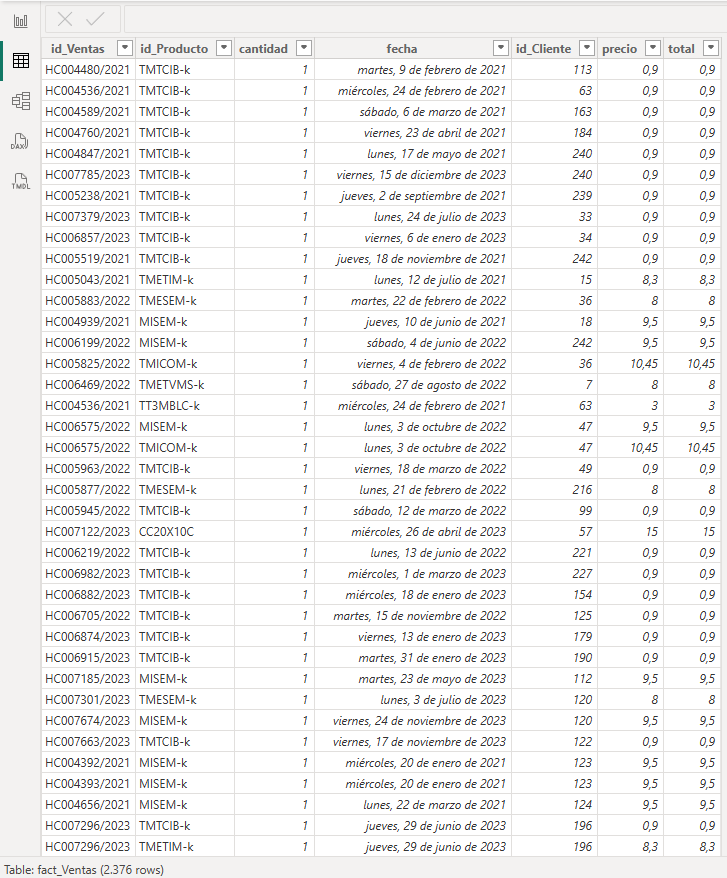
Ubicación: CD: DATA Y VISUALIZACIÓN/Proyecto Diplomado.pbix

## Tabla de Productos obtenida de la preparación de datos



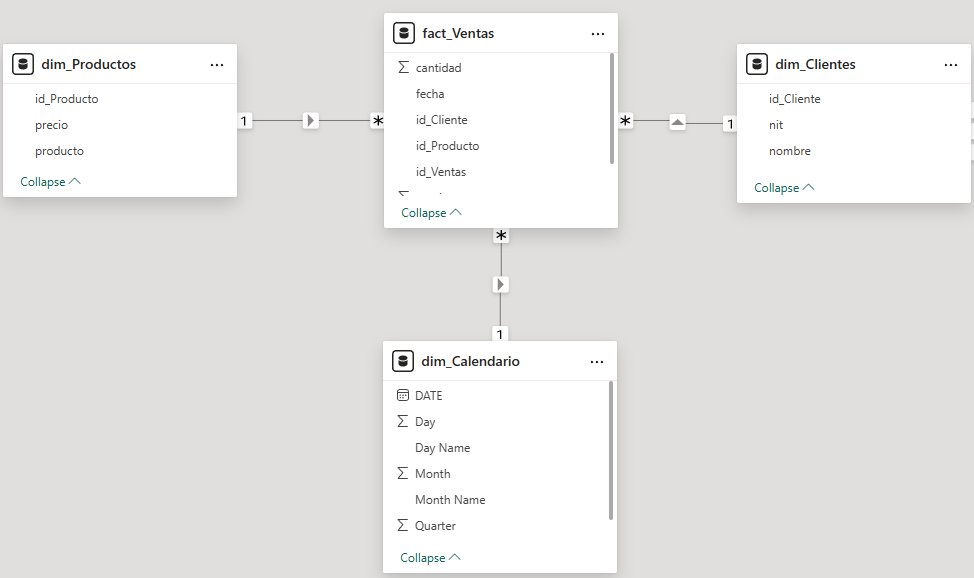
Ubicación: CD: DATA Y VISUALIZACIÓN/Proyecto Diplomado.pbix

## Tabla de Ventas obtenida de la preparación de datos



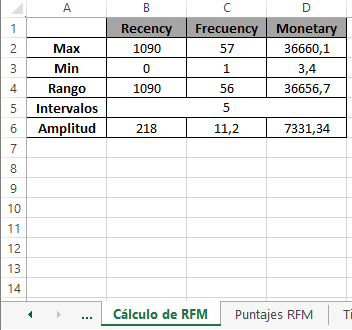
Ubicación: CD: DATA Y VISUALIZACIÓN/Proyecto Diplomado.pbix

## Modelo Relacional de las tablas Clientes, Productos y Ventas



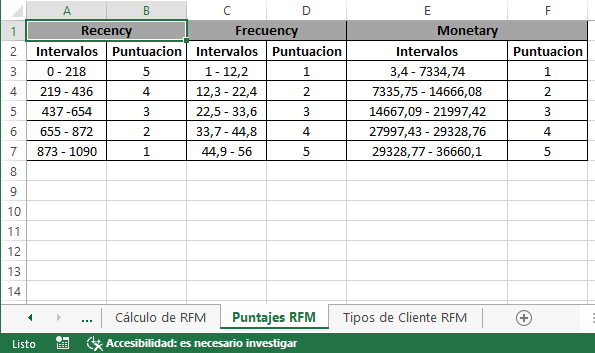
Ubicación: CD: DATA Y VISUALIZACIÓN/Proyecto Diplomado.pbix

## Tabla del cálculo de los valores RFM



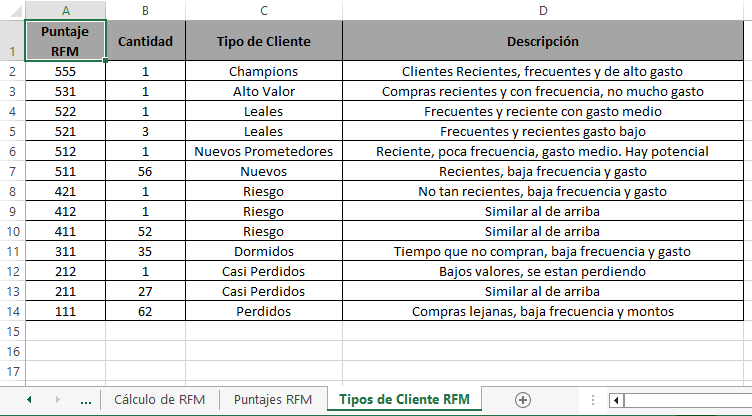
Ubicación: CD: TABLAS/Tablas de Proyecto.xlsx

## Tabla de los intervalos y puntajes RFM



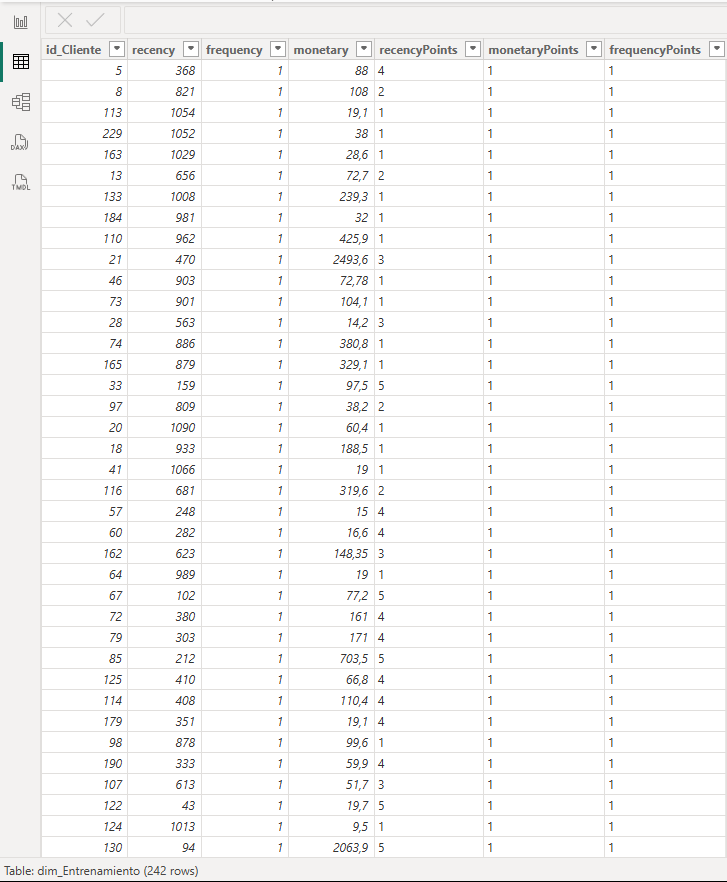
Ubicación: CD: TABLAS/Tablas de Proyecto.xlsx

## Tabla de los tipos de clientes basados en los puntajes RFM



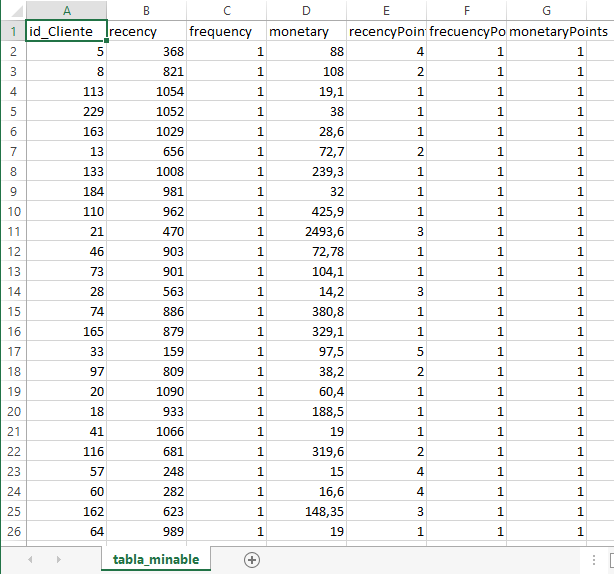
Ubicación: CD: TABLAS/Tablas de Proyecto.xlsx

## Tabla de Entrenamiento para la aplicación de los modelos



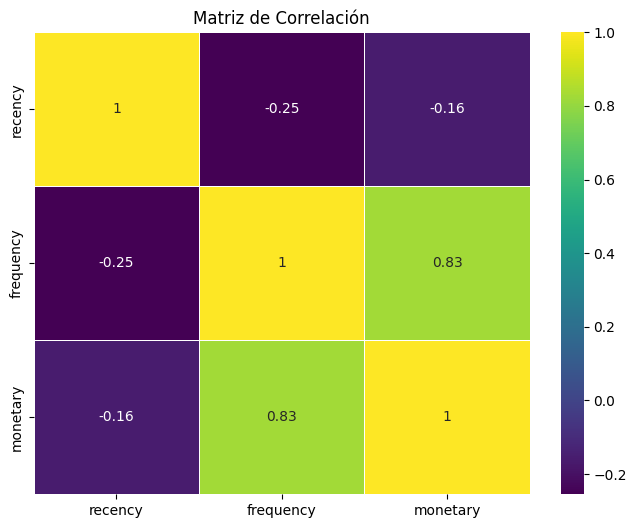
Ubicación: CD: DATA Y VISUALIZACIÓN/Proyecto Diplomado.pbix

## Tabla minable para el entrenamiento de los modelos



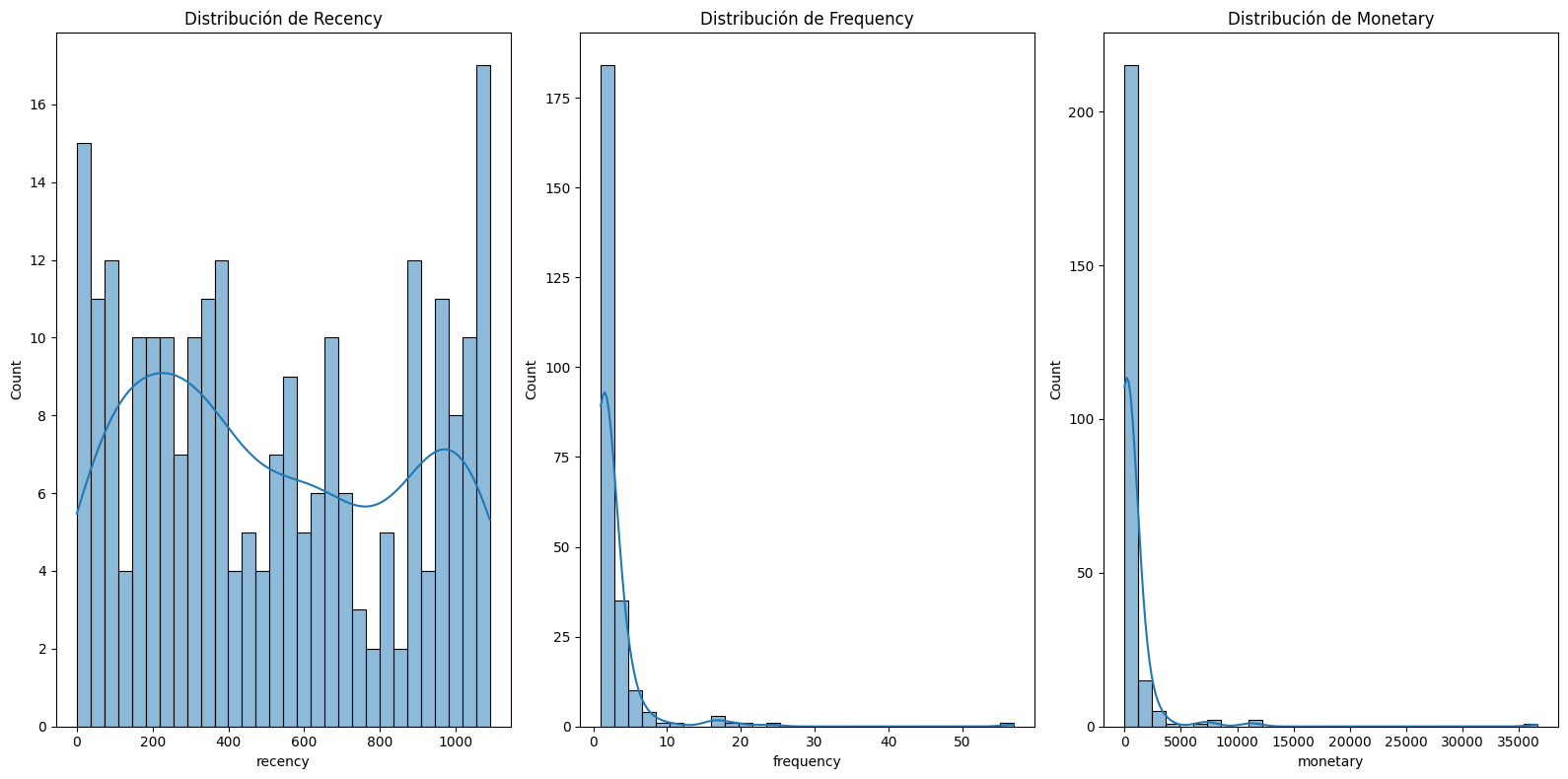
Ubicación: CD: TABLAS/tabla\_minable.csv

## Gráfica de matriz de correlación



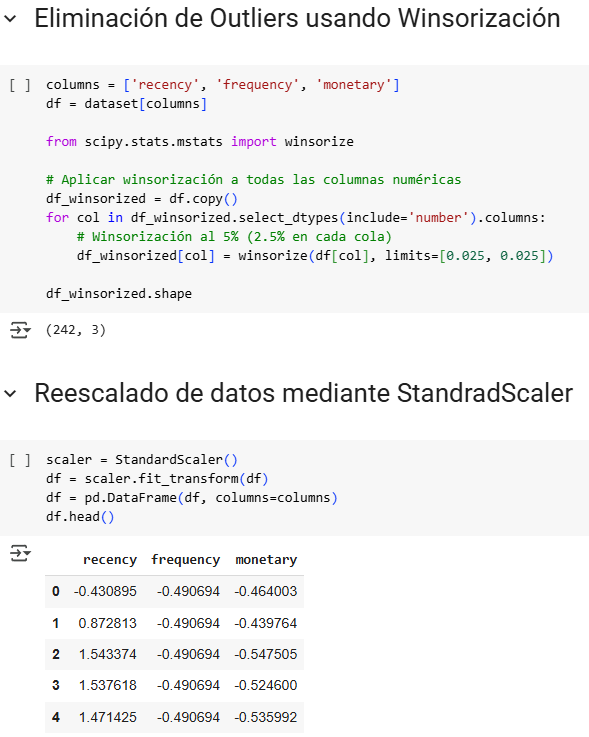
Ubicación: CD: DESARROLLO/Proyecto Winsorización.ipynb

## Gráfica de histogramas de recency, frequency y monetary



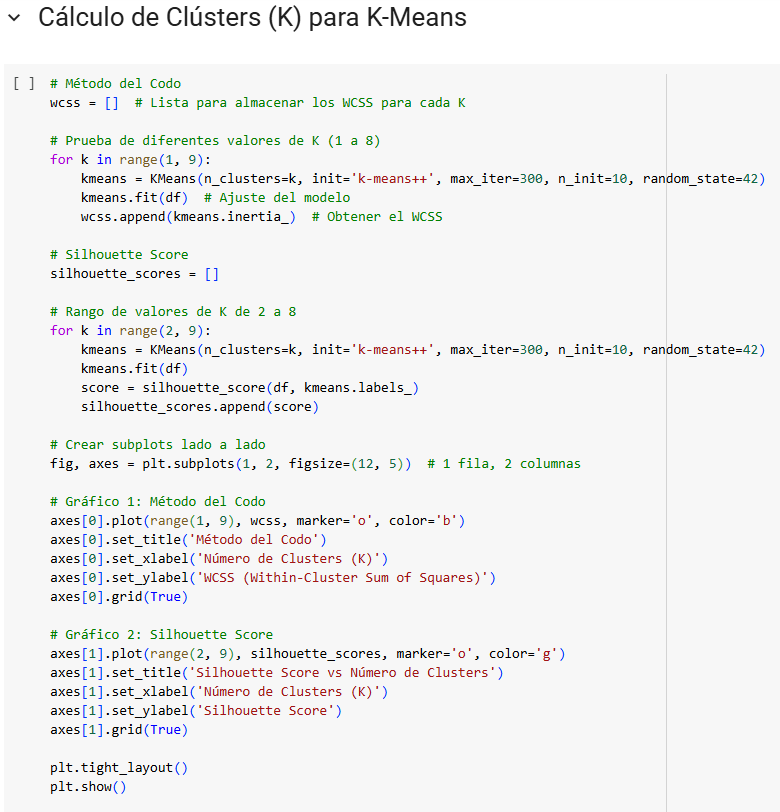
Ubicación: CD: DESARROLLO/Proyecto Winsorización.ipynb

## Código de la eliminación de outliers y re-escalado de datos



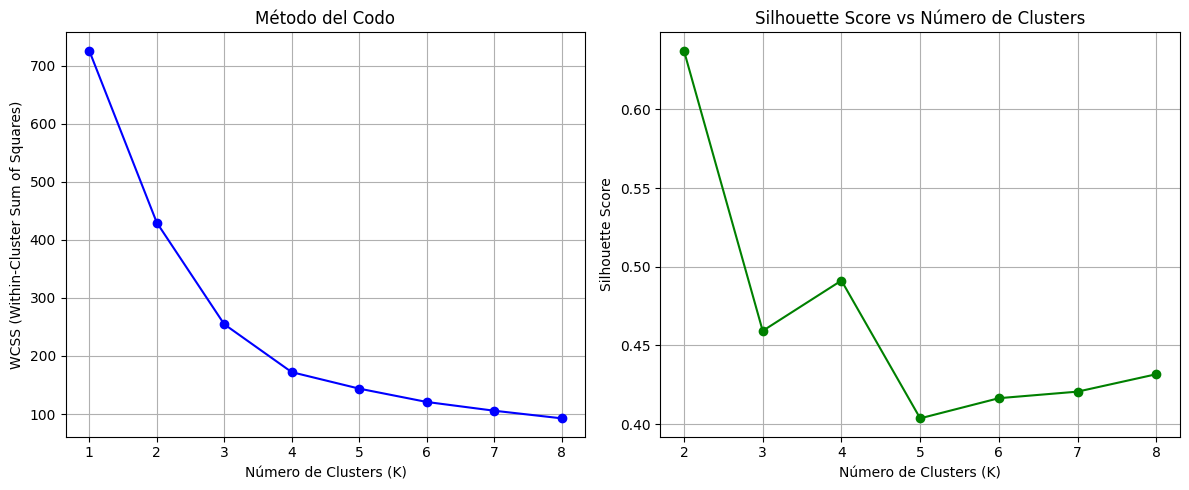
Ubicación: CD: DESARROLLO/Proyecto Winsorización.ipynb

## Código de cálculo de clústeres para K-Means



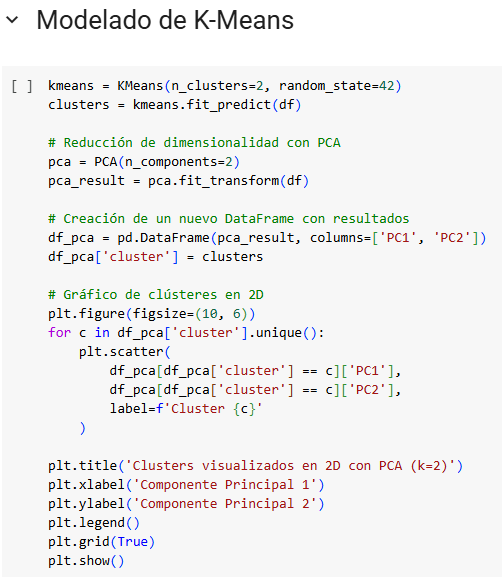
Ubicación: CD: DESARROLLO/Proyecto Winsorización.ipynb

## Valores obtenidos del Método del Codo y Silhoutte Score para K-Means



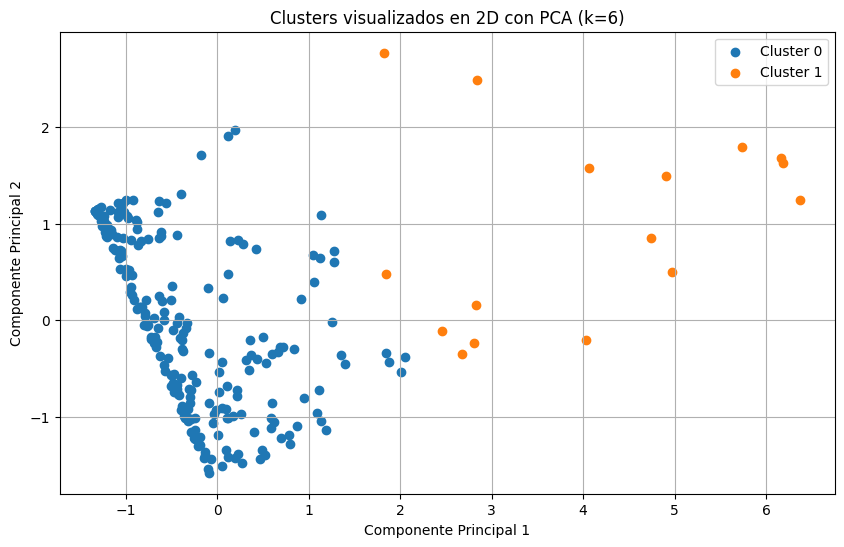
Ubicación: CD: DESARROLLO/Proyecto Winsorización.ipynb

## Modelado de K-Means



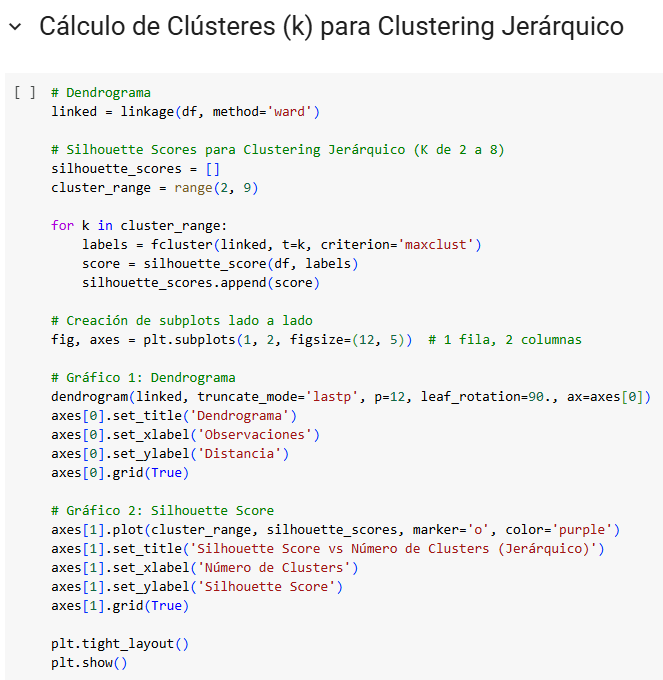
Ubicación: CD: DESARROLLO/Proyecto Winsorización.ipynb

## Resultados del modelo K-Means



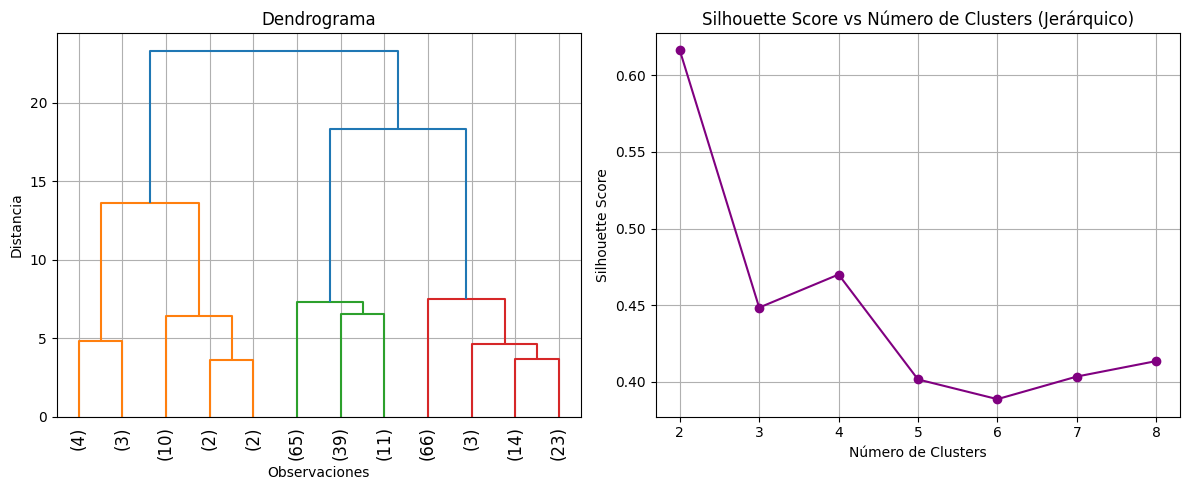
Ubicación: CD: DESARROLLO/Proyecto Winsorización.ipynb

## Código de cálculo de clústeres para Clustering Jerárquico



Ubicación: CD: DESARROLLO/Proyecto Winsorización.ipynb

## Valores obtenidos del Método del Codo y Silhoutte Score para Clustering Jerárquico



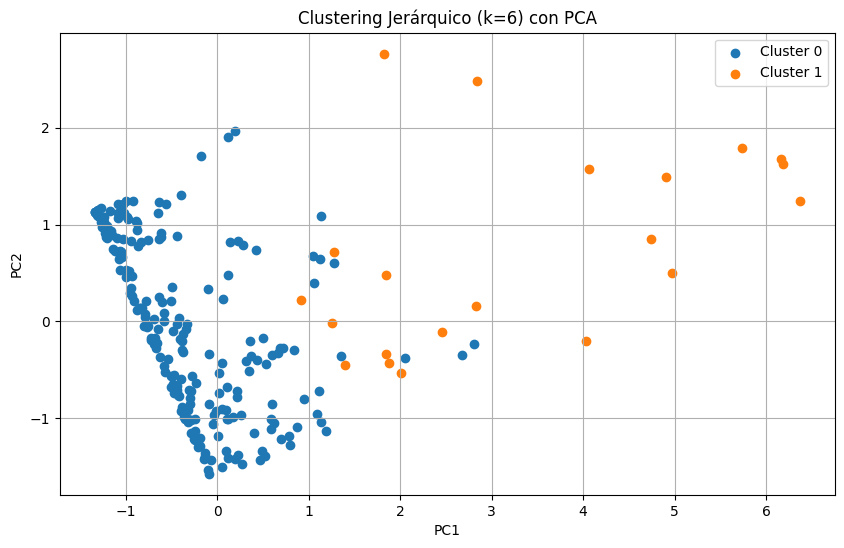
Ubicación: CD: DESARROLLO/Proyecto Winsorización.ipynb

## Modelado de Clustering Jerárquico



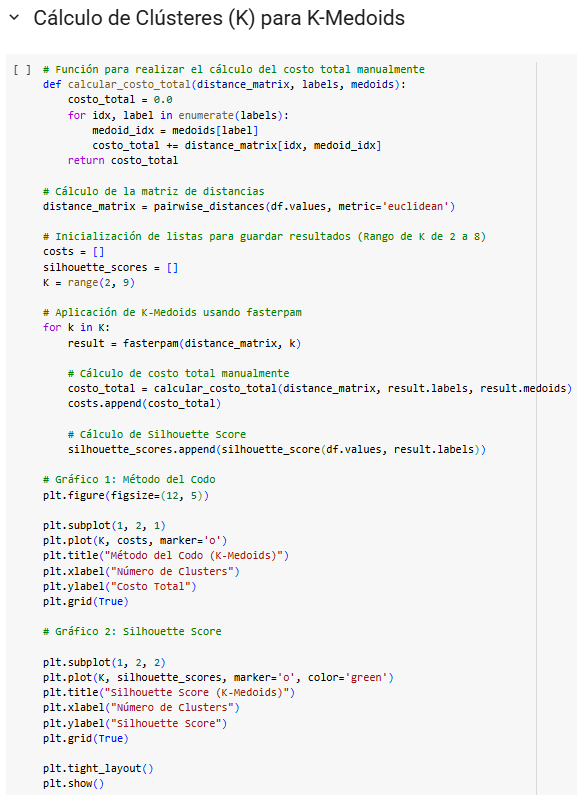
Ubicación: CD: DESARROLLO/Proyecto Winsorización.ipynb

## Resultados del modelo Clustering Jerárquico



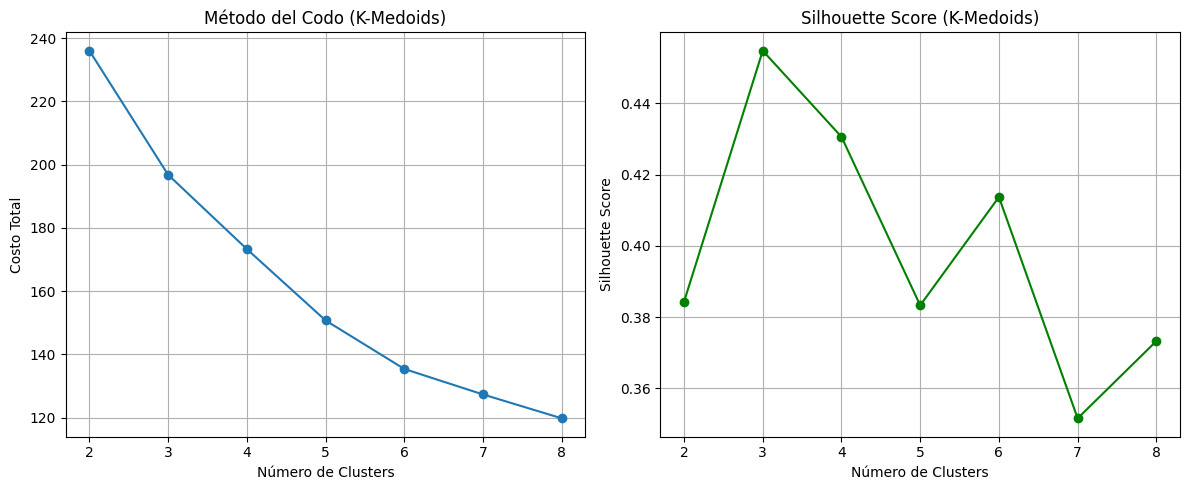
Ubicación: CD: DESARROLLO/Proyecto Winsorización.ipynb

## Código de cálculo de clústeres para K-Medoids



Ubicación: CD: DESARROLLO/Proyecto Winsorización.ipynb

## Valores obtenidos del Método del Codo y Silhoutte Score para K-Medoids



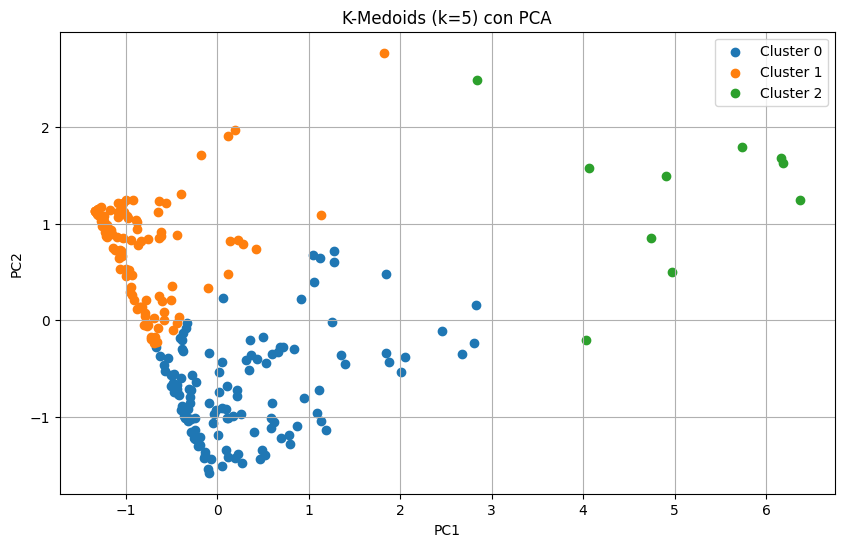
Ubicación: CD: DESARROLLO/Proyecto Winsorización.ipynb

## Modelado de Clustering K-Medoids



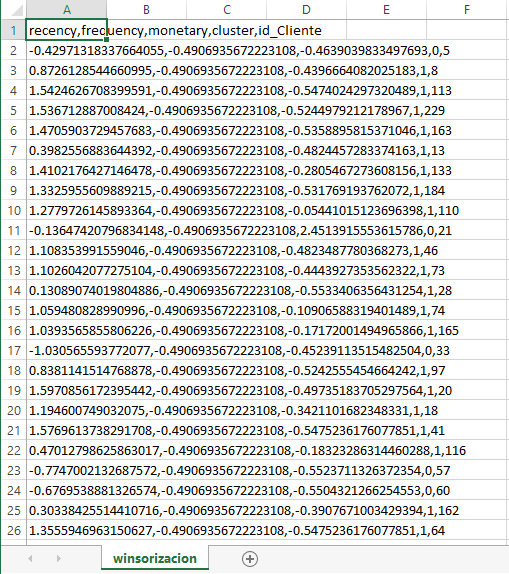
Ubicación: CD: DESARROLLO/Proyecto Winsorización.ipynb

## Resultados del modelo K-Medoids



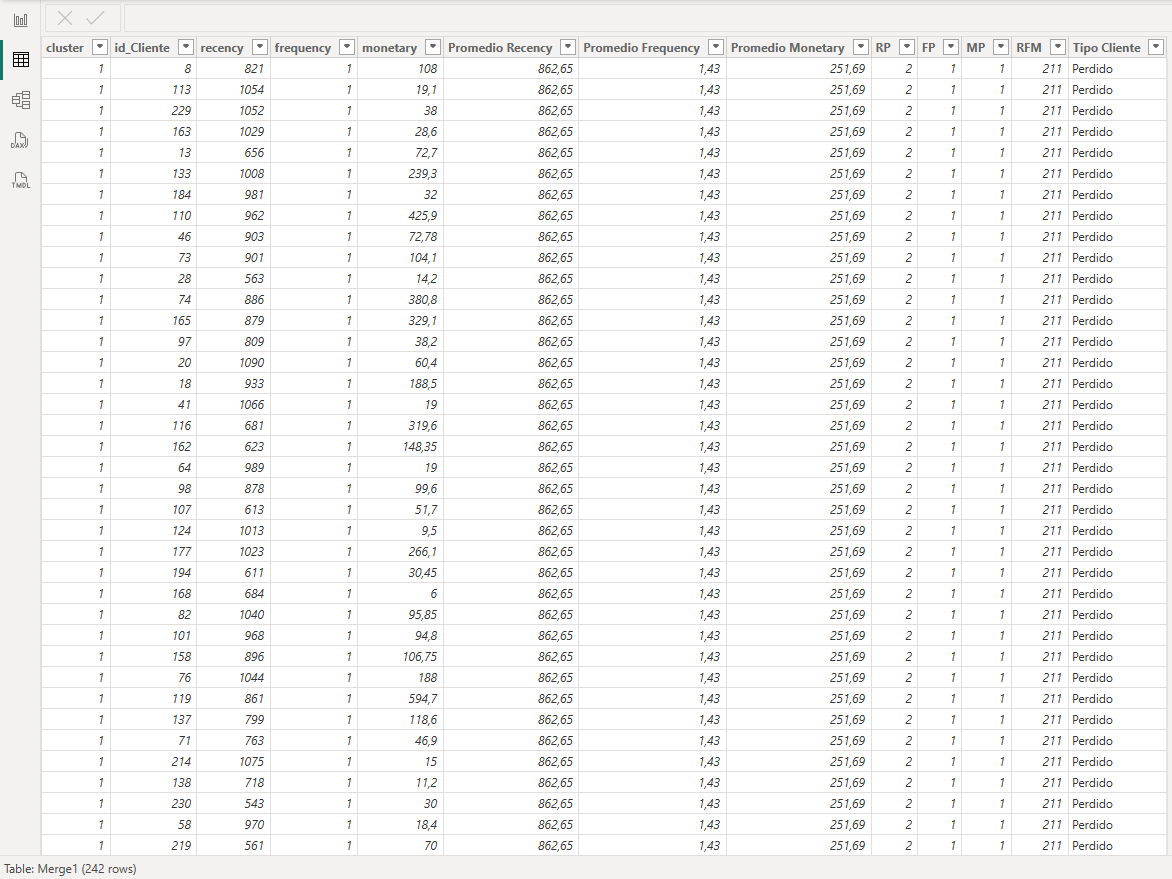
Ubicación: CD: DESARROLLO/Proyecto Winsorización.ipynb

## Tabla obtenida con los clústeres de K-Medoids



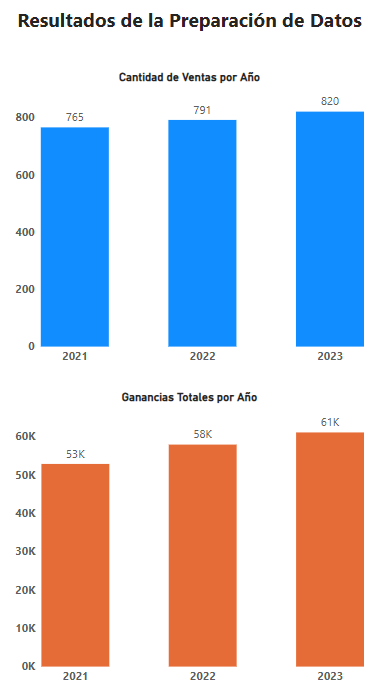
Ubicación: CD: TABLAS/winsorización.csv

## Tabla de unión con los resultados obtenidos del modelo K-medoids



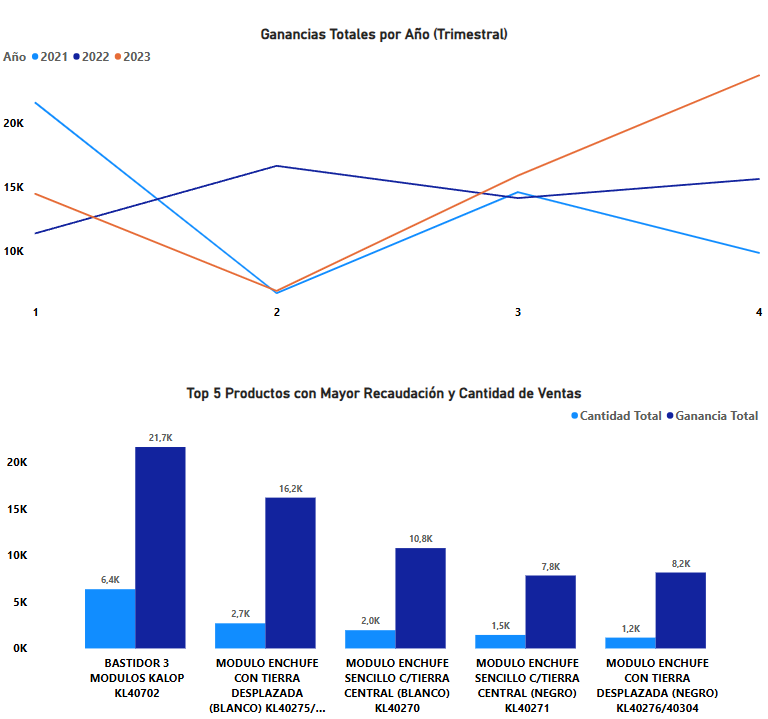
Ubicación: CD: DATA Y VISUALIZACIÓN/Proyecto Diplomado.pbix

## Gráficas obtenidas de la preparación de datos (Hoja 1)



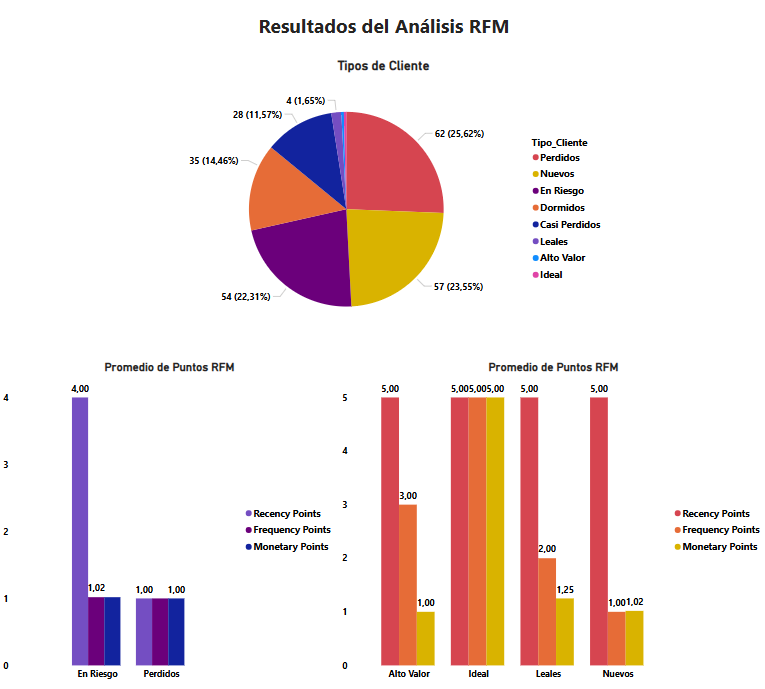
Ubicación: CD: DATA Y VISUALIZACIÓN/Proyecto Diplomado.pbix

## Gráficas obtenidas de la preparación de datos (Hoja 2)



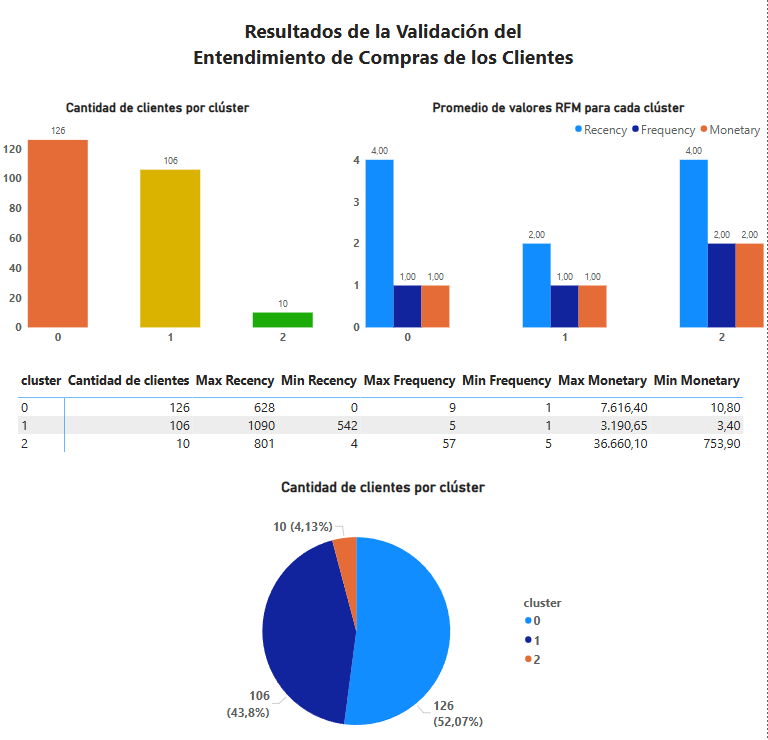
Ubicación: CD: DATA Y VISUALIZACIÓN/Proyecto Diplomado.pbix

## Gráficas obtenidas del análisis RFM



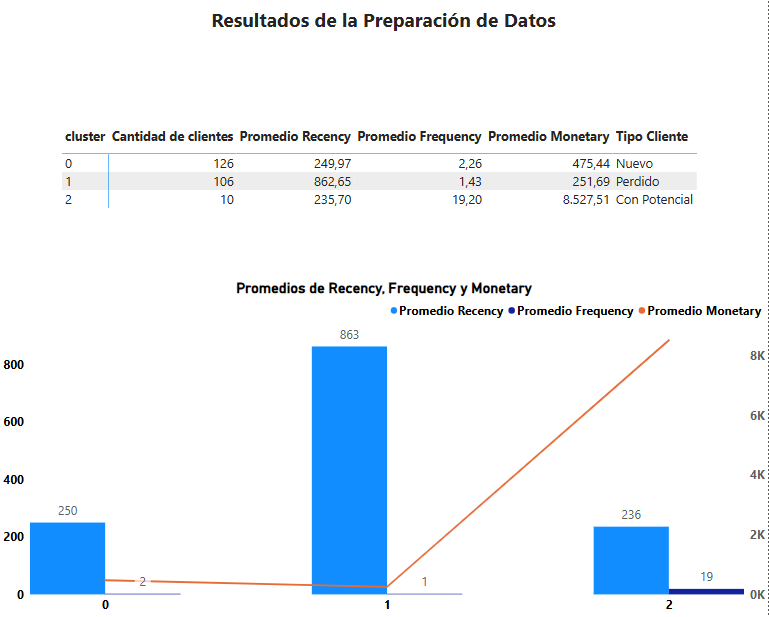
Ubicación: CD: DATA Y VISUALIZACIÓN/Proyecto Diplomado.pbix

## Gráficas obtenidas de la validación del entendimiento de compra



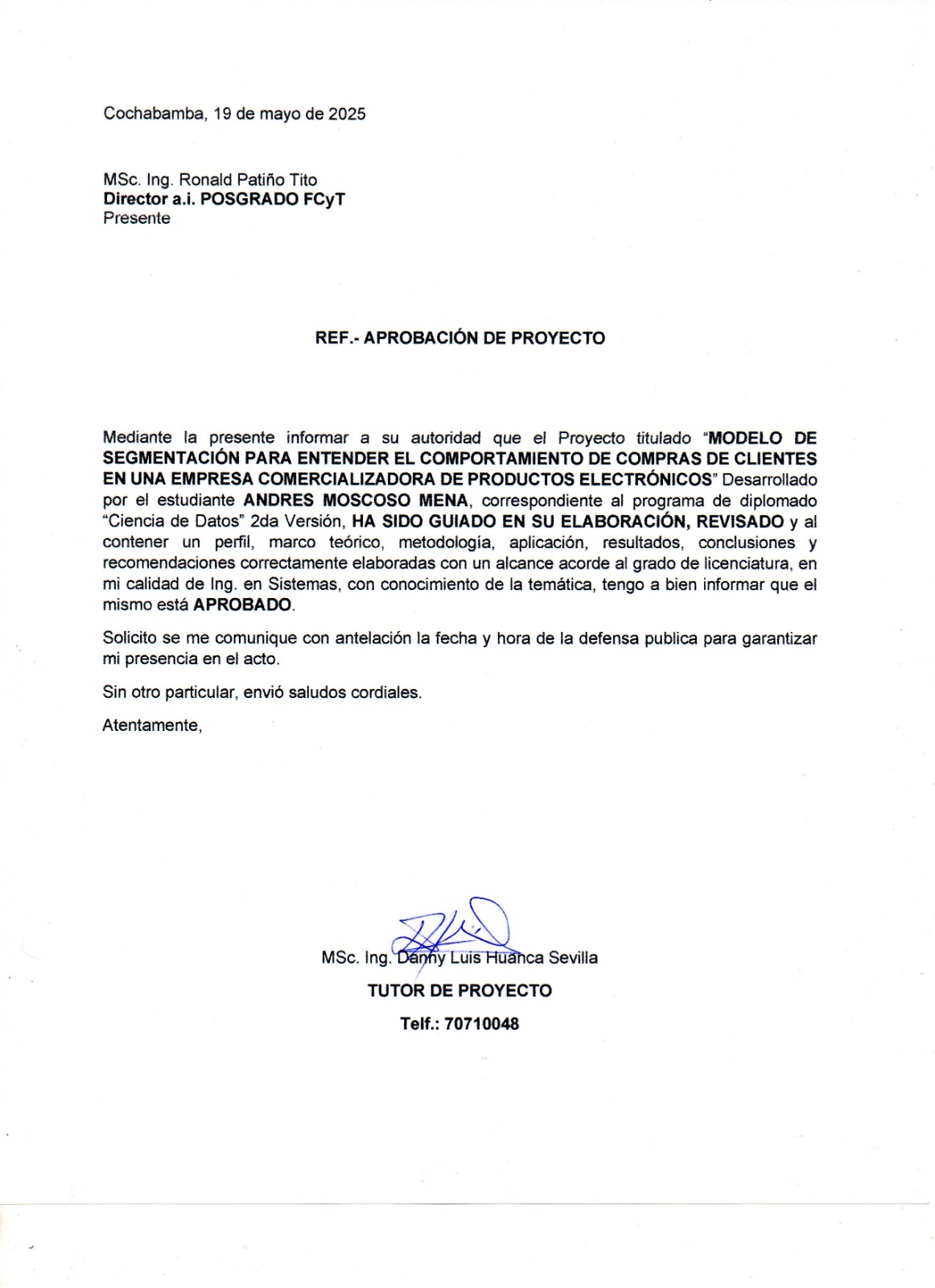
Ubicación: CD: DATA Y VISUALIZACIÓN/Proyecto Diplomado.pbix

## Gráficas obtenidas de la discusión de resultados



Ubicación: CD: DATA Y VISUALIZACIÓN/Proyecto Diplomado.pbix

## Carta de aprobación del tutor Carta de aprobación del tutor



Ubicación: CD: DOCUMENTACIÓN/Carta de Aprobación Tutor.pdf

## Código QR de enlace al repositorio del proyecto



Ubicación: CD: DOCUMENTACIÓN/Informe\_Proyecto\_AMM.docx

# Anexo PRINCIPAL: CD

El anexo CD está conformado por:

* DATA Y VISUALIZACIÓN: Carpeta donde se almacenan los archivos de los datos provistos por la empresa y el archivo de PowerBI, donde se realizó la limpieza de datos, obtención de tabla resultado y visualización de las métricas del proyecto.
* DESARRROLLO: Carpeta en la que se encuentra el archivo de Notebook donde se halla el código de la ejecución de los tres modelos usados y sus resultados.
* TABLAS: Carpeta que contiene los archivos Excel, con las diferentes tablas usadas para la clasificación de los datos, los cálculos del puntaje RFM, así como la tabla minable y la de resultados del modelo K-Medoids.
* DOCUMENTACIÓN: Carpeta que alberga el informe del proyecto realizado, y las cartas de aprobación del tutor y tribunal.
* README: Este archivo .txt contará con instrucciones para poder acceder a todos los archivos dentro de las carpetas mencionadas.

