Segmentación de Clientes Mediante Análisis de Patrones de Compra para la optimización de Estrategias Comerciales.

Jose Antonio Berrio Lasprilla.
Orlando José Olea Gómez.

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos.

Asesor
Daniel Escobar Saltarén, Magister (MsC) en Ingeniería

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2024

Cita	Berrio Lasprilla,J y Olea Gómez [1]
Referencia **Estilo IEEE (2020)**	Berrio Lasprilla,J., & Olea Gómez., O. J. (2024). Segmentación de Clientes Mediante Análisis de Patrones de Compra para la optimización de Estrategias Comerciales.]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte VII.

Grupo de Investigación Intelligent Information Systems Lab – In2Lab.

Centro de Investigación Ambientales y de Ingeniería (CIA).

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co
Rector: John Jairo Arboleda Céspedes.
Decano: Julio Cesar Saldarriaga Molina
Jefe departamento: Diego José Luis Botia Valderrama

Segmentación de Clientes Mediante Análisis de Patrones de Compra para la Optimización de Estrategias Comerciales

***Resumen* — Este trabajo aborda la necesidad de una empresa comercializadora de optimizar el tiempo empleado por su equipo de ventas en la creación de estrategias basadas en información. El objetivo principal es desarrollar un algoritmo que identifique patrones de compra en su base de clientes y los agrupe en segmentos específicos. Para lograrlo, se aplicó Kernel PCA (Análisis de Componentes Principales) para reducir la dimensionalidad de los datos y K-Means para la formación de clústeres.**

El mejor modelo obtenido fue el K-Means con kernel RBF, alcanzando un índice de silueta de 0.63, lo que permitió capturar múltiples grupos de clientes de manera efectiva. Los resultados clasificaron a los clientes en tres segmentos principales, proporcionando un recurso valioso para el diseño de iniciativas comerciales personalizadas según las características de cada grupo.

Este estudio destaca la relevancia de la ciencia de datos en pequeñas y medianas empresas, promoviendo el uso de tecnologías avanzadas en el desarrollo de estrategias que incrementen su competitividad en el mercado.

***Índice de términos* — Analítica de Datos, Clustering, Consumo Masivo, Estrategias Comerciales, Datos, K-Means, Logística, Optimización, Patrones de Compra, Segmentación, Ventas.**

I. INTRODUCCIÓN

EN la actualidad, muchas compañías en el mundo utilizan el outsourcing (o tercerización) para el ahorro dentro de sus actividades, esto quiere decir que empresas externas prestan sus servicios en nombre de ellas en los lugares que se encuentren o incluso de forma remota, ejemplo de esto son todas aquellas compañías que encargan la analítica de sus datos [1] a empresas que adquieren el compromiso de poseer los activos físicos, tecnológicos y humanos necesarios, ahorrándole a la empresa contratante la obligación de adquirir dichos elementos, permitiendo destinar este capital económico a la expansión o crecimiento del modelo de negocio.

Propiciadas estas condiciones, la analítica de datos surge como uno de los principales factores de competitividad empresarial [2]. En las empresas comercial-logísticas el uso de este activo se ve reflejado en el incremento de la eficiencia de los recursos utilizados y el ahorro de costos [3]. Esto lleva a la disminución de tiempos laborales, una mejor administración del recurso humano y optimizar la toma de decisiones, incidiendo en múltiples procesos y trayendo consigo el incremento de las utilidades [4].

Este tipo de compañías también han visto una disminución directa en los costos y gastos operativos a través de la recopilación, análisis e interpretación de la información. Proporcionando a directivos, administradores y personas que deben diseñar y ejecutar los planes de acción, una mejor comprensión del negocio o entorno económico, permitiéndoles tomar decisiones con base en mejor información [5]. A partir de esto se generan correlaciones directas que pueden verse reflejadas en el incremento de la rentabilidad de las compañías.

Para esto, se han utilizado modelos de inteligencia artificial que encuentren patrones y similitudes dentro de la maestra de clientes de tiendas en la industria minorista [1], para luego analizar matemática y estadísticamente estos grupos y generar *insights* que permitan al área comercial de la empresa crear planes de mercadeo ajustados a cada clúster. Cabe resaltar que esto contribuirá al aumento de la efectividad de las decisiones ejecutadas, pues, el empleo de la estadística en la toma de decisiones gerenciales vinculadas al marketing puede ser de bastante utilidad para disminuir la probabilidad de cometer errores [6] como campañas de mercadeo mal orientadas o inversión en proyectos poco efectivos; como también ayudará a la empresa en la generación de valor y el desarrollo de su propuesta comercial, al tiempo que contribuye a la sostenibilidad de la compañía, ya que poseer información debidamente trabajada constituye una fuente de ventaja competitiva de difícil imitación por parte de los competidores [6].

Mercantil Zafiro es una empresa comercial y logística que presta sus servicios al conglomerado transnacional del sector de consumo masivo de alimentos Grupo Nutresa, la cual hoy en día es la empresa de alimentos más grande de Colombia con presencia en países como Chile, Costa Rica, Guatemala, México, Panamá, Estados Unidos, Venezuela,

Ecuador, El Salvador, Nicaragua, Perú, República Dominicana, Malasia, las Filipinas y Sudáfrica. Dentro de sus funciones se encuentra la comercialización y distribución de una parte del portafolio de Grupo Nutresa en el departamento de Córdoba y municipios aledaños pertenecientes a los departamentos de Sucre y Antioquia; Estos productos se pueden clasificar dentro de las principales categorías: Café, Galletas, Chocolates, Pastas, Cárnicos y Snacks.

Este estudio busca aportar al desarrollo de estrategias comerciales de la empresa Mercantil Zafiro, por medio de la segmentación de sus clientes en función de las ventas realizadas a éstos durante el año 2022, de la mano de información adicional presente en los puntos de venta de cada usuario, la cual también se encuentra dentro de la base de datos.

La operación consiste en una preventa por parte del equipo comercial y una entrega a 48 horas por parte del equipo logístico. Actualmente, Mercantil Zafiro se encuentra ante el reto de incrementar la productividad de su equipo de ventas a través de un mejor uso de la información, la cual se está almacenando diariamente; sin embargo, su gestión se realiza de forma manual, consumiendo tiempo valioso de los líderes del equipo, recurso que podría ser aprovechado a través de la optimización de estos procesos [7]. Esto representa una oportunidad para la compañía, al no contar con profesionales especializados en el análisis de datos que otorguen *insights* al personal ejecutivo, que les permita tomar decisiones basadas en evidencia [8] y no asumir riesgos que pueden desembocar en actividades poco efectivas y costosas.

Considerando lo anterior, se analizarán los datos históricos comerciales del año 2022. Se aplicarán modelos de *Machine Learning* con el objetivo de identificar patrones de compra que permitan segmentar a los clientes. Este análisis proporcionará insumos valiosos para el área comercial, facilitando el desarrollo de estrategias efectivas que minimicen riesgos y reduzcan costos asociados. Además, permitirá ajustar constantemente las iniciativas para alinearse mejor con la propuesta de valor de la compañía, contribuyendo así al desarrollo sostenible de la empresa.

II. METODOLOGÍA

A. Descripción de la Base de Datos

Los datos utilizados en el estudio se obtuvieron del informe de ventas almacenado en el sistema ERP de Mercantil Zafiro (ECOM) y exportados en formato CSV. El conjunto de datos abarca información relacionada a los usuarios como mes de compra, pedidos y datos comerciales como valor de las devoluciones, presencia de activos en el punto de venta y antigüedad de los clientes. Estos registros componen la actividad comercial de la empresa durante el año 2022. Con el fin de optimizar los recursos utilizados para el manejo de la información, los datos fueron agrupados por mes y por cliente.

A continuación, en la tabla 1, se detallan una a una las variables encontradas en la base de datos original.

Variable	Descripción
Cliente	ID único para identificar a cada cliente.
Alias_asesor	ID único para identificar al asesor que realizó la venta. Se emplearon alias por razones de confidencialidad.
Alias	Hace referencia al nombre de los clientes. Se emplearon alias por razones de confidencialidad.
Negocio	Hace referencia al nombre del negocio.
Edad	Años transcurridos desde el inicio de la relación entre la empresa y el cliente hasta la fecha.
Barrio	Hace referencia al nombre del barrio donde se encuentra ubicado el negocio.
Codigo Producto	ID único para identificar al producto adquirido por el cliente.
Referencia	Nombre del producto adquirido por el cliente.
Cant. Dev. (uds)	Unidades de producto no recibidas por el cliente al momento de la entrega.
Cant. Neta (uds)	Unidades de producto recibidas por el cliente a conformidad.
Vta. - IVA ($)	Valor total de la venta en pesos colombianos (COP) antes del Impuesto al Valor Agregado (IVA).
Negocio2	Segmento de productos al cual pertenece la referencia adquirida por el cliente.
Ciudad	Ciudad en la que se encuentra ubicado el negocio.
Tamaño	Segmento de negocio al cual pertenece el establecimiento comercial según sus dimensiones físicas.
Potencial	Segmento de negocio al cual

	pertenece el establecimiento comercial según sus márgenes de compra y proyecciones a futuro.
Mes	Mes en el cual fueron realizadas las compras.
Exhibidor	Variable binaria que indica si el cliente cuenta, o no, con un activo comercial en su establecimiento.

Tabla 1. Descripción de variables en BD original

B. Generación de Base de Datos por Cliente

Para la fase de preparación de los datos, se trabajó con una base compuesta por 869,926 filas y 17 columnas, previamente descritas. Como primer paso, se diseñó un diccionario de nombres únicos con el propósito de anonimizar los datos. Luego, se realizó una agrupación de los registros para cada cliente, calculando el número de referencias y sumando el valor de sus devoluciones, compras y valor neto de la mercancía adquirida por cada mes, generando una versión transpuesta de la tabla original. Diagrama

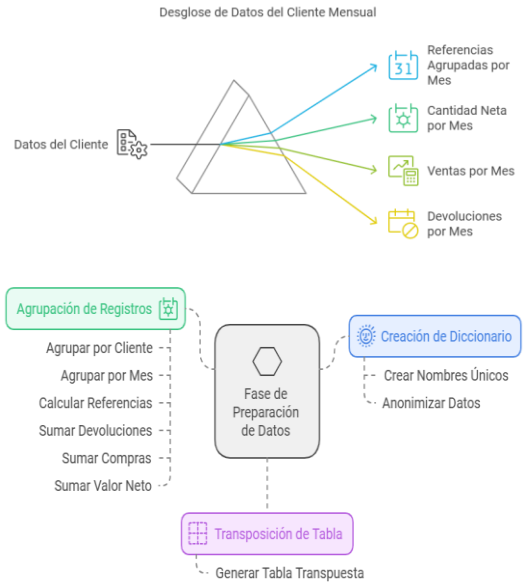

Figura 1. Diagrama de generación de bases

Este proceso consistió en operar las variables numéricas de manera mensual para cada cliente, dando como resultado una estructura más consolidada.

C. Limpieza de la Base de Datos

Se validaron los datos con el objetivo de evitar errores dentro de la información, limpiando aquellos que dificultaran el estudio, se eliminaron valores nulos y registros duplicados para mejorar la calidad y fiabilidad de la base.

Finalmente, se procedió a la eliminación de *outliers*, identificándolos a través de una caja de bigotes, donde se descartaron los que se encontraban por fuera del rango intercuartílico definido ($0.05<X<0.95$) y permitiendo un estudio con mayor claridad en la información.

Una vez completados los pasos anteriores, se procedió a estandarizar los datos utilizando la función StandardScaler, esta función se caracteriza por estandarizar las características del conjunto de datos, transformándolas para que tengan una media de 0 y una desviación estándar de 1.

D. Análisis Estadístico Exploratorio

Con la base de datos preparada, se realizó un análisis estadístico exploratorio *(EDA)* para comprender la estructura y el comportamiento de las principales variables del estudio. Se generó un resumen estadístico de las variables clave como “Cant. Neta”, “Cant. Dev”, “Vta. - IVA” y “Referencia”, utilizando métricas descriptivas como la media, mediana, desviación estándar, valores máximos y mínimos. Además, se implementaron diagramas de frecuencia para analizar la distribución mensual de las variables principales. Este enfoque facilitó la identificación de patrones específicos en el comportamiento de las ventas netas, devoluciones y ventas antes de IVA, así como de las referencias más recurrentes en diferentes periodos.

Con el objetivo de analizar la afinidad y las relaciones entre las variables, se construyó una matriz de correlación. Este análisis permitió identificar relaciones significativas que podrían influir en el modelo de segmentación.

E. Reducción Dimensional

Tras el análisis inicial, se empleó el método de Análisis de Componentes Principales *(PCA)* con el objetivo de reducir la dimensionalidad de las variables. Este procedimiento permitió transformar el conjunto de datos original en un espacio de menor dimensión, conservando la mayor parte de la variabilidad explicada, lo cual optimiza el rendimiento de los algoritmos subsecuentes.

F. Clustering

Como parte del proceso de segmentación, se implementó el método del codo para determinar el número óptimo de clústeres que mejor representara las características de los datos. Este enfoque permitió identificar el punto en el que la suma de las distancias intra - clúster deja de reducirse significativamente al incrementar el número de conjuntos, optimizando así el equilibrio entre cohesión y separación de los grupos.

Posteriormente, se aplicó el algoritmo *K-Means* para llevar a cabo la segmentación de clientes, agrupando los puntos de datos en clústeres basados en similitudes intrínsecas. Este enfoque permitió identificar grupos de clientes con características homogéneas, proporcionando una base sólida para diseñar estrategias comerciales ajustadas a cada uno de ellos, mejorando la efectividad de las decisiones tomadas.

Por último, se implementó Kernel PCA, el cual permite manejar datos no lineales por medio del uso de funciones de *kernel*, que mapean los datos a un espacio de mayor dimensión, donde los patrones no lineales pueden ser capturados, facilitando la extracción de los componentes principales en este nuevo espacio transformado.

G. Métricas de Validación

Para evaluar la calidad de los clústeres generados en el análisis de segmentación, se utilizaron tres métricas de validación: el Índice de Silueta, el Índice de Davies-Bouldin y el Índice de Calinski-Harabasz. Estas medidas permitieron analizar la cohesión interna de cada conjunto y su separación respecto a los otros grupos, proporcionando una base sólida para comparar los resultados obtenidos de los diferentes modelos con distintos números de clúster, permitiendo determinar cuál opción era la mejor.

III. RESULTADOS

C. Limpieza de la Base de Datos

Durante el análisis exploratorio, mientras se revisaba la Figura 2, se encontró que en la base de datos existían valores extremos, registros anómalos o posibles errores que, al encontrarse significativamente alejados del rango intercuartílico (IQR) definido, podrían afectar desproporcionadamente las medidas estadísticas como la media y la desviación estándar, además de los diferentes análisis de negocio que se desarrollarían a lo largo del estudio.

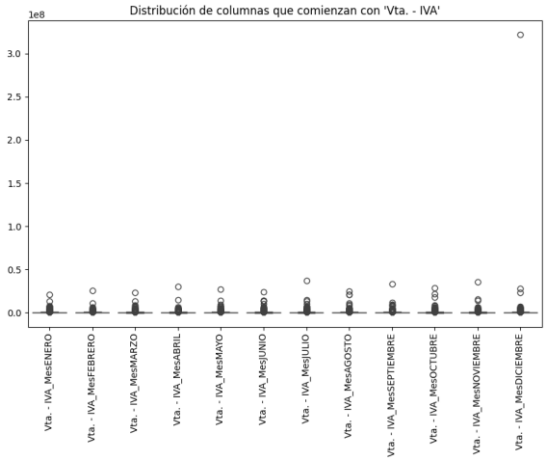

Figura 2. *Diagrama de cajas y bigotes ventas 2022*

Dada esta situación, se procedió a la eliminación de los *outliers* existentes, conservando únicamente los valores que se encontraban dentro del rango intercuartílico establecido, lo cual favoreció la claridad en estructura interna de las distribuciones de cada mes, como puede verse en la Figura 3. Esbozando un rango de valores más compacto y facilitando la comparación entre meses y la visualización de los datos.

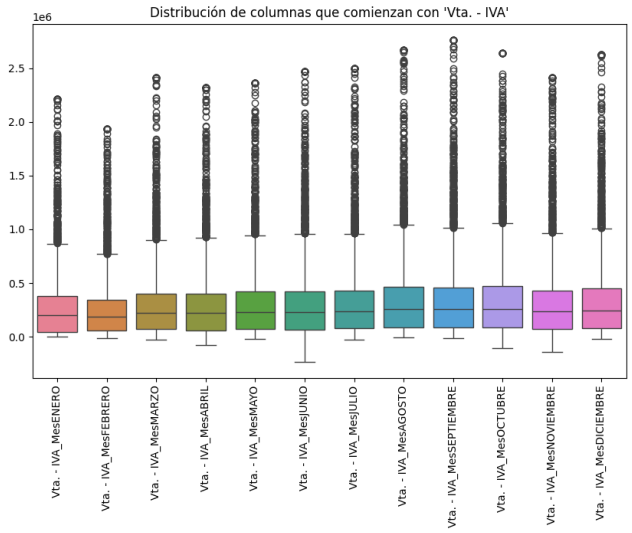

Figura 3. Diagrama de cajas y bigotes ventas 2022 sin outliers

La Figura 3 muestra distribuciones mensuales similares con medianas y rangos intercuartílicos consistentes, lo cual indica que las variaciones principales entre los meses son más homogéneas y se representa mejor el comportamiento general del conjunto de datos.

D. Análisis Estadístico Exploratorio

Luego de llevar a cabo los primeros pasos del estudio, se logró condensar la base de datos inicial a una de menor dimensión, con mayor facilidad de manejo y sin errores sustanciales que afectaran el estudio, donde la información quedó agrupada por cada uno de los clientes y de forma mensual. En la Figura 4 puede verse la manera como evolucionó la variable “Cant. Dev” para cada uno de los meses, donde se evidencia que los meses de enero y febrero son los que presentan la menor devolución de mercancía, mientras que en el resto de año existe un incremento considerable representando variaciones importantes.

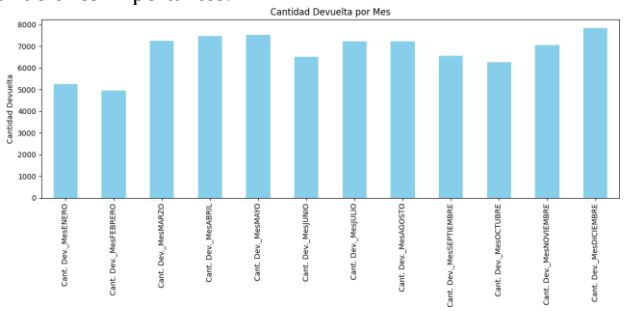

Figura 4. Evolución de las cantidades devueltas por mes

La variable “Vta. - IVA” representa las compras realizadas por los clientes a lo largo del año, en la Figura 5 se evidencia una leve tendencia alcista a lo largo del año, la cual se acentúa en los meses de agosto, septiembre y octubre, desciende en el mes de noviembre y logra un leve repunte en el último mes del año.

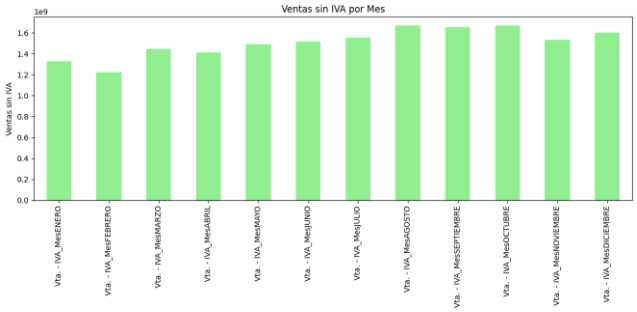

Figura 5. Ventas mes a mes 2022

Como aspecto positivo para la empresa, se destaca que en los meses de mayores ventas, la devolución de mercancía por parte de los clientes presentó un descenso considerable frente a los otros meses del año.

Con el ánimo de profundizar en la información relacionada a los clientes, se realizó un diagrama de frecuencias (Figura 6) donde se puede ver su distribución por años de relación comercial con la compañía, evidenciando que el grueso de compradores son negocios que llevan de 3 a 5 años de trabajo con la empresa, el segundo grupo con mayor número de clientes son los que llevan con la compañía de 9 a 11 años, para un estudio futuro sería valioso revisar la evolución de estos grupos en el tiempo y cómo la empresa se adapta a dichos cambios.

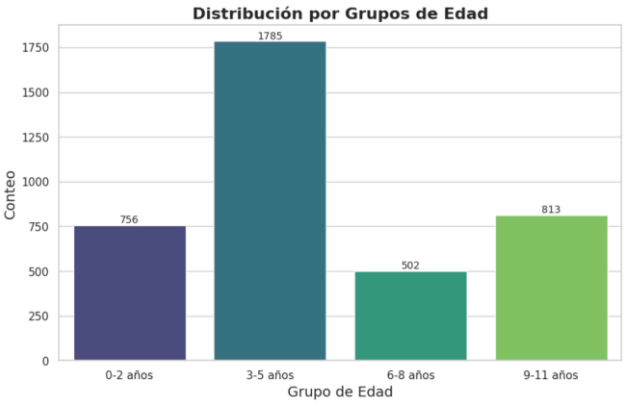

Figura 6. Diagrama de frecuencias edad de los clientes

En línea con el objetivo principal del estudio, y luego de realizar el análisis exploratorio y la estandarización de los datos, se procedió a determinar cuál sería el número óptimo de clústeres teniendo en cuenta las características de los datos (Figura 7), para esto se aplicó el Método del Codo, donde se evidencia que la inercia reduce su variación drásticamente a partir del punto número 3, por lo cual, se determina que los modelos de segmentación se evaluarán con este número de clústeres.

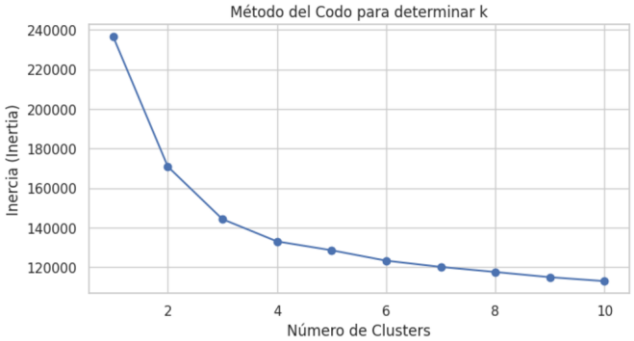

Figura 7. Método del Codo

Con el número de *clústeres* óptimo definido, procedemos al primer intento de segmentación de los clientes utilizando el Análisis de Componentes Principales (Figura 8).

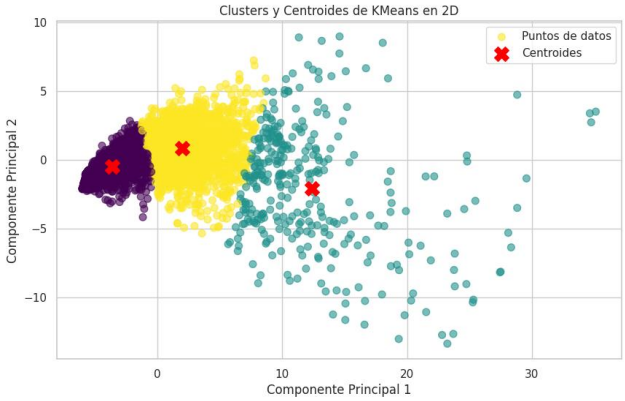

Figura 8. Visualización de Clústeres con PCA

Sin embargo, este método por sí solo no fue capaz de generar resultados trascendentales, haciendo necesario el uso de Kernel PCA, la cual permitió una mejor separación del conjunto de datos gracias a su robustez frente a la no linealidad (Figura 9).

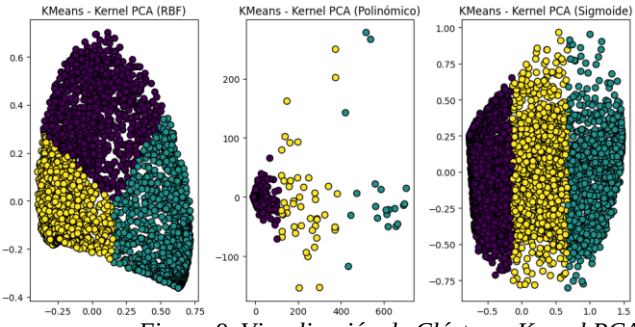

Figura 9. Visualización de Clústeres Kernel PCA

G. Métricas de Validación

Para evaluar la calidad y relevancia de los grupos formados se emplearon métricas reconocidas en el ámbito del aprendizaje no supervisado, como el índice de silueta, el índice de Davies-Bouldin y el índice de Calinski-Harabasz. Estas métricas proporcionan una visión cuantitativa de la cohesión dentro de cada clúster y la separación entre clústeres,

permitiendo identificar posibles mejoras o ajustes en el modelo.

Métrica/Modelo	K-Means
I. de Silhouette	0.2575
Davies Bouldin	1.3858
Calinski-Harabasz	1766.85

Tabla 2. Evaluación de modelos PCA

Estas métricas indican que el modelo K-Means con PCA (Tabla 2) logra una agrupación moderada, con un índice de la Silueta bajo (0.2575), lo que sugiere solapamiento entre clústeres, y un índice Davies-Bouldin (1.3858) que refleja una calidad aceptable pero mejorable y, por otro lado, el índice Calinski-Harabasz es alto (1766.85), indicando que el modelo captura cierta estructura en los datos. Estos resultados confirman la necesidad de llevar a cabo otra alternativa que ofrezca mejores indicadores.

Método	Kernel	Índice de Silueta	Índice de Davies-Bouldin	Índice de Calinski-Harabasz	Clúster 0	Clúster 1	Clúster 2
KMeans	RBF	0.6308	0.6312	8859.7827	2879	862	989
	Polinómico	0.9727	0.4492	8769.4756	4689	4	37
	Sigmoide	0.4806	0.8065	8880.3084	2228	1225	1277
Agglomerativo	RBF	0.6087	0.6858	7305.1620	916	3000	814
	Polinómico	0.9735	0.3701	8767.7751	4691	36	3
	Sigmoide	0.4555	0.8030	7853.3454	2170	1559	1001

Tabla 3. Evaluación de modelos con Kernel PCA

Con la aplicación de Kernel PCA, se observa que la mejor opción en ambos métodos (K-Means vs Aglomerativo) la ofrece el kernel polinómico, ya que tienen Índices de Silhouette altos (0.9727 en K-Means y 0.9735 en Aglomerativo), lo que indica clústeres bien definidos, y los valores más bajos de Davies-Bouldin (0.4492 y 0.3701, respectivamente), reflejando una buena separación entre clústeres. El índice Calinski-Harabasz también es alto en estos casos, confirmando una buena dispersión entre los clústeres. Por otro lado, los kernels RBF y sigmoide muestran buen desempeño, con índices de Silhouette y Davies-Bouldin altos.

El paso siguiente es entrar a analizar la distribución de los *clusters* y ver qué información puede ser relevante para las conclusiones del estudio.

Cluster	KMeans_RBF	KMeans_poly	KMeans_sigmoid
0	2879	4689	2228
1	862	4	1225
2	989	37	1277

Tabla 4. Número de Clientes por Clúster usando K-Means

Cluster	Agglo_RBF	Agglo_poly	Agglo_sigmoid
0	916	4691	2170
1	3000	36	1559
2	814	3	1001

Tabla 5. Número de Clientes por Clustering Aglomerativo

La distribución de los clústeres muestra que el kernel polinómico concentra la mayoría de los clientes en un solo grupo, lo cual va en contravía de lo que se había planteado anteriormente, esto puede indicar *overfitting* o una estructura dominante. El kernel RBF logra una distribución más equilibrada entre conjuntos, ideal para identificar múltiples patrones lo cual es el objetivo principal del estudio, por otro lado, que el kernel sigmoide ofrece una distribución moderadamente equilibrada, aunque con menor calidad de agrupamiento según las métricas lo cual puede ser contraproducente en el futuro.

F. Perfilamiento de Clúster

Para el perfilamiento de los Clusters tenemos nuestros datos normalizados entre 0 y 1

Clúster 0

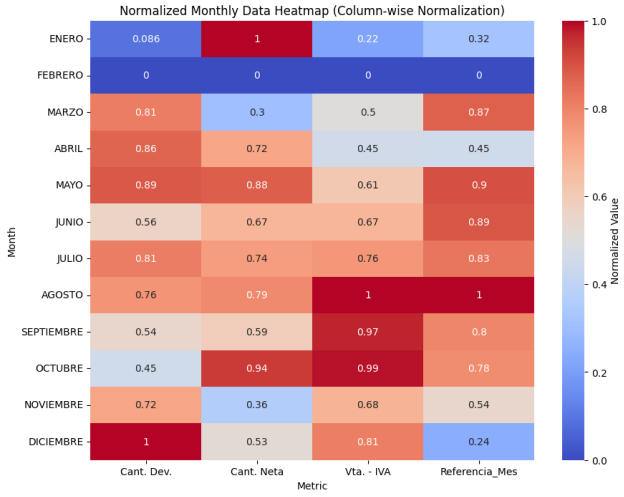

Figura 10. Matriz de Correlación Clúster 0

El Clúster 0 muestra un comportamiento **muy variable a lo largo del año**. En el **primer trimestre (Q1)**, enero destaca por un máximo en **Cant. Neta** (1), pero febrero tiene valores extremadamente bajos (0 en todas las métricas). Marzo se recupera parcialmente, con un valor alto en **Cant. Dev.** (0.81). En el **segundo trimestre (Q2)**, abril sobresale por valores equilibrados y altos en **Cant. Dev.** (0.86) y **Cant. Neta** (0.72), mientras que junio registra un pico en **Ref. Mes** (0.89). El **tercer trimestre (Q3)** es el periodo más fuerte para este clúster, con máximos en **Vta.-IVA** y **Ref. Mes** en agosto y septiembre (1 y 0.97, respectivamente), consolidándose como el clúster más destacado del trimestre. En el **cuarto trimestre (Q4)**, este clúster mantiene su dominancia, alcanzando máximos en **Cant. Neta** en octubre y diciembre (0.94 y 1), junto con altos valores en **Vta.-IVA**, lo que refleja su liderazgo en ventas y actividad general al cierre del año.

Cluster1

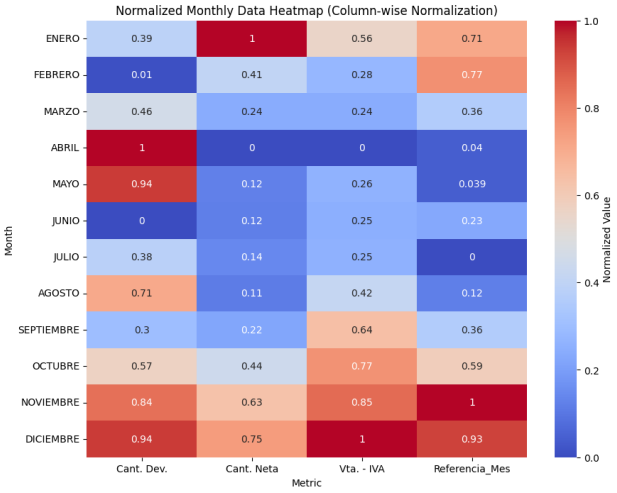

Figura 11. Matriz de Correlación Clúster 1

En el **primer trimestre (Q1)**, el Clúster 1 tiene un rendimiento bajo en casi todas las métricas, con un único punto destacado en enero, donde la **Cant. Neta** alcanza su máximo valor (1). Febrero y marzo muestran bajos niveles de actividad, reflejando un comportamiento irregular en este periodo. Durante el **segundo trimestre (Q2)**, este clúster presenta un contraste significativo: mientras que en abril **Cant. Dev.** alcanza un máximo absoluto (1), las demás métricas permanecen extremadamente bajas. Mayo y junio son meses con poca actividad en general. En el **tercer trimestre (Q3)**, el rendimiento sigue siendo limitado, con valores bajos en julio y agosto, aunque hay un ligero repunte en **Vta.-IVA** en septiembre (0.64). Finalmente, en el **cuarto trimestre (Q4)**, se observa una mejora significativa, destacando con máximos en **Ref. Mes** en noviembre y diciembre (1 y 0.93, respectivamente). Aunque las demás métricas siguen siendo moderadas, este trimestre representa el periodo de mejor desempeño para este clúster.

Cluster2

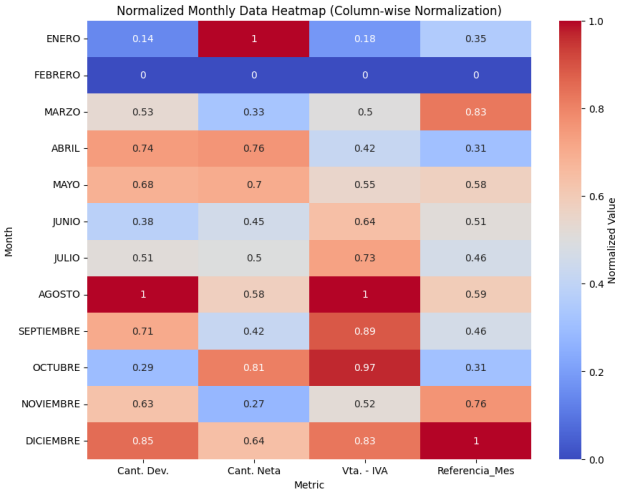

Figura 12. Matriz de Correlación Clúster 2

El Clúster 2 se caracteriza por su **comportamiento estable y balanceado** a lo largo del año. En el **primer trimestre (Q1)**, destaca en enero con un máximo en **Cant. Neta** (1), pero febrero muestra valores bajos en todas las métricas. En marzo, se recupera con altos valores en **Ref. Mes** (0.83) y niveles moderados en las demás métricas. Durante el **segundo trimestre (Q2)**, el clúster mantiene una estabilidad notable, con métricas equilibradas (~0.5) en abril, mayo y junio, aunque **Ref. Mes** es ligeramente más baja en abril (0.31). En el **tercer trimestre (Q3)**, este clúster sobresale con altos valores en **Vta.-IVA**, alcanzando 0.89 en septiembre, y métricas consistentes en **Cant. Neta** y **Ref. Mes**. Finalmente, en el **cuarto trimestre (Q4)**, este clúster tiene su mejor desempeño del año: en diciembre, todas las métricas alcanzan valores altos, con un máximo absoluto en **Ref. Mes** (1), consolidándose como el clúster más balanceado en el cierre del año.

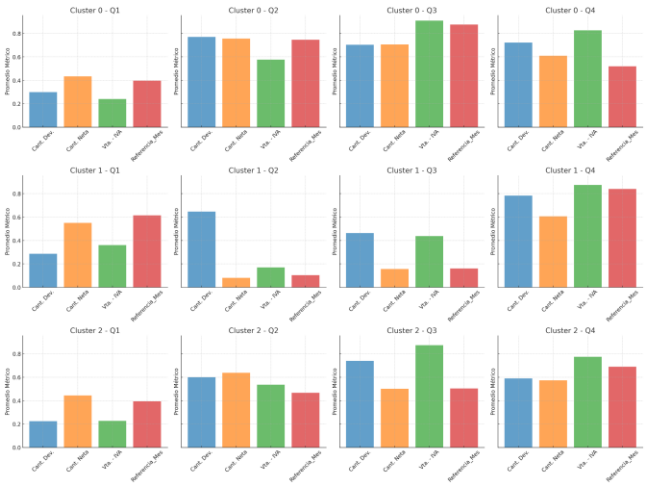

Figura 13. Diagrama de frecuencia de los cluster por trimestre

- El Clúster 0 tiene un comportamiento consistente con un aumento en las ventas netas y el volumen neto hacia el final del año.
- El Clúster 1 Los valores altos en Q1 indican un mejor rendimiento en el inicio del año.Muestra un desempeño irregular, con ventas netas y cantidades netas bajas en algunos trimestres (especialmente Q2 y Q3)
- El Clúster 2 representa clientes con un buen desempeño y estabilidad a lo largo del tiempo.

IV. CONCLUSIONES

Existe mucho valor en la información comercial de la empresa Mercantil Zafiro, es importante que con el tiempo vayan adoptando una filosofía de trabajo orientada a los datos, teniendo en cuenta que hay un importante campo de acción en la información generada dentro de su operación diaria; no conformarse solamente a nivel de ventas sino también explorar la posibilidad de llevar la analítica de datos a las demás áreas de la compañía.

Gracias a la información que se tiene de la antigüedad de los clientes se puede concluir que el 46.3% tienen entre 3 y 5 años de relación con la empresa, seguido del grupo de clientes que se encuentran dentro de los 9 a 11 años con una participación del 21.1%, se evidencia una disminución importante entre ambos grupos, es decir, los clientes que tienen una relación comercial de entre 6 a 8 años. Es importante, que la empresa evalúe las razones de este fenómeno y si puede evitar el decrecimiento de su grupo de clientes más representativos a través de estrategias que estén a su alcance.

En cuanto a las ventas, como se mencionó anteriormente, presentan un leve crecimiento a lo largo del año, los meses con las ventas más altas del año fueron agosto, septiembre y octubre.

La devolución de mercancía fluctuó más que las ventas, el mes con mayor cantidad de devolución fue diciembre, seguido de marzo, abril y mayo. Esto, dentro del contexto de negocio, se entiende debido al número de días festivos, pues muchas familias en semana santa y diciembre se van para otras ciudades o a vacacionar al campo, lo cual representa una disminución de las ventas en el segmento tienda a tienda.

Aunque inicialmente los algoritmos *K-Means* Polinómico y Aglomerativo Polinómico presentaban las mejores métricas, se evidenció en la distribución de *clústeres* que la agrupación que éstos realizaban no era la más adecuada para el estudio, por lo cual, en función de éstas distribuciones, se estableció que el algoritmo más adecuado es el *K-Means RBF*, ya que logra una distribución más equilibrada entre conjuntos, característica que jugará a favor en el análisis de los clientes y en el desarrollo de estrategias comerciales enfocadas en éstos.

La segmentación de datos mediante la identificación de clústeres ha demostrado ser una herramienta valiosa para analizar grupos con características únicas, proporcionando una comprensión más precisa de cómo diferentes clientes o productos contribuyen al rendimiento general. Este análisis reveló patrones específicos relacionados con la estacionalidad, el volumen de actividad y la estabilidad en las métricas, lo que permite diseñar estrategias más enfocadas y efectivas.

Cada clúster requiere un enfoque diferenciado:

- Clúster 0: Estrategias para maximizar ingresos en picos estacionales.
- Clúster 1: Reactivar el desempeño en periodos bajos.
- Clúster 2: Estabilidad y potenciar resultados.

En resumen, el análisis por clústeres no sólo aporta un recurso importante para la toma de decisiones estratégicas, sino que también maximiza los resultados al permitir acciones adaptadas a las particularidades de cada segmento identificado.

V. REFERENCIAS

[1] K. Agarwal, P. Jain y M. Rajnayak, “Comparative Analysis of Store Clustering Techniques in the Retail Industry”, in DATA 2019 - Proceedings of the 8th International Conference on Data Science, Technology and Applications, 2019, pp. 65-66.

[2] H. M. Figueroa, “Mejores prácticas en analítica de datos para la optimización de procesos de comercialización en empresas cárnicas ubicadas en el departamento de Córdoba”, Tesis de grado, Dep. de Adm. de Emp., CESA, 2023.

[3] O.S. Pardo y D.M. Navarro, “Analítica de datos para toma de decisiones en las pymes y los micro establecimientos del sector turístico de Colombia 2015–2019”, Fundación Universitaria Compensar, Villavicencio, Colombia, Informe Final de Investigación, 2020.

[4] R. Treviño-Reyes, F. S. Rivera-Rodríguez, y J. A. Garza-Alonso, “La analítica de datos como ventaja competitiva en las organizaciones”, VinculaTégica, vol. 6, n.º 2, pp. 1063–1074, dic. 2020.

[5] D. R. Anderson, D. J. Sweeney y T. A. Williamns, *Estadística para Administración y Economía*, 10a ed, México, D.F: Cengage Learning, 2008.

[6] D.A. Villegas. “Importancia de la estadística aplicada para la toma de decisiones en marketing”, Revista Boliviana de Administración, vol. 3, no. 2, pp. 63-74, jul-dic, 2021.

[7] G. Petersen, *High-Impact Sales Force Automation: A Strategic Perspective*, USA, CRC Press, 2023, pp. xiii

[8] M. Easterby-Smith, R. Thorpe, P.R. Jackson, L.J. Jaspersen, *Management and Business Research*, 7a ed, NY: Sage, 2021, pp. 171