SEGMENTACIÓN AUTOMATICA DE LOS CLIENTES DE PHARMADERM Y SKINDRUG, UTILIZANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

JOHN DIAZ ALONSO

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

GLORIA INES ALVAREZ
Director

GERARDO MAURICIO SARRIA
Jurado

DIEGO LINARES
Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ingeniería de Software

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Santiago de Cali, 12 de Septiembre del 2022

Acta de Correcciones al Documento de Trabajo de Grado

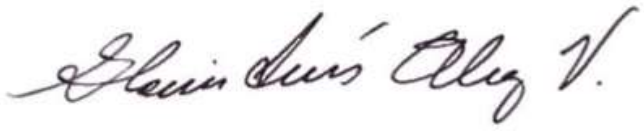
Santiago de Cali, septiembre 8 de 2022

Autor: John Diaz Alonso

Título del Trabajo de Grado: "SEGMENTACIÓN AUTOMATICA DE LOS CLIENTES DE PHARMADERM Y SKINDRUG, UTILIZANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO"

Director: Gloria Inés Alvarez V.

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.



Firma del Director del Trabajo de Grado

DATOS DEL ESTUDIANTE

Nombre Completo: JOHN DIAZ ALONSO
Dirección: Calle 142C N.141B-60 – Agrupación 9 - Casa 108 – Bogotá D.C.
Teléfonos: 3204216833 - 6014793295
Correos Electrónicos: johndiazalonso@javerianacali.edu.co,
john.diaz.alonso@outlook.com
Profesión: Ingeniera de Sistemas: Universidad Cooperativa de Colombia, 2015
Empresas: PHARMADERM S.A. y SKINDRUG S.A.
Cargo: Gerente de TI

SEGMENTACIÓN AUTOMÁTICA DE LOS CLIENTES DE PHARMADERM Y SKINDRUG, UTILIZANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

JOHN DÍAZ ALONSO
Código 8961548

Proyecto de grado para optar al título de Magíster en Ingeniería de Software

Director(a)
GLORIA INÉS ÁLVAREZ, PhD

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN INGENIERÍA DE SOFTWARE
SANTIAGO DE CALI, JULIO 25 DE 2022

TABLA DE CONTENIDO

ABSTRACT

The marketing department of PHARMADERM S.A. and SKINDRUG S.A., does not have a tool to identify customers with similar needs to carry out targeted marketing, with the purpose of developing new specific marketing strategies, and thus increase customer purchases.

For this reason, the project aims to develop a conglomerate identification model of the database of the laboratories' clients, through the preparation of the data, the models of grouping of the clients, the evaluation and recommendation of the selected model. and a display means for the results obtained.

In the preparation of the data, a dataset with 17 attributes and 852,448 records, corresponding to 683 clients, was unified, and subsequently 6 datasets were generated (gifts, type of client, demographic, cosmetics, magistrals and medicines) according to the business approach. determined by the sales and marketing department.

Subsequently, hyper-parameters were selected for the training models, k-means and hierarchical with the elbow method, and the silhouette coefficient, and for the DBSCAN model, the k neighbors' algorithm closest neighbors (k-nearest neighbors' algorithm).

Finally, in the evaluation and selection of the model, the K-means training model is chosen, determined with the Davies-Bouldin index, with the indices closest to zero (0) being the best, indicating more compact clusters, and with the centers of each cluster well separated. Additionally, an analysis of the models obtained with the radar chart, also known as the spider chart or star chart, is carried out, accompanied by the marketing and sales department for their assessment.

RESUMEN

El departamento de mercadeo de los laboratorios PHARMADERM S.A. y SKINDRUG S.A., no cuenta con una herramienta que permita identificar clientes con necesidades simulares para hacer un *marketing* dirigido, con el propósito de desarrollar nuevas estrategias de mercadeo específicas, y así aumentar las compras de los clientes.

Por esta razón, el proyecto tiene como objetivo desarrollar un modelo de identificación de conglomerados de la base de datos de los clientes de los laboratorios, mediante la preparación de los datos, los modelos de agrupamiento de los clientes, la evaluación y recomendación del modelo seleccionado y un medio de visualización para los resultados obtenidos.

En la preparación de los datos se unifico un *dataset* con 17 atributos y 852.448 registros, correspondientes a 683 clientes, y posteriormente se generaron 6 *datasets* (obsequios, tipo de cliente, demográfico, cosméticos, magistrales y medicamentos) de acuerdo con el enfoque de negocio que determino el departamento de mercadeo y ventas.

Posteriormente, se seleccionaron los hiper-parámetros para los modelos de entrenamiento, *k-means* y jerárquico con el método del codo (*elbow method*), y el coeficiente de la silueta (*silhouette coefficient*), y para el modelo DBSCAN, el algoritmo k vecinos más cercanos (*k-nearest neighbors algorithm*).

Finalmente, en la evaluación y selección del modelo se escoge el modelo de entrenamiento K-means, determinado con el índice de Davies Bouldin (*Davies-Bouldin index*), con mejores los índices más cercanos a cero (0), señalando clústeres más compactos, y con los centros de cada clúster bien separados. Adicional, se realiza un análisis de los modelos obtenidos con el gráfico del radar también conocido como gráfico de araña o grafico de estrella, acompañado con el departamento de mercadeo y ventas para su valoración.

INTRODUCCIÓN

Los datos extraídos de las ventas, como las características de los pedidos y las devoluciones de los clientes, se han convertido en uno de los activos más importantes de las compañías, debido a que, contienen información y características propias, que con un procesamiento adecuado se puede convertir en información estratégica, y con el análisis correcto y las técnicas apropiadas pueden llegar a ser información relevante para segmentar características únicas de los clientes.

Por esta razón, en la actualidad, ha crecido el análisis de los datos con *Machine Learning* o aprendizaje automático, que ofrece alternativas eficientes para capturar el conocimiento de los datos, y mejorar gradualmente el rendimiento de los modelos de *Clustering* y tomas de decisiones basadas en esos datos.

Cabe destacar que en Colombia el crecimiento de la analítica de datos, apoyado en la ciencia de los datos y la inteligencia artificial, ha venido creciendo aceleradamente, existen comunidades de expertos como Colombia AI, con 5.700 miembros, que se dedican a realizar eventos, compartir experiencias y conocimiento de *Data Science*, Inteligencia Artificial y *Machine Learning* [1]. En el ámbito educativo, las universidades también han visto el futuro con la analítica de los datos, y según el Sistema Nacional de Información de la Educación Superior SNIES, existen activos e inscritos 5 pregrados, 9 especializaciones y 15 maestrías, como la Maestría en Ciencia de Datos de la Pontificia Universidad Javeriana.

Finalmente, los laboratorios dermatológicos PHARMADERM y SKINDRUG, buscan mejorar la segmentación de los clientes para las campañas de *marketing*, con el fin de descubrir cuál podría ser un agrupamiento estratégico de los clientes, para generar estrategias específicas enfocadas a cada grupo de clientes.

1. DEFINICIÓN DEL PROBLEMA

Los laboratorios PHARMADERM S.A. y SKINDRUG S.A., son un grupo empresarial que tienen como actividad económica fabricar y comercializar productos dermatológicos. Entre las categorías que los laboratorios comercializan están los cosméticos, que sirven para cuidar y embellecer la piel y el cabello; los medicamentos, para el tratamiento y prevención de patologías cutáneas; los magistrales, fórmulas destinadas a pacientes específicos; las materias primas, insumo para la fabricación de los productos; y el arrendamiento, de bodegas y oficinas. Al respecto, el mayor porcentaje de ventas lo integran las categorías de cosméticos y medicamentos, con sus líneas de productos más significativas que son los protectores solares, limpiadores faciales, champús anticaídas e hidratantes faciales y corporales, con sus marcas más relevantes Sunstop, Sunface, Synderm, Gelclin, Piloskin, entre otros. Además, cuenta con clientes a nivel local en la ciudad de Bogotá, donde tiene su sede principal y clientes a nivel nacional e internacional en el país de Panamá.

Adicionalmente, la gerencia de mercadeo tiene como estrategia generar promociones, descuentos, publicidad, y otros planes de *marketing* de manera general, pero no de forma dirigida a sus clientes, debido a que, los sistemas de información actuales del laboratorio no permiten realizar una segmentación de los clientes, lo que ayudaría como punto de partida para el análisis del departamento de mercadeo sobre sus estrategias, como, por ejemplo, recomendar el producto adecuado a los clientes.

1.1. PLANTEAMIENTO DEL PROBLEMA

El departamento de mercadeo de PHARMADERM y SKINDRUG tiene el objetivo de generar estrategias de mercadeo específicas para aumentar las compras de los clientes. En la actualidad, para determinar qué productos van a comprar los clientes se analizan los datos disponibles de manera manual, lo cual es costoso y tarda mucho tiempo, pues no se cuenta con una herramienta que permita identificar clientes con necesidades simulares para hacer un mercadeo dirigido, proponiendo diferentes ofertas a cada grupo de clientes. Este sistema sería de gran importancia debido a que generaría un ahorro en tiempo y costos. Además, permitiría desarrollar nuevas estrategias de mercadeo, que en el momento no puede implementar.

2. FORMULACIÓN DEL PROBLEMA

De acuerdo con lo anterior, surge la siguiente pregunta:

- ¿Cómo realizar un análisis de conglomerados de los clientes de los laboratorios PHARMADERM y SKINDRUG, para apoyar las estrategias definidas por el departamento de mercadeo?

  - ¿Qué datos son accesibles y cómo se prepararán para realizar la segmentación de los clientes?
  - ¿Cómo construir un modelo de agrupación, que sirva mejor al propósito de análisis?
  - ¿Cómo se evaluará la calidad de los modelos desarrollados?
  - ¿Cómo visualizar los resultados obtenidos para facilitar su interpretación?

2.2. OBJETIVOS DEL PROYECTO

2.2.1. Objetivo General

Desarrollar un modelo de identificación de conglomerados de la base de datos de los clientes de los laboratorios PHARMADERM y SKINDRUG, para apoyar las estrategias del departamento de mercadeo.

2.2.2. Objetivos Específicos

1. Preparar los datos para realizar la segmentación de los clientes.
2. Desarrollar modelos de agrupación mediante la aplicación de técnicas basadas en centroides, jerárquicas, basadas en densidad, para que el departamento de mercadeo pueda generar nuevas estrategias.
3. Evaluar los modelos desarrollados para seleccionar el modelo a recomendar.
4. Desarrollar una visualización de los resultados obtenidos que faciliten su interpretación.

2.2.3. Resultados Esperados

- Base de datos de clientes preparada para realizar al análisis de conglomerados.
- El modelo de agrupación de clientes debidamente entrenado y probado.
- Documento con el desarrollo del proyecto y los resultados de los análisis realizados.

— Medio de visualización que permita utilizar el modelo de agrupamiento de los clientes y visualizar los resultados.

2.3. ALCANCE

Se busco la segmentación de los clientes, incluyendo los modelos de aprendizaje automático, que permitieron realizar una agrupación de los clientes, para el departamento de mercadeo de los laboratorios dermatológicos PHARMADERM y SKINDRUG S.A., los alcances para este proyecto son:

— El proyecto se desarrolló utilizando la información de los clientes contenida en la base de datos del ERP de los laboratorios PHARMADERM y SKINDRUG.

— Para la construcción del Modelo de Clustering, se exploró 3 algoritmos, correspondientes a los enfoques más reconocidos en el área. Los algoritmos son: K-Means, DBSCAN (Density-based Spatial Clustering of Applications with Noise) y Hierarchical clustering.

— Las ventas entre el grupo empresarial y el arrendamiento a terceros, no se tuvieron en cuenta para el modelo, debido a que no son considerados clientes estratégicos para análisis del departamento de mercadeo.

2.4. JUSTIFICACIÓN

Dada la necesidad de los laboratorios PHARMADERM y SKINDRUG de realizar la segmentación de los clientes de modo automático, se estableció que es necesario desarrollar un modelo de segmentación, apoyado en la concepción y el modelo de negocio de las compañías, permitiendo, fijarse en un grupo de clientes con características similares, para identificar patrones que no son visibles y construir conocimiento orientado al mercadeo.

Adicionalmente, se dio inicio al proyecto y garantizo el desarrollo de este, se cuenta con la autorización para acceder a la base de datos con la que se va a construir el modelo de segmentación. Del mismo modo, el acceso a los recursos tecnológicos, como desktop y acceso a internet en la compañía. Así mismo, los laboratorios suministraron los recursos económicos en el caso de algún tipo de licenciamiento, previo de una aprobación y autorización de la gerencia general para dar viabilidad al proyecto.

Por otro lado, el departamento de mercadeo tiene mejores herramientas en el mercadeo dirigido gracias al desarrollo de este proyecto. Adicionalmente, esto permitirá posicionar a los laboratorios entre las compañías más innovadoras que incorporan tecnologías como Machine Learning para tener una ventaja competitiva ante sus competidores.

3. MARCO TEÓRICO DE REFERENCIA

3.1. CIENCIA DE DATOS

Inicialmente, el núcleo donde comienza el estudio de los datos es determinado por la Ciencia de los Datos, según Drew Conway se trata de un tema interdisciplinario compuesto por tres disciplinas: la Estadística, que modela y resume el conjunto de datos; la Ciencia Informática que diseña y usa los algoritmos para almacenar, procesar y visualizar los datos y la experiencia en un campo determinado, el cual es un conjunto de habilidades que se puede aplicar en un área de especialización [2], de aquí es donde entran varios factores como el Análisis de Datos.

3.2. ANALITÍCA DE DATOS

La Analítica de Datos, con el apoyo de la Inteligencia Artificial busca mecanismos que ayuden a comprender la inteligencia y realizar modelos y simulaciones de estos [3], específicamente con *Machine Learning* que es un tipo de inteligencia artificial mediante el cual un algoritmo o método extrae patrones de los datos.

3.3. INTELIGENCIA ARTIFICIAL

Además, la Ciencia de la Informática tiene el campo de la Inteligencia Artificial, que en los años cincuenta John McCarthy lo definió como la simulación de la inteligencia humana por un computador, con el fin de hacer que la máquina sea capaz de identificar y utilizar el conocimiento en una etapa determinada de la solución de un problema planteado y en el análisis de información, su objetivo es el tratamiento de forma masiva y automática con una potencia de cálculo suficiente de información que pudiera ser compleja [4].

3.4. MACHINE LEARNING

Adicionalmente, con *Machine Learning* una de las ramas de la Inteligencia Artificial y según Dan Fagella, es la ciencia que permite que las computadoras aprendan y actúen como lo hacen los seres humanos, mejorando su aprendizaje a lo largo del tiempo de una forma autónoma, alimentándolas con datos e información en forma de observaciones e interacciones con el mundo real, resolviendo algunos problemas generales [5], como los que se enuncian en la tabla a continuación:

Tabla 1: Los problemas que puede resolver Machine Learning [5].

Problemas	Categorías de Machine Learning
Ajustar algunos datos a una función o aproximación de función.	Aprendizaje supervisado.
Averiguar cuáles son los datos sin ningún comentario.	Aprendizaje no supervisión.

Maximizar las recompensas a lo largo del tiempo.	Aprendizaje reforzado.

En *Machine Learning* existen tres tipos o categorías principales de aprendizaje automático: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje reforzado. En relación con el siguiente proyecto, este se enfocará en el aprendizaje no supervisado:

3.5. APRENDIZAJE NO SUPERVISADO

Modela las características de un conjunto de datos sin hacer referencia a ninguna etiqueta y, a menudo, se describe como "dejar que el conjunto de datos hable por sí mismo". Estos modelos incluyen tareas como agrupamiento y reducción de dimensionalidad. Los algoritmos de agrupamiento identifican distintos grupos de datos, mientras que los algoritmos de reducción de dimensionalidad buscan representaciones más concisas de los datos [6].

Por consiguiente, se desarrollará el proyecto tomando las técnicas de *Clustering* o agrupamiento:

3.6. AGRUPAMIENTO (*CLUSTERING*)

Es el análisis de grupo (*cluster analysis*), que permite descubrir estructuras ocultas en datos en los cuales no conocemos la respuesta correcta por adelantado. El objetivo del agrupamiento o *clustering* es encontrar cómo agrupar de forma natural los datos de manera que los elementos del mismo grupo sean más parecidos unos a otros que aquellos de diferentes grupos [6].

Cabe destacar que, para la construcción del Modelo de Clustering, se van a explorar 3 algoritmos, correspondientes a los enfoques más reconocidos en el área. Los algoritmos son: K-Means, DBSCAN (Density-based Spatial Clustering of Applications with Noise) y Hierarchical clustering.

3.6.1. K-Means

El algoritmo k-means pertenece a la categoría de agrupamiento basados en prototipos, lo que significa que cada clúster está representado por un prototipo, que puede ser tanto el centroide (promedio) de puntos similares con características continuas como el medoide (el punto más representativo o que parece con más frecuencia), uno de los inconvenientes es que tenemos que especificar el número de grupos [6].

Módulo cluster de scikit-learn [7]:

$$KMeans(n\_clusters = 3, init = k - means + +, n\_init = 10)$$

Híper-parámetros utilizados de la librería scikit-learn, para la construcción del modelo;

— n_clusters: Número de conglomerados que se formarán, como el número de centroides que se generarán, de tipo int.
— Init: Método de inicialización. *Predeterminado* = k-means++ o random.
— n_init: Número de veces que se ejecutará el algoritmo de k-medias con diferentes semillas de centroide. *Predeterminado = 10*.

3.6.2. DBSCAN (*Density-based Spatial Clustering of Applications with Noise*)

El DBSCAN, agrupamiento espacial basado en densidad de aplicaciones con ruido, como el nombre lo indica, el agrupamiento basado en densidad asigna etiquetas de grupos basadas en regiones de puntas densas. En DBSCAN, la noción de densidad se define como el número de puntos dentro de un radio especifico [6].

Módulo cluster de scikit-learn [8]:

$$DBSCAN\ (\ eps\ =\ 0.5\ ,\min\ \_samples=\ 5\ ,metric\ =\ 'euclidean')$$

Híper-parámetros utilizados de la librería scikit-learn, para la construcción del modelo;

— eps: La distancia máxima entre dos muestras para que una se considere próxima a la otra. Este no es un límite máximo en las distancias de los puntos dentro de un grupo. Este es el parámetro DBSCAN más importante para elegir adecuadamente para su conjunto de datos y función de distancia. Predeterminado = 0.5, de tipo float.

— min_saples: El número de muestras (o peso total) en una vecindad para que un punto se considere como un punto central. Predeterminado = 5, de tipo int.

— metric: La métrica que se utilizará al calcular la distancia entre instancias en una matriz de características. Predeterminado = 'euclidean', de tipo str.

3.6.3. Jerárquico (*Hierarchical clustering*).

El agrupamiento jerárquico aglomerativo, divide el conjunto de datos en jerarquías que requieren un punto de corte manual, lo que permite elegir el número de grupos que se quieren devolver, es útil si se quiere podar el árbol de grupos jerárquicos [6].

Módulo cluster de scikit-learn [9]:

$$AgglomerativeClustering(n\_clusters=2, affinity='euclidean')$$

Híper-parámetros utilizados de la librería *scikit-learn*, para la construcción del modelo;

— n_clusters: El número de agrupaciones que se van a encontrar, de tipo int.
— affinity: Métrica utilizada para calcular el vínculo. Puede ser " euclidean", "l1", "l2", " manhattan", " cosine" o " precomputed", es de tipo str.
— Linkage: Criterio de vinculación determina qué distancia usar entre conjuntos de observación.

3.7. MÉTRICAS DE VALIDACIÓN.

Los ingenieros de datos emplean una gran cantidad de tiempo mejorando la calidad de los resultados de los modelos obtenidos, normalmente no es posible saber visualmente si un modelo es mejor que otro; resulta necesario procesar algunas medidas de calidad para conocer la exactitud de los resultados [10]. En efecto, para la estimación de parámetros de los modelos desarrollados, se utilizaron el método del codo (*Elbow method*), el coeficiente de la silueta (*Silhouette Coefficient*) y el algoritmo k vecinos más cercanos (k-nearest neighbors algorithm) y para la evaluación de estos, el índice de Davies Bouldin (*Davies-Bouldin index)*, explicados a continuación:

3.7.1. Método del Codo (*Elbow Method*).

El método del codo ayuda a elegir un número apropiado de clústeres para agrupar los datos, el cual utiliza los valores de la inercia obtenidos tras aplicar el K-means a diferente número de clústeres, desde 1 hasta N clústeres, siendo la inercia la suma de las distancias al cuadrado de cada objeto del clúster a su centroide, calculado como:

$$Inertia = \sum_{i=0}^{N} ||xi - \mu||^2$$

La representación generalmente es por medio de una gráfica lineal de la inercia respecto al número de clústeres, en la cual se apreciará un cambio brusco en la evolución de la inercia, la cual indicará un número óptimo de clústeres [11].

3.7.2. Coeficiente de la Silueta (*Silhouette Coefficient*).

El análisis de silueta es una medida intrínseca para evaluar la calidad de un agrupamiento, limitado a un rango de -1 a 1, calculado como:

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

Basados en la ecuación anterior, el coeficiente de silueta es 0 si la separación y la cohesión del grupo son iguales ($b^{(i)} = a^{(i)}$) . Además, cuando se acerca a un coeficiente de silueta 1 si $b^{(i)} \gg a^{(i)}$, puesto que $b^{(i)}$ cuantifica la desigualdad de una muestra frente a otros grupos, y $a^{(i)}$ indica lo igual que es a las otras muestras de su mismo grupo[6].

3.7.3. Índice de Davies Bouldin (*Davies-Bouldin inde*x).

El índice se define como la medida de similitud promedio de cada grupo con su grupo más similar, donde la similitud es la relación entre las distancias dentro del grupo y las distancias entre grupos. Por lo tanto, los grupos que están más separados y menos dispersos darán como resultado una mejor puntuación [12], calculado como:

$$DB = \frac{1}{k}\sum_{i=1, i\neq j}^{k} \max\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

En el cual k es el número de clústeres, σ_i es la distancia promedio entre cada punto en el clúster i y el centroide del clúster, σ_j es la distancia promedio entra cada punto del clúster j y el centroide del clúster, y $d(c_i, c_j)$ es la distancia entres los centroides de los 2 clústeres.

3.7.4. Algoritmo k vecinos más cercanos (*k-nearest neighbors algorithm*).

El algoritmo de k vecinos más cercanos proporciona una funcionalidad para métodos de aprendizaje no supervisados para implementar búsquedas de vecinos. El número de muestras puede ser una constante definida por el usuario (k-aprendizaje del vecino más cercano) [13].

Módulo de vecinos más cercanos de scikit-learn

$$NearestNeighbors(n\_neighbors = 5, metric = 'minkowski', p = 2)$$

Híper-parámetros utilizados de la librería *scikit-learn*, para la construcción del modelo [14];

- n_neighbors: Número de vecinos a usar por defecto para las consultas para kneighbors
- metrics: Métrica de distancia que se usará para el árbol.
- p: Parámetro para la métrica

3.8. TRABAJOS RELACIONADOS

- El trabajo de grado aplicado: "Desarrollo de un modelo de clusterización de los taxistas para la rentabilidad del segmento corporativo de Smart Taxi", de la Maestría en Analítica para la Inteligencia de Negocios, de la Pontificia Universidad Javeriana sede Bogotá D.C., se enfoca en construir una segmentación de taxistas según el historial de sus carreras para identificar los atributos que los caracterizan y que permitan ofrecer mejores niveles de servicios a los clientes corporativos a partir de la construcción de estrategias basadas en sus comportamientos y hábitos [15]. Las técnicas de modelado utilizadas en este trabajo de grado fueron, la técnica de *K-means* y *Mclust* y las conclusiones referentes a las técnicas utilizadas fueron que el modelo *K-means* fue el mejor debido a que se identifican mejor los grupos, que dan sentido al negocio.
- Otro trabajo de grado revisado fue en la tesis sobre analítica de datos, con el nombre: "Modelo de planeación de inventarios para *E-commerce*, utilizando herramientas de inteligencia artificial para hacer pronósticos de demanda y clasificación de inventarios", de la Maestría en Ingeniería Industrial de la Universidad de los Andes, y lo que busca es la creación de un modelo de planeación de

inventarios *E-commerce*, resaltando el impacto económico que tiene el inventario en las finanzas de la compañía analizada. Las técnicas de modelado utilizadas en este trabajo de grado fueron*, K-means, Hierarchical Clustering, DBSCAN, Gaussian Mixture Model* y las conclusiones referentes a los modelos utilizados fueron que el modelo *K-means* fue el mejor debido a tener mejores resultados[16].

3.9. TÉRMINOS Y DEFINICIONES.

- Cohesión: Los objetos de cada clúster debe ser lo más cercano posible a los otros objetos del mismo clúster.
- Separación: Los clústeres deben estar ampliamente separados entre ellos. Existen varios enfoques para medir esta distancia entre clúster: distancia entre el miembro más cercano, distancia entre los miembros más distantes o la distancia entre los centroides.
- Scikit-Learn: Librería de Python de código abierto para el aprendizaje automático, que ayuda en el preprocesamiento, la reducción de dimensionalidad (selección de parámetros), la clasificación, la regresión, la agrupación y la selección de modelos. Entre sus principales características se destacan Clustering.
- Parámetro: en un modelo, son las variables que se estiman durante el proceso de entrenamiento con los conjuntos de datos.
- Hiperparámetro: en un modelo, son los valores de las configuraciones utilizadas durante el proceso de entrenamiento.
- Centroide: Promedio de puntos similares con características continuas como el medoide, el punto más representativo o que aparece con más frecuencia.
- Puntos de ruido: Se consideran todos los puntos que no son ni núcleo ni borde.
- Maldición de la dimensionalidad: Es un problema que se presenta si se quieren tener en cuenta todas las características o atributos posibles en el sistema.
- Distancia Euclídea Raíz cuadrada del cuadrado de la distancia (L2).
- Distancia Manhattan Valores absolutos de la distancia entre los puntos cartesianos (L1).
- Distancia de Minkowski: Esta distancia puede considerarse una generalización de las distancias euclideas y Manhattan.
- StandardScaler: Método para la estandarización de las características eliminando la media y escalando a la varianza de la unidad.
- NearestNeighbors: Método de aprendizaje no supervisado para implementar búsquedas de vecinos.

4. DESARROLLO DEL PROYECTO

En esta sección se describe: la preparación de los datos, la selección de registros, la limpieza de los datos, la selección de hiper-parámetros por medio de experimentos, el entrenamiento de los modelos k-means, DBSCAN y Hierarchical y las dificultades presentadas y estrategias implementadas para solucionarlas.

4.1. PREPARACIÓN DE LOS DATOS

En la preparación de los datos, se debe llevar a cabo numerosas tareas específicas tales como la identificación de la fuente de los datos, su limpieza, la eliminación de información que está fuertemente correlacionada, la búsqueda de información sesgada, y la realización de las normalizaciones necesarias [10].

4.1.1. Identificación de los datos.

Inicialmente, los datos fueron extraídos del sistema ERP de la compañía, concretamente de la tabla de ventas, con las transacciones de los clientes, como las facturas de venta, las devoluciones realizadas, con un periodo de tiempo desde el año 2014 hasta el año 2021, y con un total de 852.448 registros.

Los atributos del *dataset* inicial con el que se dispone a realizar la selección y limpieza de los datos, es:

Tabla 2: Atributos *Dataset* Inicial

	Mes_Compra	Items	Cant_Total_Venta	Descuento_%	Valor_Venta	Valor_Descuento	CANAL	CATEGORIA	MARCA	Ciudad	Condicion_Pago	Cupo_Crédito	Grupo_Dscto	Departamento	Cant_Obsequio	ZONA	Client
0	1	4383	12.0	0.0000	225000.0	0.0	1	23	15	11001	30	20000000	1	11	0	1	626
1	1	4384	12.0	0.0000	225000.0	0.0	1	23	15	11001	30	20000000	1	11	0	1	626
2	1	3486	20.0	0.5000	391500.0	391500.0	1	22	30	11001	30	20000000	3	11	10	1	83
3	1	2701	15.0	0.3333	212011.0	105989.0	1	23	15	66001	30	60000000	1	66	5	2	96
4	1	2701	15.0	0.3333	212011.0	105989.0	1	23	15	11001	30	30000000	1	11	5	1	265
5	1	2701	7.0	0.2857	106002.0	42398.0	1	23	15	11001	30	8000000	3	68	2	1	20
6	1	2284	6.0	0.0000	343800.0	0.0	1	3	6	11001	30	8000000	3	68	0	1	20
7	1	21	30.0	0.3333	241012.0	120488.0	1	1	9	11001	30	8000000	3	68	10	1	20
8	1	3784	13.0	0.2307	273025.0	81875.0	1	6	18	11001	30	8000000	3	68	3	1	20
9	1	1859	10.0	0.0000	212000.0	0.0	1	23	15	54001	30	45000000	1	54	0	6	407

De manera que, la descripción de los atributos, son:

- Mes_Compra: Mes en que se realizaron las compras por parte de las farmacias o droguerías de cadena, como Olímpica, Cafam, o Colsubsidio, entre otros.
- Items: Código interno del producto de venta, asignado automáticamente en sistema ERP.
- Cantidad_Total_Venta: Cantidades totales de venta de cada producto en registro de venta.
- Descuento_%: Descuento otorgado por producto, según política de venta.
- Valor_Venta: Valor total de la venta de cada registro.
- Valor_Descuento: Descuento otorgado por producto en valores.
- CANAL: Identificación asignada a los clientes, como farmacias, institucional, etc.

- CATEGORIA: Identificación asignada a los productos, como protección solar, hidratante facial, antiacné oral, entre otras.
- MARCA: Identificación de cada producto según a la marca que corresponda, como HYDRASKI, SUNFACE, SUNSTOP, etc.
- Ciudad: Código de la ciudad de venta en cada registro.
- Condicion_Pago: Plazo máximo de pago otorgado a cada cliente, como 0, 30, 45 y 60 días.
- Cupo_Credito: Cupo máximo asignado a cada cliente en valores.
- Grupo_Descuento: Clasificación de los descuentos, otorgados a cada cliente.
- Departamento: Código del departamento de venta en cada registro.
- Cant_Obsequida: Unidades obsequiadas en producto.
- Zona: Código de las regiones de venta.
- Cliente: Numeración asignada al NIT del cliente (Número de Identificación Tributaria).

Tabla 3: Estadísticas descriptivas del *dataset* inicial

	Mes_Compra	Items	Cant_Total_Venta	Descuento_%	Valor_Venta	Valor_Descuento	CANAL	CATEGORIA	MARCA	Ciudad	Condicion_Pago	Cupo_Crédito	Grupo_Dscto	Departamento	Cant_Obsequio	ZONA	Client
count	852448.000000	852448.000000	8.524480e+05	852448.000000	8.524480e+05	8.524480e+05	852448.000000	852448.000000	852448.000000	852448.000000	852448.000000	8.524480e+05	852448.000000	852448.000000	852448.000000	852448.000000	852448.000000
mean	6.558761	4050.131580	1.510020e+02	0.143850	3.959502e+05	1.062494e+05	2.347512	18.865984	18.360804	20476.781133	33.395353	1.486308e+08	2.831296	19.631115	2.806927	1.551593	303.100696
std	3.307888	2706.233725	1.312679e+04	0.199793	1.798677e+06	4.613407e+05	2.508126	6.040935	7.979203	20924.564671	14.012546	2.037324e+08	3.616356	19.581002	11.926599	1.189267	184.489658
min	1.000000	1.000000	-8.500000e+05	0.000000	-6.794600e+08	-1.985600e+07	1.000000	1.000000	1.000000	5001.000000	0.000000	0.000000e+00	0.000000	5.000000	-360.000000	1.000000	1.000000
25%	4.000000	1501.000000	1.000000e+00	0.000000	3.360000e+04	0.000000e+00	1.000000	17.000000	15.000000	11001.000000	30.000000	2.000000e+07	1.000000	11.000000	0.000000	1.000000	161.000000
50%	7.000000	4617.000000	5.000000e+00	0.000000	1.105000e+05	0.000000e+00	1.000000	22.000000	15.000000	11001.000000	30.000000	5.000000e+07	3.000000	11.000000	0.000000	1.000000	243.000000
75%	9.000000	6792.000000	1.200000e+01	0.333300	2.958570e+05	7.400000e+04	1.000000	23.000000	25.000000	11001.000000	30.000000	2.000000e+08	3.000000	11.000000	2.000000	1.000000	484.000000
max	12.000000	8827.000000	6.300000e+06	1.000000	6.794600e+08	2.915000e+07	9.000000	25.000000	37.000000	99999.000000	90.000000	1.000000e+09	40.000000	99.000000	855.000000	7.000000	683.000000

Por otro lado, se verifica el resumen de las estadísticas descriptivas de los atributos del *dataset* inicial, como: el número de ítems, el cual corresponde a los creados en el ERP, al igual que la condición de pago, que su máximo plazo es de 90 días, y la media está alrededor de los 30 días, la cual encaja con el plazo generalmente otorgado. Los valores negativos en los registros de venta, cantidades de compra, entre otros, pertenecen a devoluciones realizadas por los clientes.

En general, se van a analizar 852.448 registros correspondientes a 683 clientes, contenidos en el *dataset* inicial.

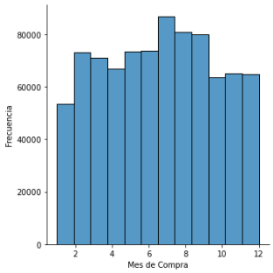


Imagen 1: Histograma con la frecuencia de registros por mes.

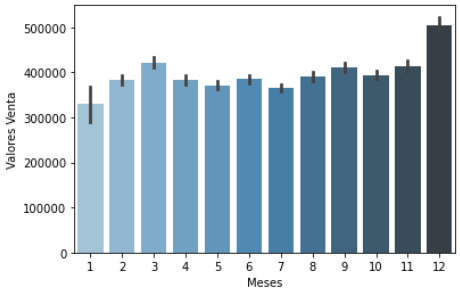


Imagen 2: Diagrama de ventas mensuales expresando en valores.

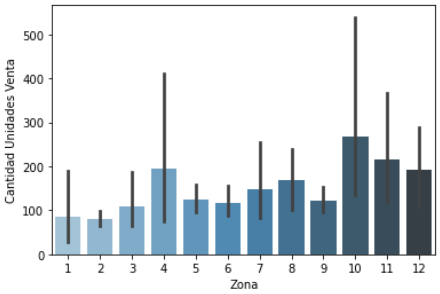


Imagen 3: Diagrama de unidades vendidas mensualmente.

Por otra parte, en la Imagen 1, podemos observar el número de registros por mes, se observa claramente cómo en enero hay menos registros, lo cual se explica por el hecho que en este mes solo se trabajan 15 días debido al periodo de vacaciones

colectivas para los empleados. En la Imagen 2, se puede observar que diciembre él es mes con las mayores ventas, debido a la temporada y el aumento del precio para el próximo año, y por último en la Imagen 3, refleja en cuatro meses, un aumento en las unidades vendidas, en abril se debe a que el inventario comprado por los clientes a finales del año disminuye en los tres primeros meses del año, por lo que deben volver a comprar, y octubre, noviembre y diciembre se debe a la temporada de final de año.

4.1.2. Selección de registros y limpieza de los datos.

En la preparación de los modelos se utilizaron cuatro *Datasets* obtenidos del ERP Siesa Enterprise, dos con la información de los movimientos de las ventas de los clientes y otros dos con la información propia de cada cliente, uno para cada laboratorio PHARMADERM S.A. y SKINDRUG S.A., los cuales se concatenaron en un solo *Dataset* utilizando como llave el NIT (Número de Identificación Tributaria) del cliente.

Después de realizar la verificación del d*ataset*, se procedió a:

- Eliminación de registros descrito en el alcance, como las ventas entre el grupo empresarial y el arrendamiento a terceros, que no se tendrá en cuenta para el modelo, debido a que no son considerados clientes estratégicos para el análisis del departamento de mercadeo, descritos en la tabla a continuación:

Tabla 4: Registros eliminados

Tercero	Registro
PHARMADERM SA	830058969
SKINDRUG SA	900098045
LABORATORIO DERMAELITE SAS	900392555
RTS SAS	805011262

- Eliminación de información que está correlacionada, como atributos con el código y descripción en diferentes campos y atributos no significativos.
- Renombramientos de variables con espacios y con signos de puntuación.
- Asignación a los campos con valores NaN, basados en las sugerencias del departamento de Mercadeo y Ventas.
- Creación de los seis *Datasets*, según el enfoque de negocio determinado por el departamento de Mercadeo y Ventas.

Por último, a partir del *dataset* original, se hace la creación de los 6 *datasets* finales, según el enfoque de negocio sugerido por el departamento de Mercadeo y Ventas, en donde se filtraron los registros y se eliminan los atributos de acuerdo con el criterio del experto de Mercadeo y Ventas de la compañía, de la siguiente manera:

- dataObsequios: Se filtraron y se dejan los productos con valores positivos o negativos en el atributo Cant_Obsequio, y se eliminan los atributos Cant_Total_Venta, Valor_Venta, CANAL, Condición_Pago, Cupo_Credito, Grupo_Dscto y ZONA.

- dataTipoCliente: Se excluyen los clientes de uso institucional en el atributo Client y se eliminan los atributos Mes_Compra, Items, Cant_Total_Venta, Descuento_%, Valor_Descuento, Valor_Descuento, CANAL, CATEGORIA y MARCA.
- DataDemográfico: Se excluyen los registros con las ventas realizadas al país de Panamá, y se eliminan los atributos Mes_Compra, Descuento_%, Valor_Descuento, CANAL, CATEGORIA, MARCA, Condición_Pago, Cupo_Credito, Grupo_Dscto y Cant_Obsequio.
- dataCosméticos: Se filtran los registros con los productos cosméticos clasificados el atributo CATEGORIA, y se eliminan los atributos Descuento_%, Valor_Descuento, Condicion_Pago, Cupo_Credito, Grupo_Dscto, Cant_Obsequio, y ZONA.
- dataMagistrales: Se filtran los registros con los ítems de fórmulas magistrales clasificados el atributo items, y se eliminan los atributos Mes_Compra, Items, Descuento_%, Valor_Descuento, CANAL, CATEGORIA, MARCA, Condicion_Pago, Cupo_Credito, Grupo_Dscto y Cant_Obsequio.
- dataMedicamentos: Se filtran los registros con los productos medicamentos clasificados el atributo CATEGORIA, y se eliminan los atributos Valor_Descuento, Descuento_%, Condición_Pago, Grupo_Dscto, Cupo_Credito, Cant_Obsequio y ZONA.

En la tabla 5 se describen los *datasets* generados para el entrenamiento de los modelos:

Tabla 5: *Datasets* Finales por enfoque de Negocio

Nombre	Registros	Atributos	Descripción
dataObsequios	307.807	10	Obsequios y descuentos otorgados a los clientes según políticas comerciales.
dataTipoCliente	835.412	8	Características del cliente, como: Condición de pago, grupo de descuento al que pertenecen.
dataDemográfico	839.487	7	Zona demográfica del cliente, como: Ciudad, departamento.
dataCosméticos	359.003	10	Productos dermocosméticos, como: protectores solares, limpiadores e hidrantes corporales y faciales.
dataMagistrales	349.289	6	Productos especializados que varían en su composición según la patología.
dataMedicamentos	126.918	10	Medicamentos, prescritos y venta baja formula médica.

4.2. DESARROLLO DE LOS MODELOS.

El modelado se suele ejecutar en múltiples iteraciones. Normalmente, los analistas de datos ejecutan varios modelos utilizando los parámetros predeterminados y ajustan los parámetros o vuelven a la fase de preparación de datos para las manipulaciones necesarias del modelo [17].

4.2.1. Selección de hiper-parámetros por medio de experimentos.

En esta sección se describe la selección de los hiper-parámetros para los modelos de *clustering* descritos, por medio de experimentos y validados a través del método del Codo (*Elbow Method*), el Coeficiente de la Silueta (*Silhouette Coefficient*) y para el modelo DBSCAN el algoritmo k vecinos más cercanos.

El propósito de aplicar el método del Codo en el modelo k-means es elegir el número apropiado de clústeres para agrupar los datos y en el caso del Coeficiente de la Silueta se busca scores que estén cercanos a 1, para garantizar que los puntos estén muy cerca de su propio clúster y lejos de los otros clústeres y así optimizar la elección de los hiper-párametros.

Previamente, se crean 6 *datasets* de muestra, de cada uno de los *datasets* del enfoque de negocio, con el 10% de los registros, aleatorios, dada la función *random state*, y con reinicio del Índex.

a) Modelo K-means

En el modelo de K-means, se eligieron y variaron los siguientes hiper-parámetros:

- *n_clusters:* Validación de 3 a 40 clústeres.
- *init:* Los métodos de inicialización, con *K-means++* y *random.*
- *n_init:* El número de veces que se ejecutó el algoritmo con semillas de centroide diferentes, el cual fue de 10 para buscar mayor exactitud.

Dataset: muestraObsequios

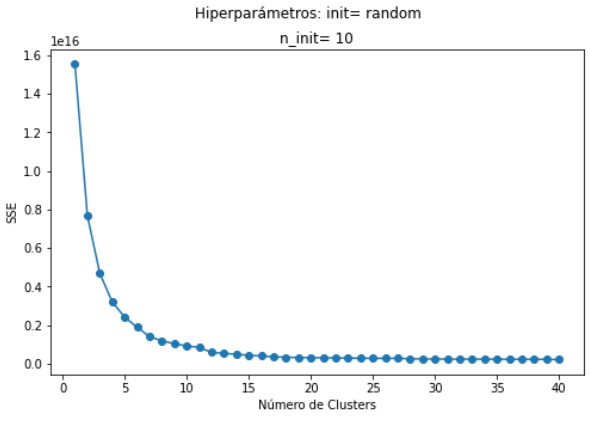


Imagen 4: *Elbow Method (random)* del dataset muestraObsequios

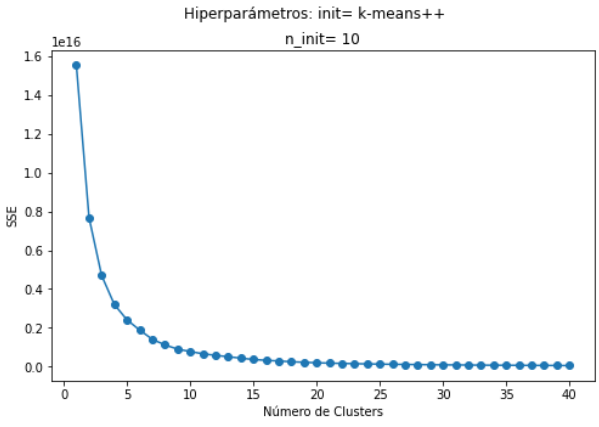


Imagen 5: *Elbow Method (k-means++)* del dataset muestraObsequios

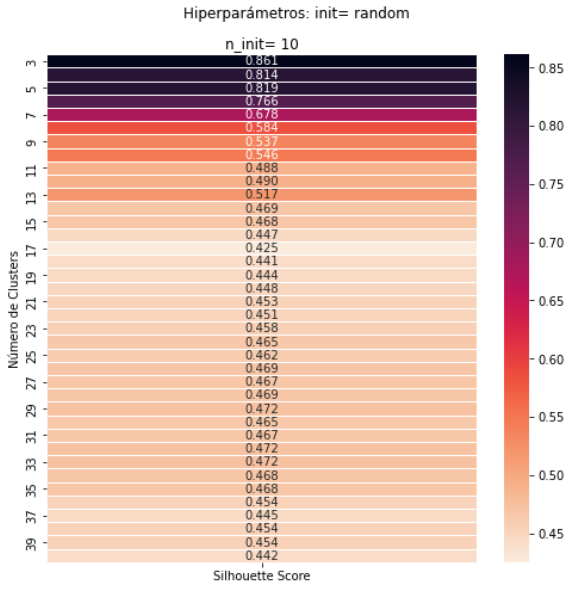


Imagen 6: *Silhouette score (random)* del dataset muestraObsequios

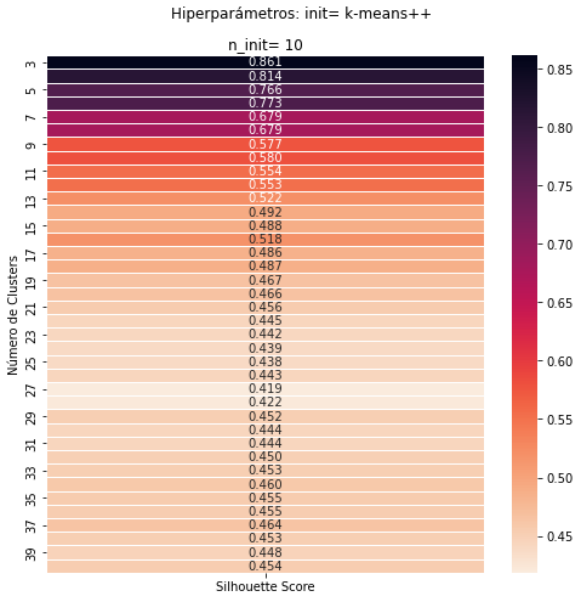


Imagen 7: *Silhouette score (k-means++)* del dataset muestraObsequios

En este experimento se observa que para el *dataset* muestraObsequios, en las imágenes 4 y 5 el método del codo indica que el número ideal de clústeres son 3 para los dos modelos y en las Imágenes 6 y 7 el coeficiente de siluete muestran un resultado *score* de 0.861 para el mismo número de clústeres 3, con una cercanía a 1, garantizando que los puntos estén muy cerca de su propio clúster y lejos de los otros clústeres, por lo que el hiper-parámetro *n_clusters* es 3 y en los métodos de inicialización *init* se puede aplicar los métodos dos para este *dataset*.

Dataset: muestraTipoCliente

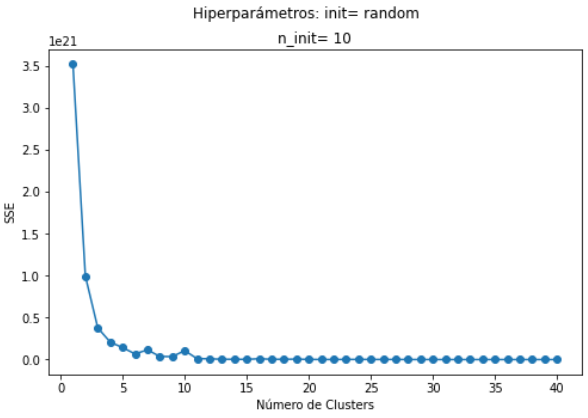


Imagen 8: *Elbow Method (random)* del dataset muestraTipoCliente

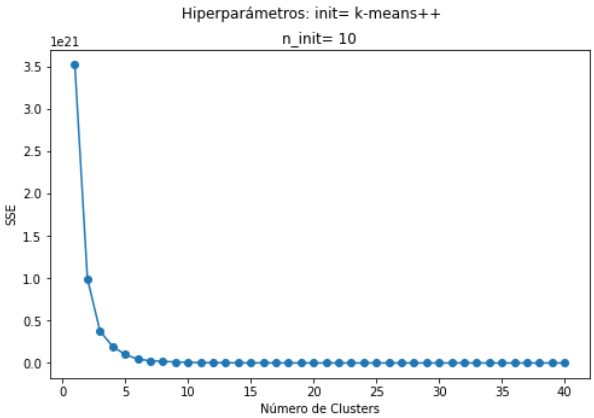


Imagen 9: *Elbow Method (k-means++)* del dataset muestraTipoCliente

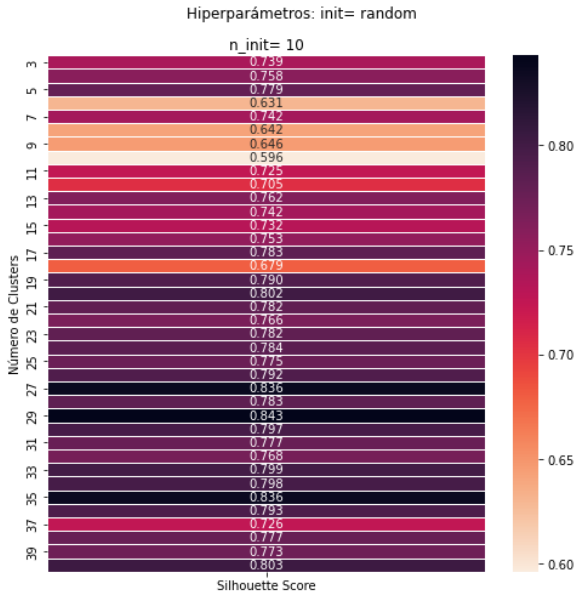


Imagen 10: *Silhouette score (random)* del dataset muestraTipoCliente

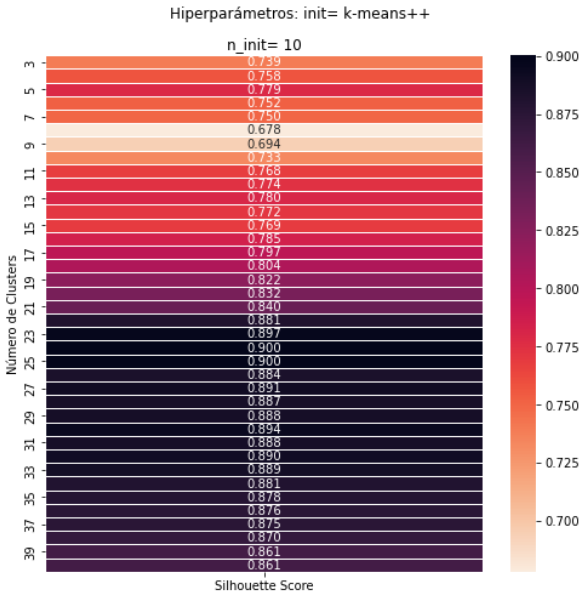


Imagen 11: *Silhouette score (k-means++)* del dataset muestraTipoCliente

En este *dataset* en las Imágenes 8 y 9, el método del codo indica que para los dos modelos el número ideal de clústeres son 3 y en la Imagen 10 el coeficiente de la silueta se observa que con *random* tiene varios sobre picos con mayor cercanía a 1 en el clúster 29 con 0,843 de score y en la Imagen 11, al contrario, con *k-means++*, muestra una mejor tendencia con un *score* más alto e ideal de 0.900 en el clúster 24, garantizando que los puntos estén muy cerca de su propio clúster y lejos de los otros clústeres, por lo que el hiper-parámetro *n_clusters* se van a evaluar en 3 y 24. Y en los métodos de inicialización *init* se aplicará el método de k-means++ para este *dataset*.

Dataset: muestraDemográfico

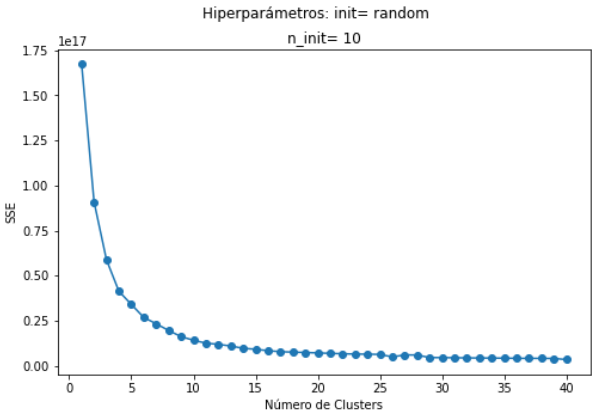


Imagen 12: *Elbow Method (random)* del dataset muestraDemográfico

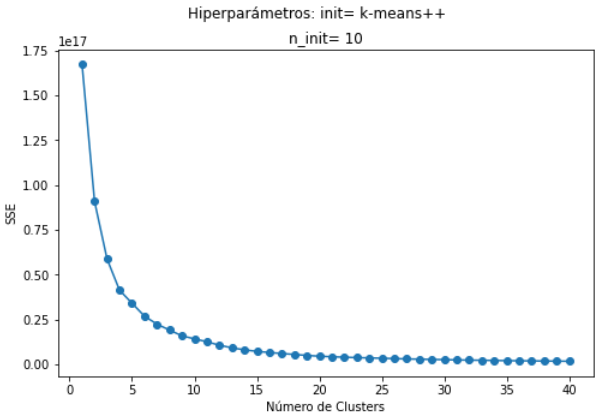


Imagen 13: *Elbow Method (k-means++)* del dataset muestraDemográfico

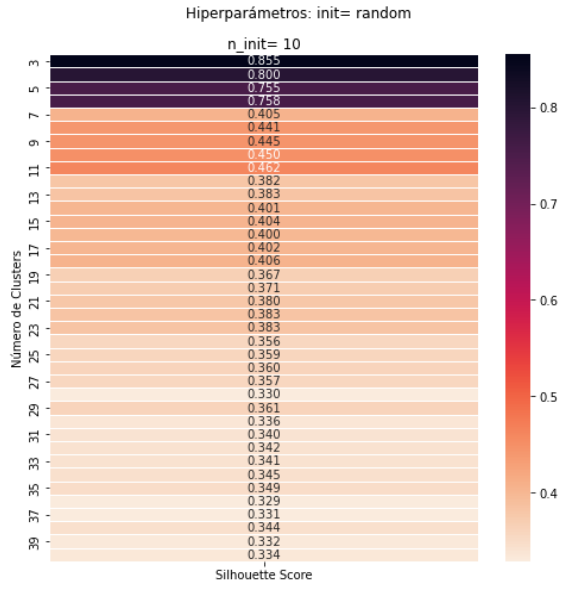


Imagen 14: *Silhouette score (random)* del dataset muestraDemográfico

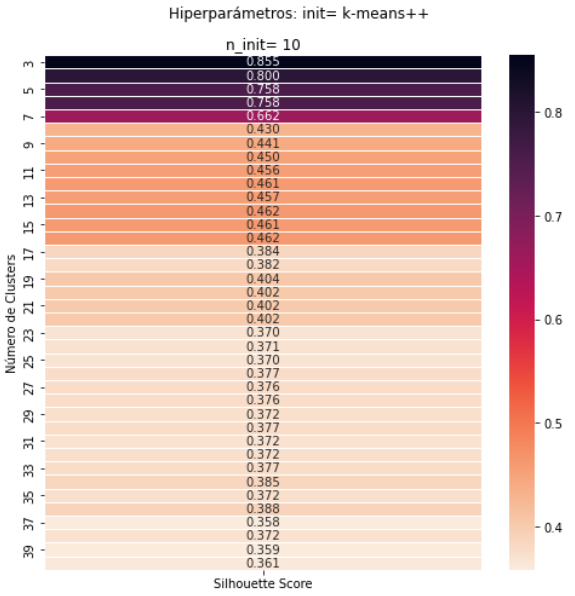


Imagen 15: *Silhouette score (k-means++)* del dataset muestraDemográfico

En este experimento se observa en las Imágenes 12 y 13 una similitud como los resultados del *dataset*, en donde el método del codo indica un número ideal de clústeres en 3 y en las Imágenes 14 y 15 el coeficiente de silueta muestran la misma similitud en los resultados con un *score* de 0.855, con una cercanía a 1, garantizando que los puntos estén muy cerca de su propio clúster y lejos de los otros clústeres, por lo que el hiper-parámetro *n_clusters* es 3 y en los métodos de inicialización *init* se puede aplicar los métodos dos para este *dataset*.

Dataset: muestraCosméticos

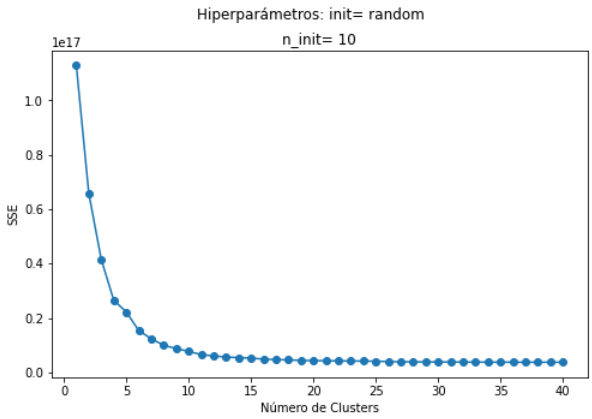


Imagen 16: *Elbow Method (random)* del dataset muestraCosméticos

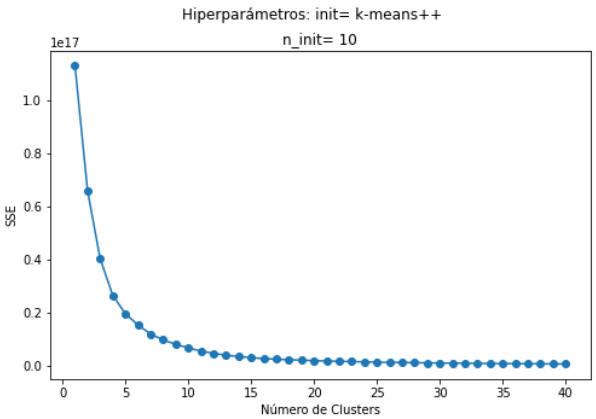


Imagen 17: *Elbow Method (k-means++)* del dataset muestraCosméticos

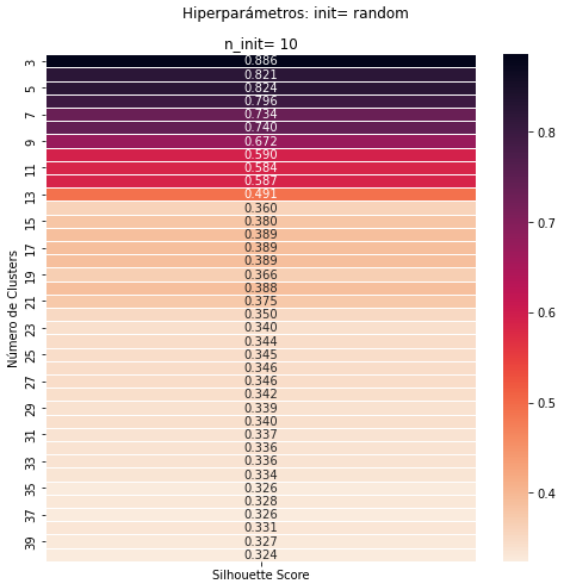


Imagen 18: *Silhouette score* (*random*) del dataset muestraCosméticos

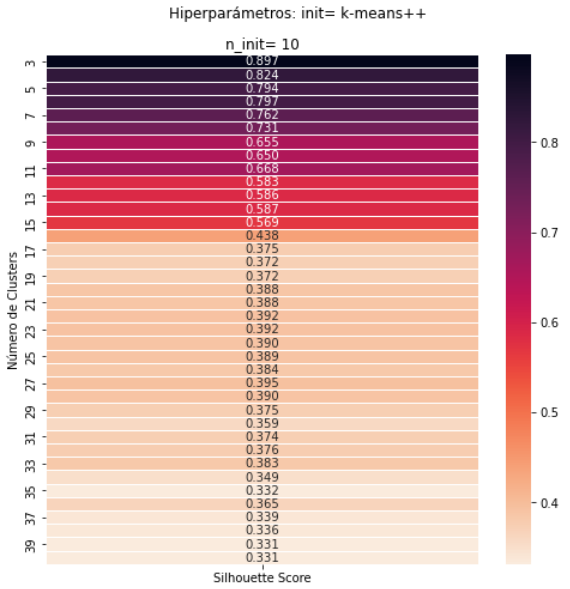


Imagen 19: *Silhouette score* (*k-means++*) del dataset muestraCosméticos

Dada la similitud en las Imágenes 16 y 17 en los resultados de este *dataset* con el método del codo, para un número ideal de clúster en 3 y aunque con una variación en el coeficiente de silueta en el mismo número de clúster 3, pero en la Imagen 19 el más cercano a 1 e ideal es *k-means++* con de 0.897, que garantiza que los puntos estén muy cerca de su propio clúster y lejos de los otros clústeres, dejando como hiperparámetro *n_clusters* es 3 y en los métodos de inicialización *init* k-means++.

Dataset: muestraMagistrales

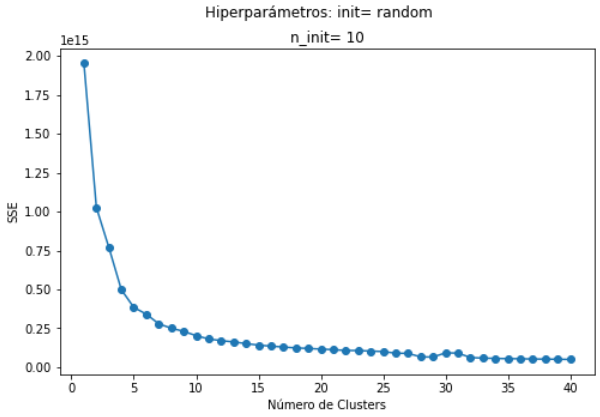


Imagen 20: *Elbow Method* (*random*) del dataset muestraMagistrales

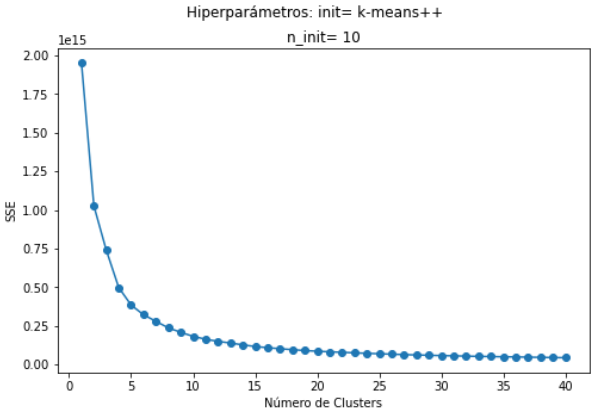


Imagen 21: *Elbow Method* (*k-means++)* del dataset muestraMagistrales

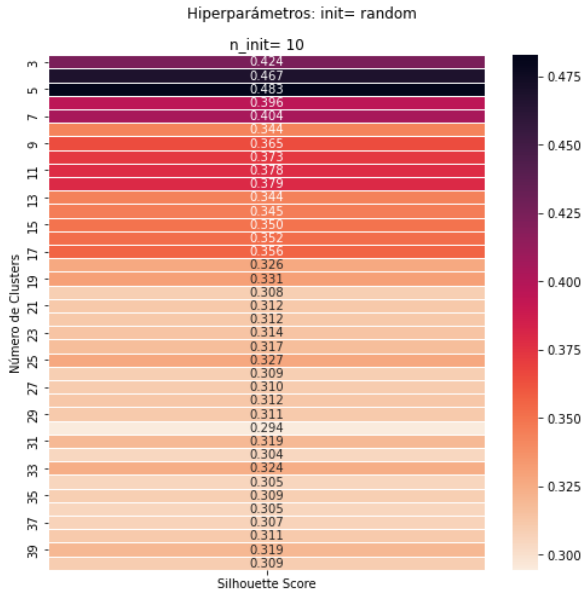


Imagen 22: *Silhouette score* (*random*) del dataset muestraMagistrales

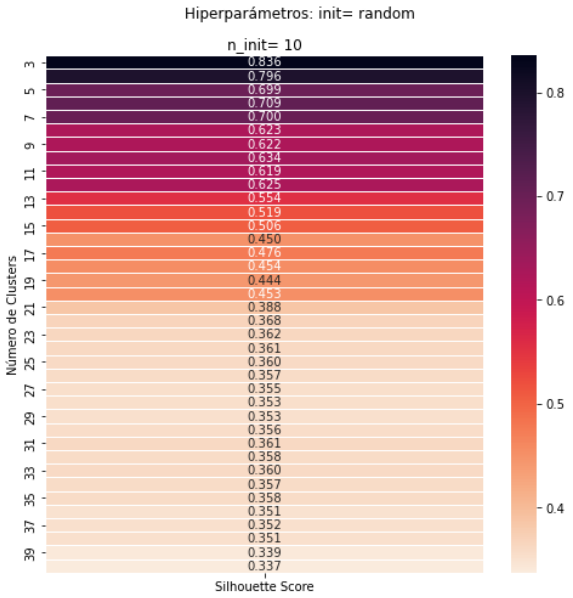


Imagen 23: *Silhouette score* (*k-means++*) del dataset muestraMagistrales

En relación con este experimento se observa en las Imágenes 20 y 21 que, en el método del codo en los dos modelos se determina un número e ideal de clúster en 5, aunque en la Imagen 22 con *random* en el coeficiente de silueta un menor *score* con 0.483 en mismo número de clústeres, a diferencia de la Imagen 23 el modelo con *k-means++* con un mejor score de 0.836 para 3 clústeres, lo que sugiere para este *dataset* los hiper-parámetros en *n_clusters* de 3 y 5. Y en el método de inicialización *init* k-means++.

Dataset: muestraMedicamentos

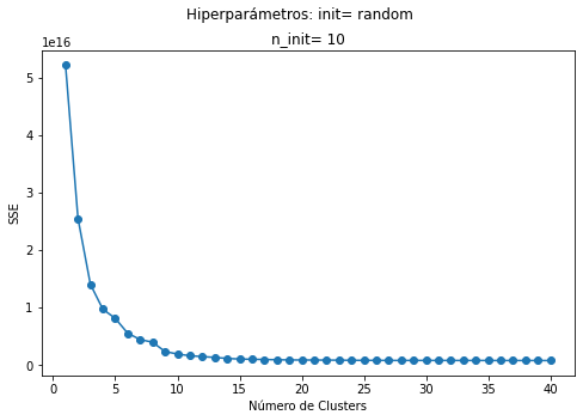


Imagen 24: *Elbow Method* (*random*) del dataset muestraMedicamentos

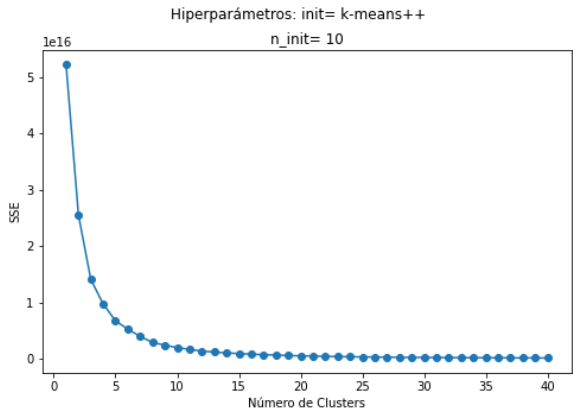


Imagen 25: *Elbow Method* (*k-means++* del dataset muestraMedicamentos)

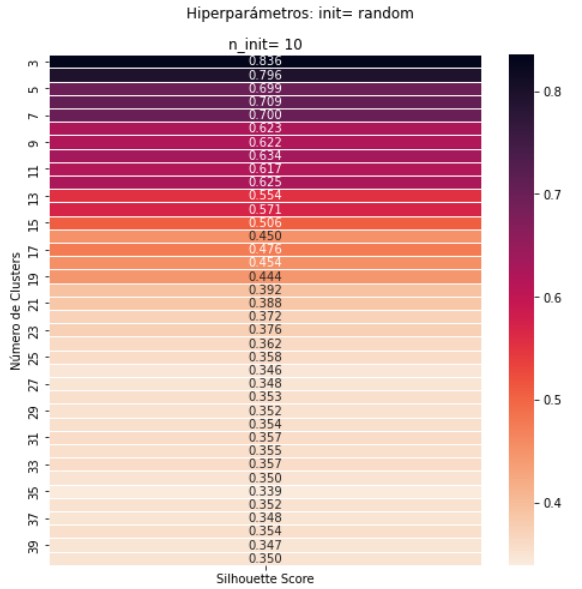


Imagen 26: *Silhouette score (random)* del dataset muestraMedicamentos

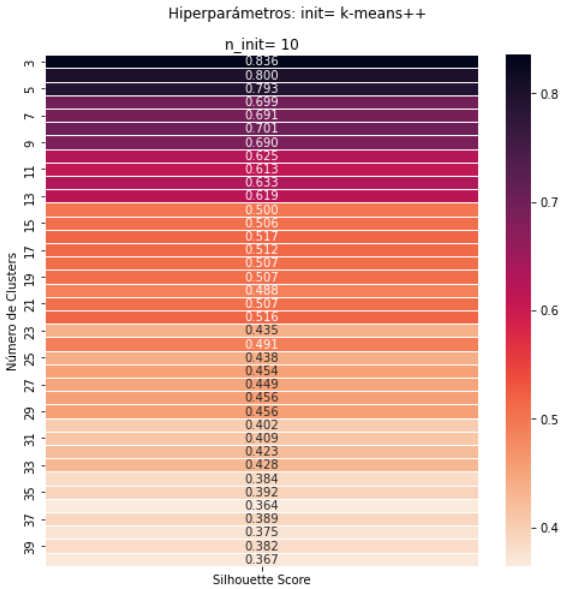


Imagen 27: *Silhouette score (k-means++)* del dataset muestraMedicamentos

Realizado el experimento, en las Imágenes 24 y 25 el método del codo sugiere que es 3 el número ideal de clústeres y en las Imágenes 26 y 27 el coeficiente de siluete muestran una similitud en los resultados con el mismo *score* de 0.836, indicando que en los hiper-parámetros, *n_clusters* es 3 y en el método de inicialización *init* tienen el mismo impacto los dos métodos.

— Resultados:

Tabla 6: *silhouette score (random vs k-means++)*

Datasets	**Hiper-parámetros**		
	n_clusters	*random*	*k-means++*
muestraObsequios	3	**0.861**	**0.861**
muestraTipoCliente	24	0.784	**0.900**
muestraDemográfico	3	**0.855**	**0.855**
muestraCosméticos	3	0.886	**0.897**
muestraMagistrales	3	0.424	**0.836**
muestraMedicamentos	3	**0.836**	**0.836**

Aplicado el método del codo, se destaca una considerable similitud entre las variaciones de los hiper-parámetros y los *datasets* evaluados. En relación con el coeficiente de la silueta en la Tabla N. 6, se resumen los mejores resultados, cercanos a 1, que indica que los puntos están muy cerca de su propio clúster y lejos de los otros clústeres, destacando el Hiper-parámetro de K-means++ con mejores puntuaciones.

b) Modelo DBSCAN (Density-based Spatial Clustering of Applications with Noise)

El modelo DBSCAN se seleccionarán los hiper-parámetros con el algoritmo k vecinos más cercanos (*k-nearest neighbors algorithm*), variado los parámetros del algoritmo para cada *dataset, de la siguiente manera*:

- n_neighbors: Número de vecinos a usar 500, 1.000, 3.000, 8.000, 12.000, 15.000 y 20.000, de acuerdo con el tamaño del *dataset*.
- metrics: Las métricas de distancia utilizadas son, euclidiana, manhattan y minkowski.
- p: El parámetro para la métrica fue de 2, 3, 4 y 5.

Dataset: muestraObsequios

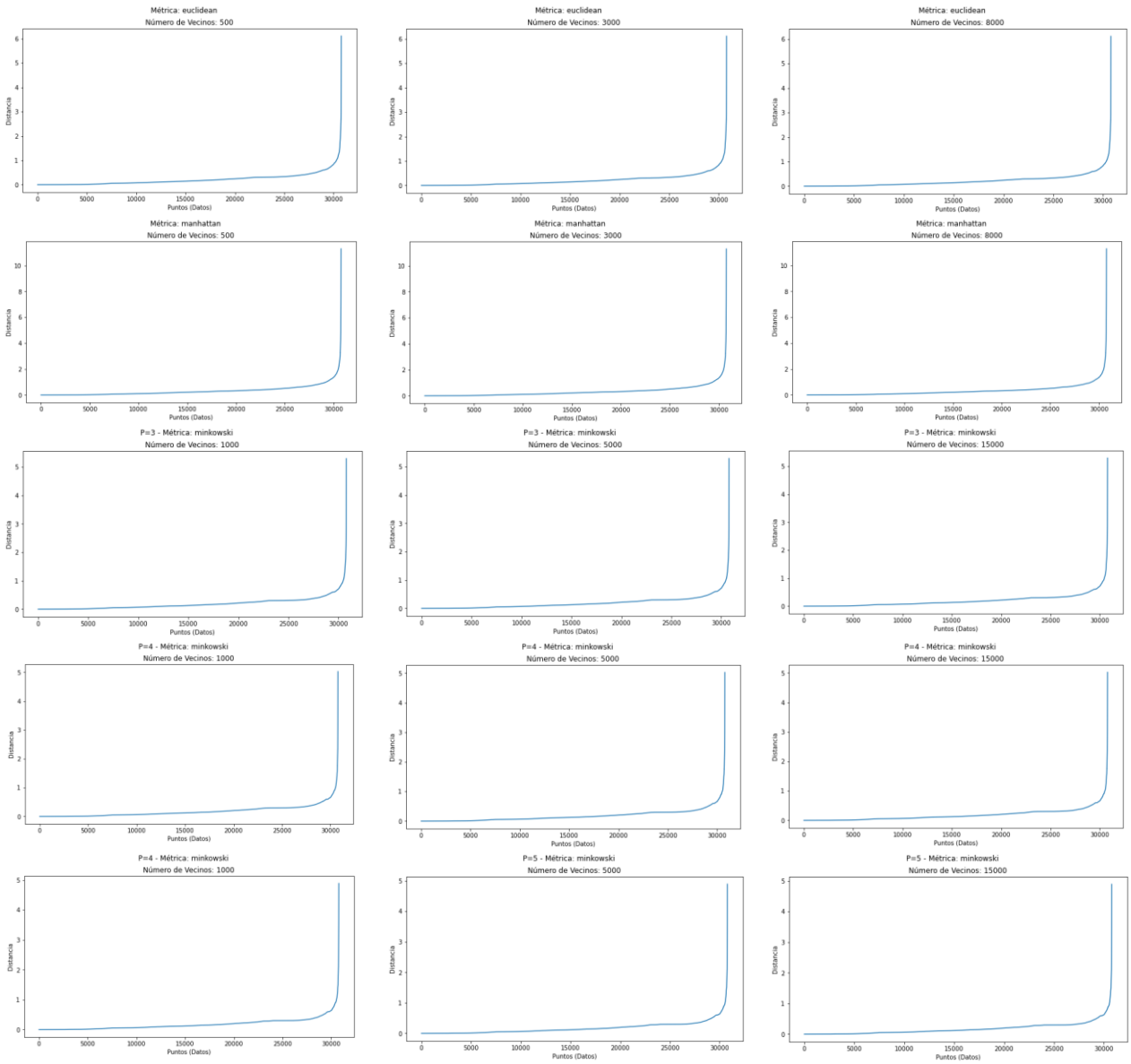


Imagen 28: Diagrama *k nearest neighbor* del dataset muestraObsequios

En este *dataset* en la Imagen 28, con una de muestra de 30.871 registros, se analizaron la distancia euclidiana y manhattan con (*p*) igual a 2, y con (*n_neighbors*) igual 500, 3.000 y 8.000, en el caso de la distancia minkowski con (*p*) igual a 3, 4 y 5, con (*n_neighbors*) igual 1.000, 5.000 y 15.000.

Dataset: muestraTipoCliente

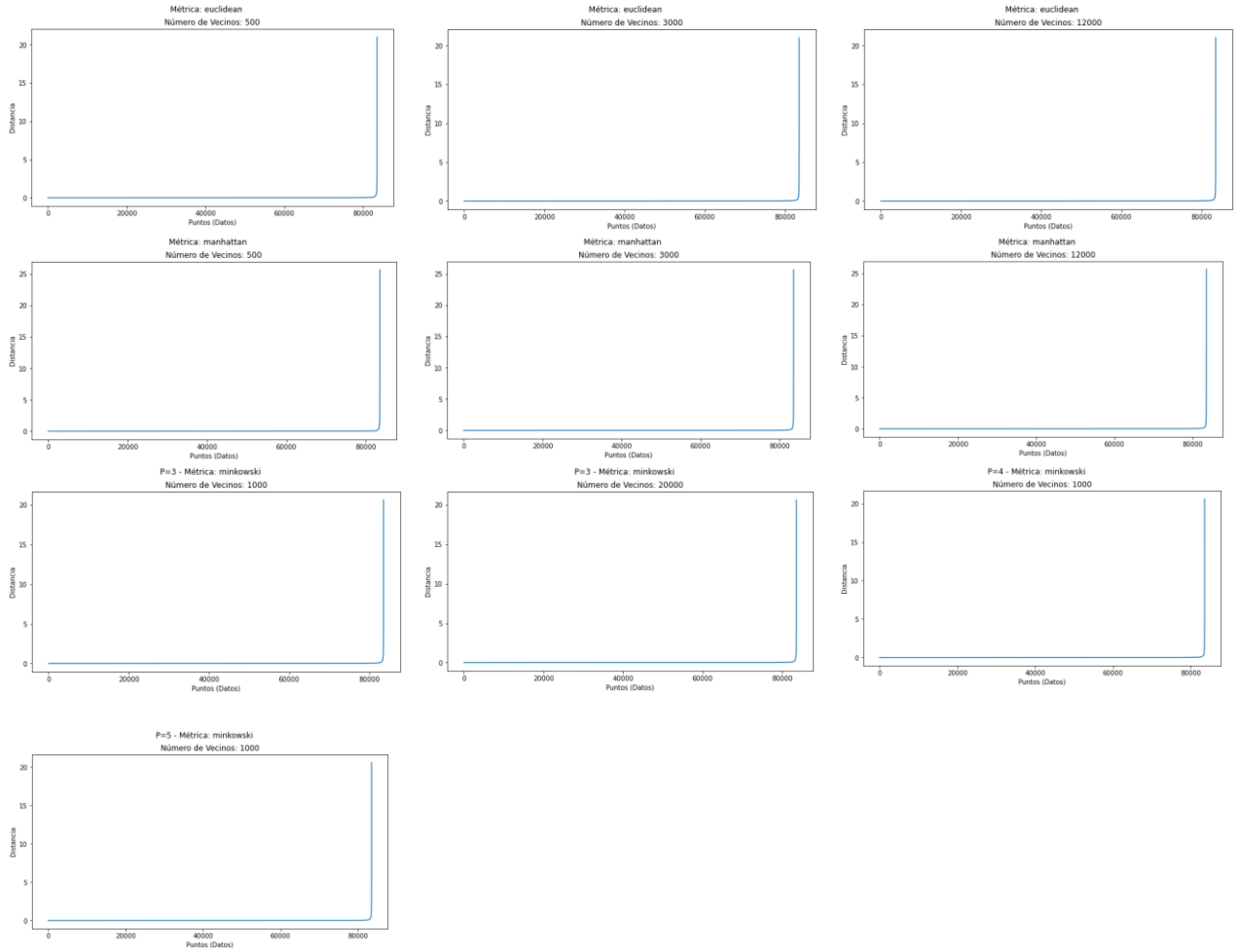


Imagen 29: Diagrama *k nearest neighbor* del dataset muestraTipoCliente

En este *dataset* en la Imagen 29, con una de muestra de 83.541 registros, se analizaron la distancia euclidiana y manhattan con (*p*) igual a 2, y con (*n_neighbors*) igual 500, 3.000 y 12.000, en el caso de la distancia minkowski con (*p*) igual a 3, 4 y 5, con (*n_neighbors*) igual 1.000 y 20.000.

Dataset: muestraDemográfico

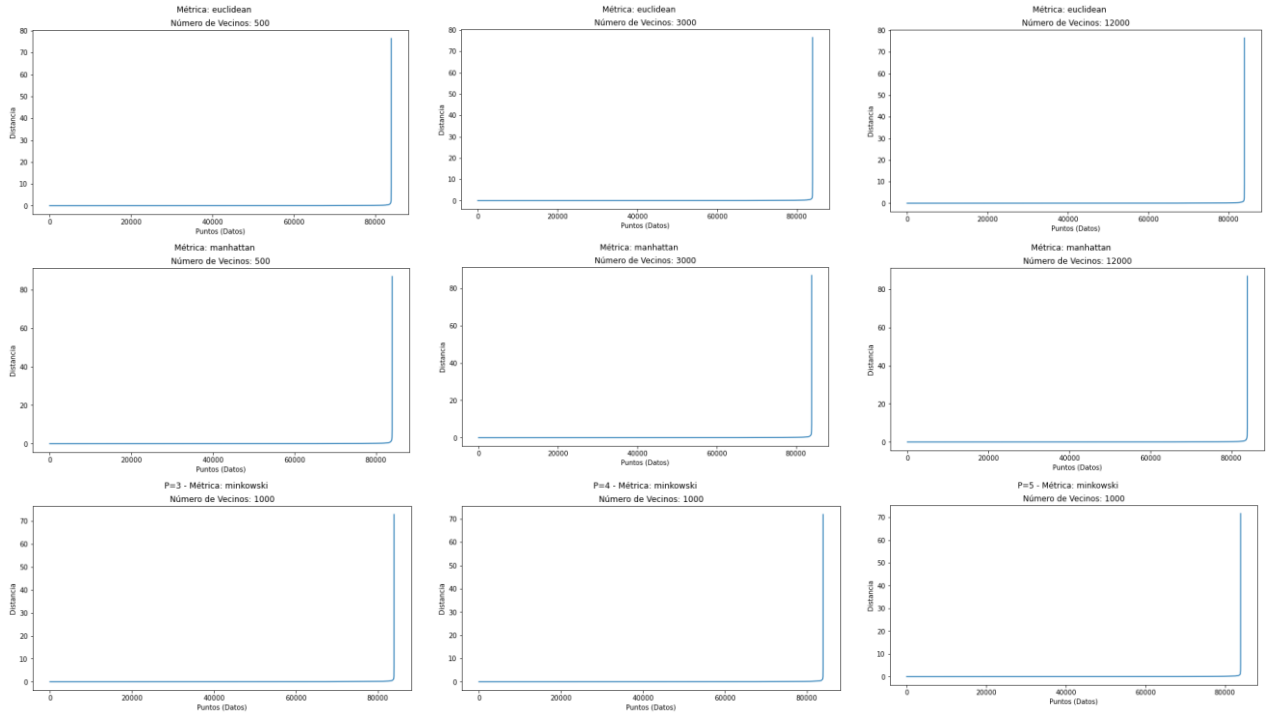
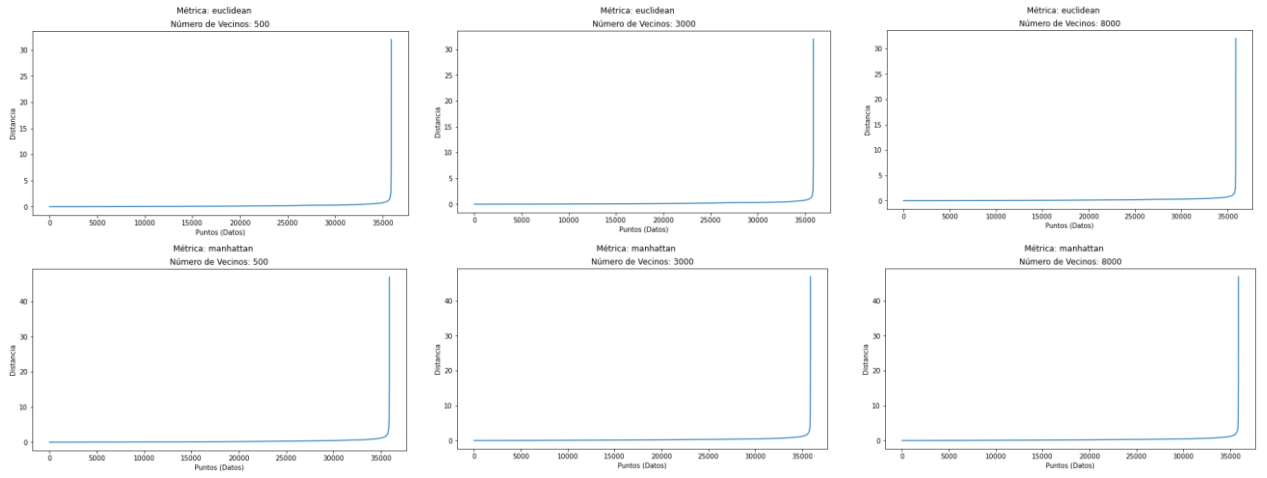


Imagen 30: Diagrama *k nearest neighbor* del dataset muestraDemográfico

En este *dataset* en la Imagen 30, con una de muestra de 83.949 registros, se analizaron la distancia euclidiana y manhattan con (*p*) igual a 2, y con (*n_neighbors*) igual 500, 3.000 y 12.000, en el caso de la distancia minkowski con (*p*) igual a 3, 4 y 5, con (*n_neighbors*) igual 1.000.

Dataset: muestraCosméticos



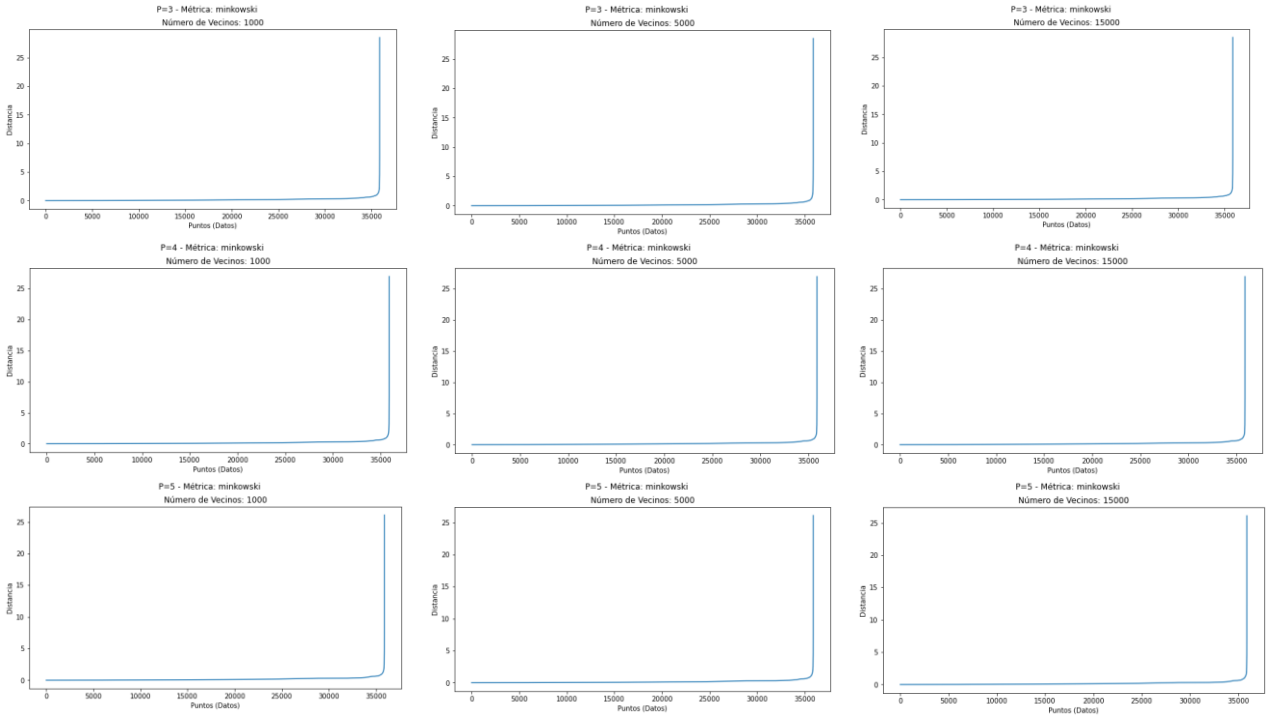
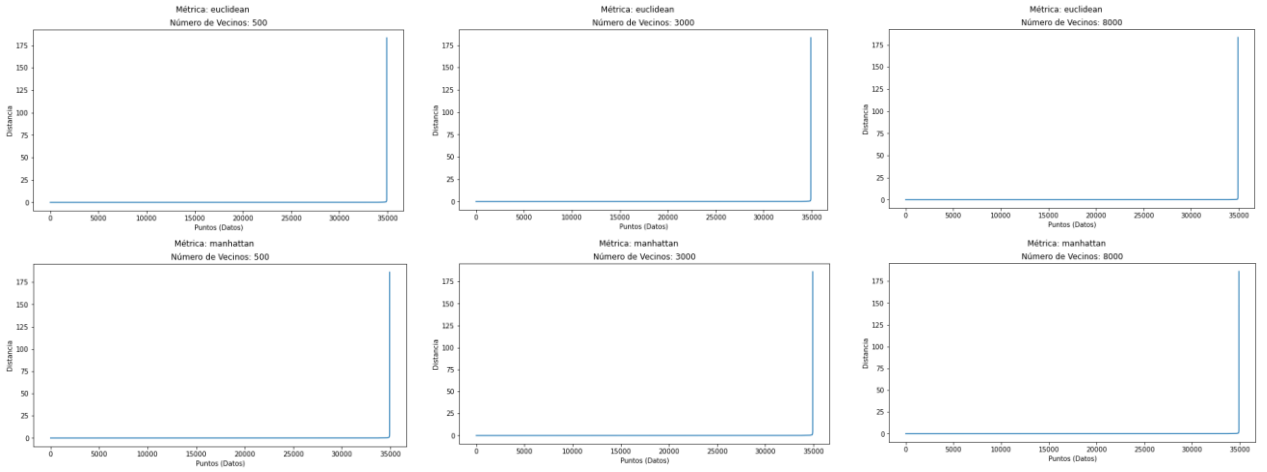


Imagen 31: Diagrama *k nearest neighbor* del dataset muestraCosméticos

En este *dataset* en la Imagen 31, con una de muestra de 35.900 registros, se analizaron la distancia euclidiana y manhattan con (*p*) igual a 2, y con (*n_neighbors*) igual 500, 3.000 y 8.000, en el caso de la distancia minkowski con (*p*) igual a 3, 4 y 5, con (*n_neighbors*) igual 1.000, 5.000 y 15.000.

Dataset: muestraMagistrales



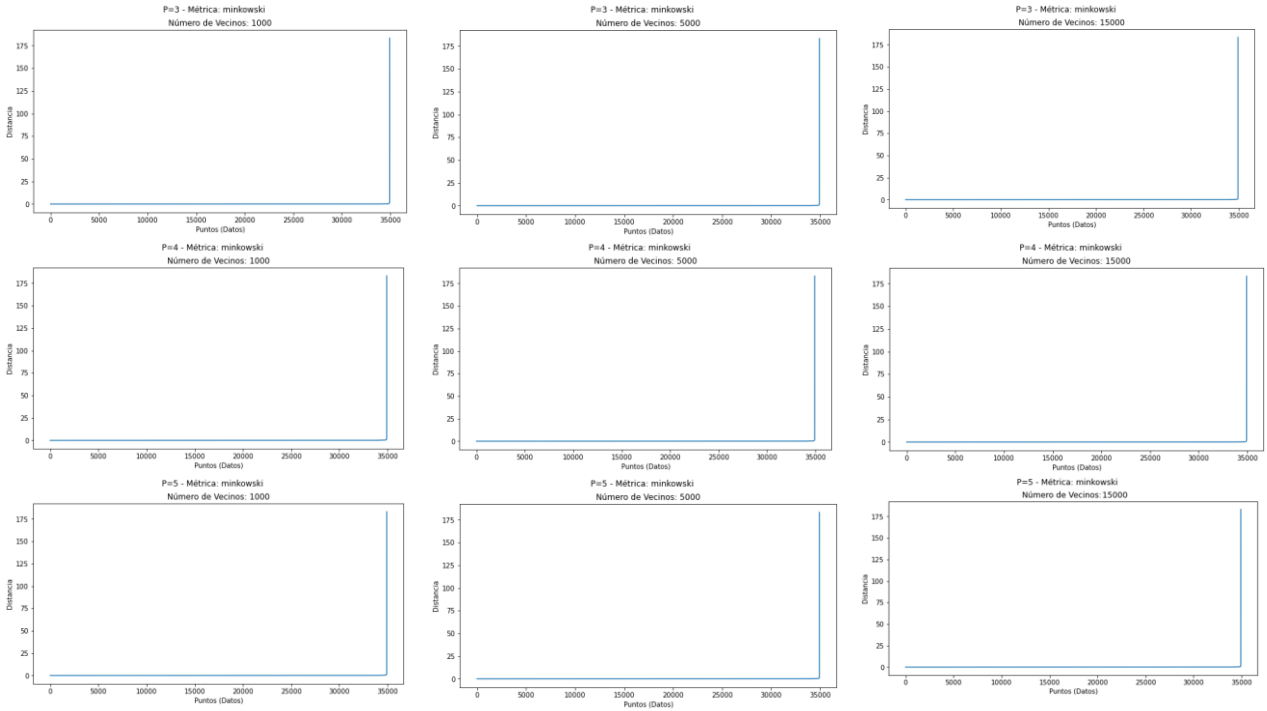
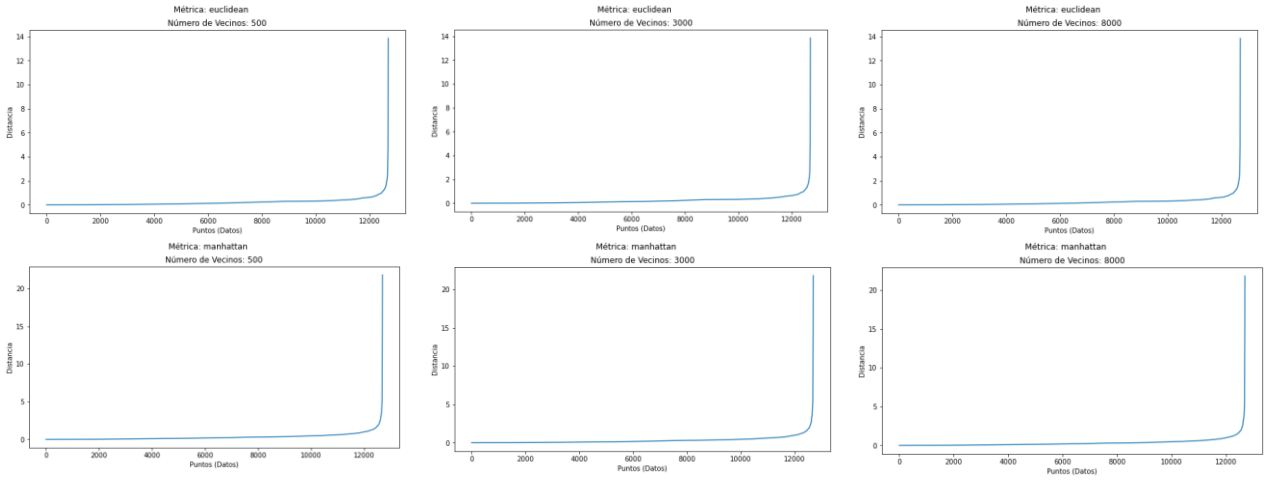


Imagen 32: Diagrama *k nearest neighbor* del dataset muestraMagistrales

En este *dataset* en la Imagen 32, con una de muestra de 34.929 registros, se analizaron la distancia euclidiana y manhattan con (*p*) igual a 2, y con (*n_neighbors*) igual 500, 3.000 y 8.000, en el caso de la distancia minkowski con (*p*) igual a 3, 4 y 5, con (*n_neighbors*) igual 1.000, 5.000 y 15.000.

Dataset: muestraMedicamentos



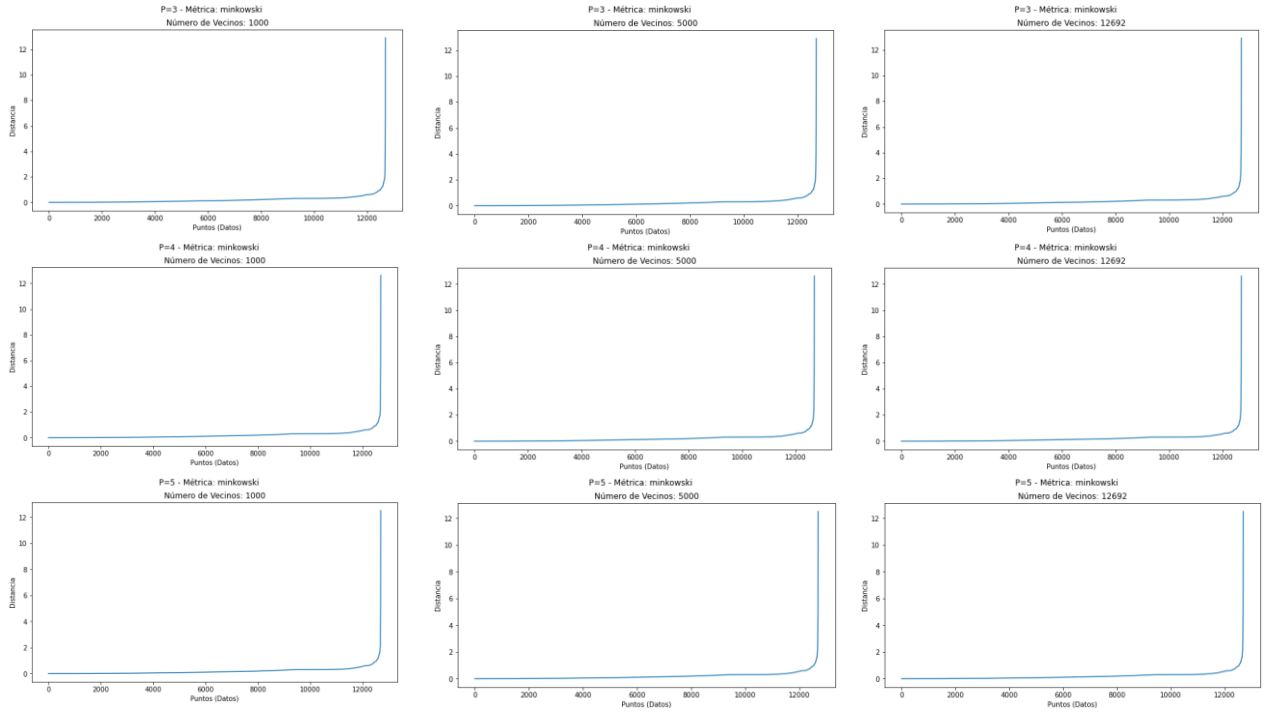


Imagen 33: Diagrama *k nearest neighbor* del dataset muestraMedicamento

En este *dataset* en la Imagen 33, con una de muestra de 12.692 registros, se analizaron la distancia euclidiana y manhattan con (*p*) igual a 2, y con (*n_neighbors*) igual 500, 3.000 y 8.000, en el caso de la distancia minkowski con (*p*) igual a 3, 4 y 5, con (*n_neighbors*) igual 1.000, 5.000 y 12.692.

— Resultados:

Finalizado las verificaciones de las búsquedas de vecinos, con todas las variaciones en el algoritmo, como, número de vecinos, métricas de distancia y los parámetros para la métrica, se observa que el modelo DBSCAN la densidad es extremadamente alta, para crear clusteres diferenciables, en donde por ejemplo en el diagrama de *k nearest neighbor* del dataset muestraMedicamento se observa que aunque se variaron las distancias euclidianas manhattan y minkowski, los parametros para la métrica y el número de vecinos se obtuvo un solo clúster.
Por lo que este modelo queda desestimado, para continuar con las ejecuciones de entrenamiento en los datasets completos.

c) Jerárquico (Hierarchical clustering).

En el modelo Jerárquico es utilizaron los hiper-parámetros:

— n_clusters: Validación de 3 a 30 clústeres.
— affinity: Métrica utilizada, *euclidean* y *manhattan*
— Linkage: Criterio de vinculación es *ward* para *euclidean,* la cual solo se acepta *euclidean* y *average* para *manhattan*.

Datasets: muestraObsequios

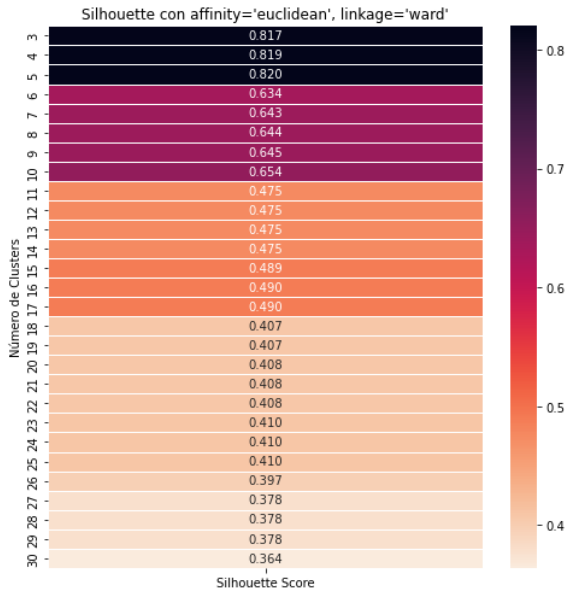


Imagen 34: *Silhouette score* (*euclidean*)

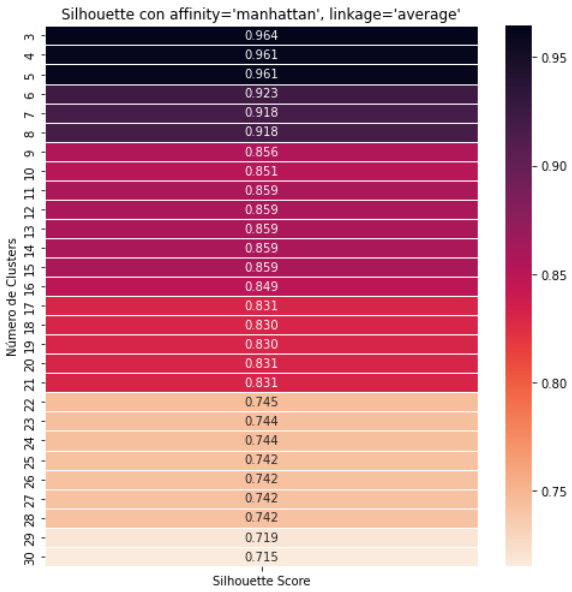


Imagen 35: *Silhouette score* (*manhattan*)

Dada la similitud en los resultados en las Imágenes 32 y 35, de este *dataset* en 3, 4 y 5 clústeres, el más cercano a 1 e ideal, es *manhattan* con *average* con un *silhouette score* de 0.964 para el clúster 3.

Datasets: muestraTipoCliente

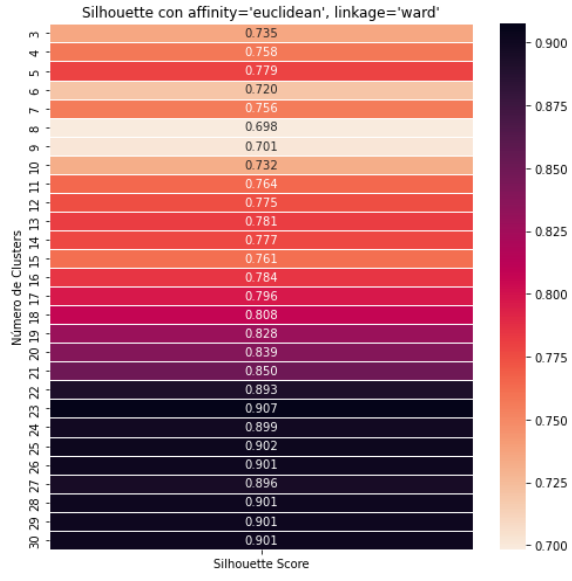


Imagen 36: *Silhouette score* (*euclidean*)

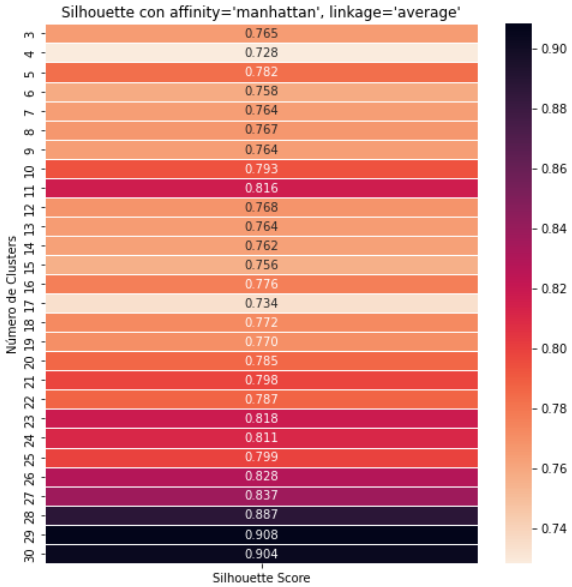


Imagen 37: *Silhouette score* (*manhattan*)

En este *dataset* se observan en las Imágenes 36 y 37 puntajes más altos con mayores clústeres, y el mejor resultado de este *dataset* está en el 29 clúster, el más cercano a 1 e ideal, con *manhattan* con *average* con un *silhouette score* de 0.908, para el clúster 29.

Dataset: muestraDemográfico

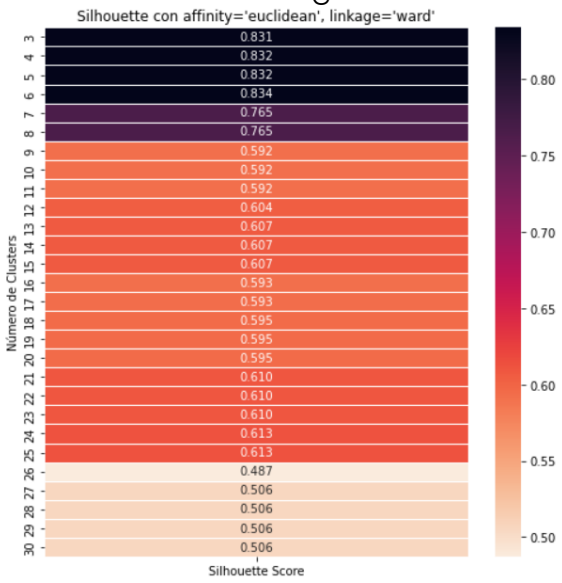


Imagen 38: *Silhouette score* (*euclidean*)

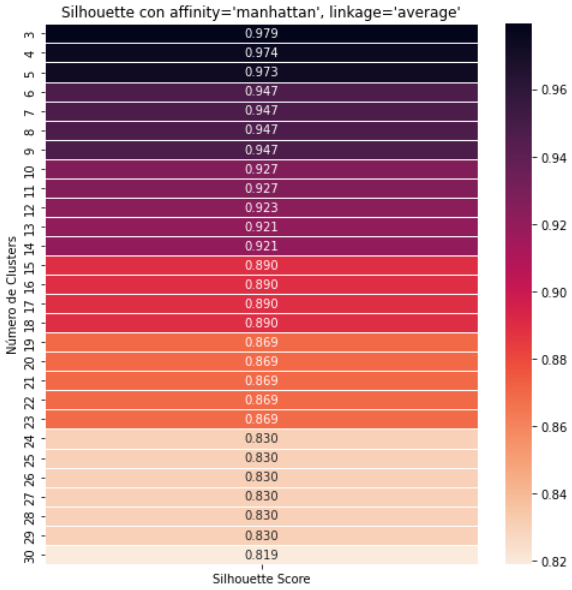


Imagen 39: *Silhouette score* (*manhattan*)

Dada la similitud en los resultados en las Imágenes 38 y 39 de este *dataset* en 3 clústeres, el más cercano a 1 e ideal, es *manhattan* con *average* con un *silhouette score* de 0.979, para el clúster 3.

Dataset: muestraCosméticos

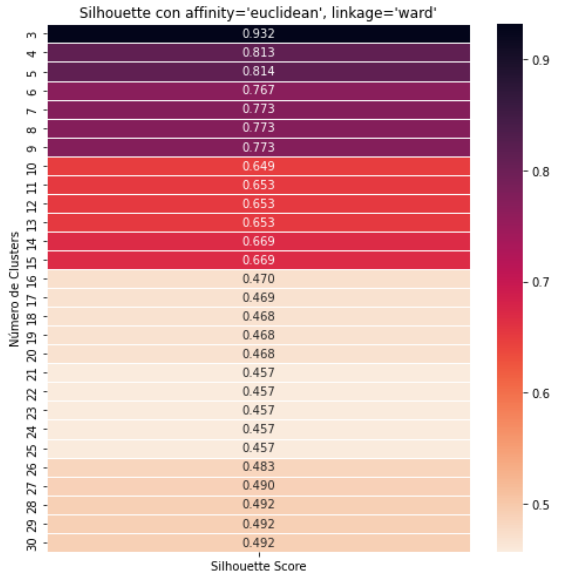


Imagen 40: *Silhouette score* (*euclidean*)

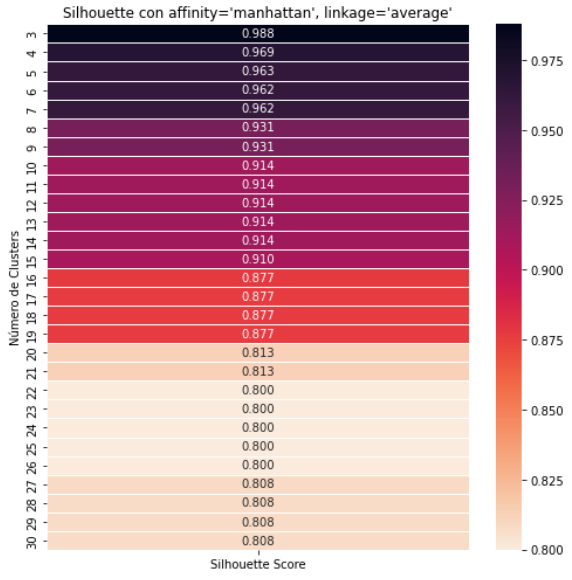


Imagen 41: *Silhouette score* (*manhattan*)

Dada la similitud en los resultados en las Imágenes 40 y 41 de este *dataset* en el tercer clúster, el más cercano a 1 e ideal, es *manhattan* con *average* con un *silhouette score* de 0.988, para el clúster 3.

Dataset: muestraMagistrales

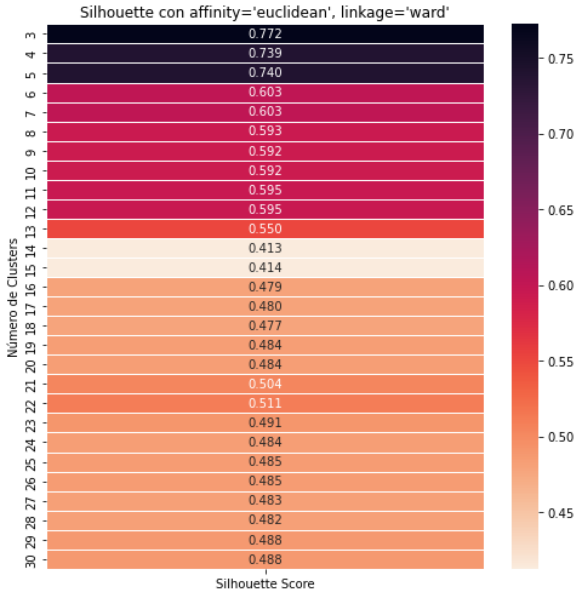


Imagen 42: *Silhouette score* (*euclidean*)

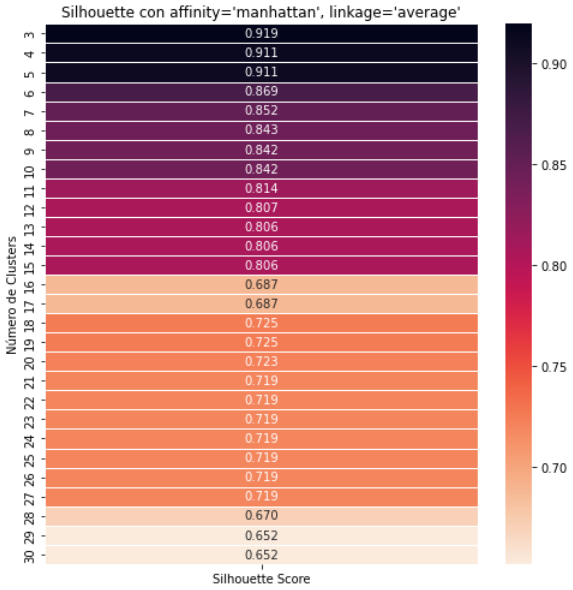


Imagen 43: *Silhouette score* (*manhattan*)

Dada la similitud en los resultados en las Imágenes 42 y 43 de este *dataset* en el tercer clúster, el más cercano a 1 e ideal, es *manhattan* con *average* con un *silhouette score* de 0.919, para el clúster 3.

Dataset: muestraMedicamentos

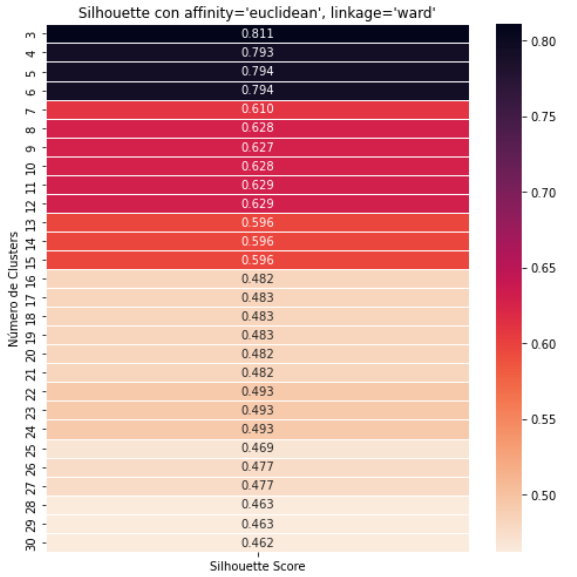


Imagen 44: *Silhouette score* (*euclidean*)

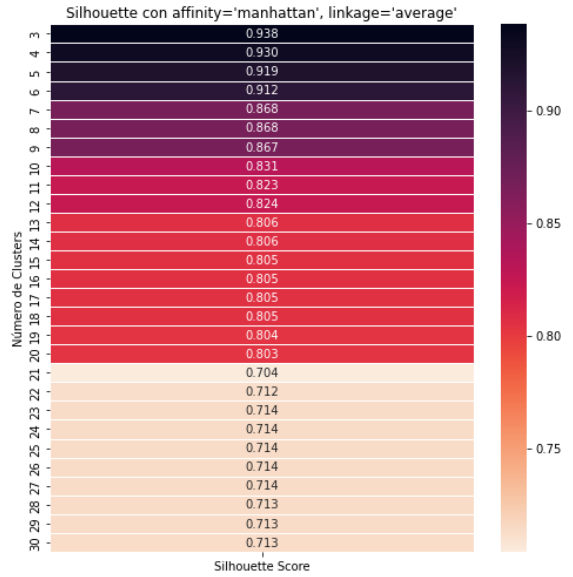


Imagen 45: *Silhouette score* (*manhattan*)

Dada la similitud en los resultados en las Imágenes 44 y 45 de este *dataset* en el tercer clúster, el más cercano a 1 e ideal, es *manhattan* con *average* con un *silhouette score* de 0.938, para el clúster 3.

— Resultados:

Tabla 7: *silhouette score (euclidean con ward vs manhattan con average)*

Datasets	**Hiper-parámetros**		
	n_clusters	*euclidean - ward*	*manhattan - average*
muestraObsequios	3	0.817	**0.864**
muestraTipoCliente	29	0.901	**0.908**
muestraDemográfico	3	0.831	**0.979**
muestraCosméticos	3	0.932	**0.988**
muestraMagistrales	3	0.772	**0.919**
muestraMedicamentos	3	0.811	**0.938**

En relación con el coeficiente de la silueta en la Tabla N. 7, se resumen los mejores resultados, cercanos a 1, que indica que los puntos están muy cerca de su propio clúster y lejos de los otros clústeres, destacando el hiper-parámetro *manhattan* con el criterio de vinculación *average*.

4.2.2. Entrenamiento de los modelos K-means y Hierarchical.

a. *Datasets* para el entrenamiento.

Los *datasets* definidos para el entrenamiento se efectuarán con dos versiones de *datasets*, uno con el 100% de los registros y otro con un porcentaje dado de los *datasets* dadas las capacidades de cómputo con las que se dispone, que se explicaran en el punto 4.2.3. (Impacto por las dificultades presentadas).

Tabla 8: *Datasets para el entrenamiento*

Nombre	**Tipo 1**		**Tipo 2**	
	% datos	**Registros**	**% datos**	**Registros**
dataObsequios	100%	307.807	41%	126.201
dataTipoCliente	100%	835.412	15%	125.312
dataDemográfico	100%	839.487	15%	125.923
dataCosméticos	100%	359.003	35%	125.651
dataMagistrales	100%	349.289	36%	125.744
dataMedicamentos	100%	126.918	-	-

b. Hiper-parámetros

Seleccionados de los resultados de los experimentos realizados, los hiper-parámetros para los dos modelos a entrenar, serán los descritos a continuación:

Modelo K-means:

- n_clusters: Número de clústeres; entre 2 a 4 clústeres, basado en el *silhouette score* de los experimentos.
- init: Método de inicialización; *K-means++*.
- n_init: Número de veces que se ejecutara el algoritmo; *10*.

Modelo Jerárquico:

- n_clusters: Numero de clústeres; entre 2 a 4 clústeres, basado en el *silhouette score* de los experimentos.
- affinity: Métrica; *manhattan*
- Linkage: Criterio de vinculación; *average*

4.2.3. Dificultades presentadas y estrategias implementadas

El modelo Jerárquico (Hierarchical clustering) presentó dificultades de rendimiento deficiente en los *datasets* con registros superiores a 127.000, en las capacidades de cómputo de memoria y rendimiento.

En la publicación *The (black) art of runtime evaluation: Are we comparing algorithms or implementations?,* examinan el rendimiento de diferentes algoritmos de *clustering*, y en el caso del algoritmo Jerárquico, tiende a quedarse sin memoria antes de las 100.000 instancias, para lo que necesitaría 38 GB de RAM, y para un millón de instancias se necesitarían 3,7 TB de RAM, en función de la implementación determinada [18].

Presentada esta dificultad, se aprovisionó una instancia con 16 vCPUs y 128 GB de RAM en Google Cloud Platform en modo de prueba, lo que permitía ejecutar modelos Jerárquicos con 127.000 registros máximos.

La primera solución fue dividir el *dataset* original en partes iguales o inferiores a 127.000 registros, que permitiera ejecutar el algoritmo a cada una de las partes del *dataset* subdividido, y posteriormente se realizó un análisis de reagrupamiento y se conformaron los clústeres finales.

La segunda solución fue crear *datasets* cercanos a los 127.000 registros, de forma aleatoria con la función *random state*, y reiniciando el Índex, descritos en la Tabla N.8 de tipo 2, y posteriormente ejecutar el algoritmo de conglomerado.

Cabe destacar que en el caso del *dataset* de Medicamentos, no fue necesario crear otro *dataset* debido a que contiene 126.918 registros y no presenta dificultades para el modelo jerárquico.

5. EVALUACIÓN Y SELECCIÓN DEL MODELO

Inicialmente, en esta sección se evaluará la calidad del agrupamiento de los modelos k-means y jerárquico, valorando la separación inter-grupos y la cohesión intra-grupos, para lo cual existen diversas métricas en la evaluación de *clustering,* como el coeficiente de la Silueta, con el que se seleccionaron los hiper-parámetros, el índice de Calinski y Harabasz, entre otros, y el índice de Davies-Bouldin, elegido y adecuado para realizar la evaluación cuantitativa del agrupamiento realizado por los algoritmos. Después, se llevará a cabo el análisis de los resultados de los modelos obtenidos a través de la visualización de gráficos de radar y finalmente, con el apoyo del departamento de mercadeo y ventas, se analizarán los resultados obtenidos.

5.1. Comparativo entre *K-means* y Jerárquico (*Hierarchical*).

En el comparativo entre el entrenamiento de los modelos K-means y el Jerárquico, se evaluó con el índice de Davies-Bouldin, generando los siguientes resultados, para los *datasets* descritos en el numeral 4.2.2. punto a. *Datasets* para el entrenamiento.

a. *Datasets* de Tipo 1, con el 100% de los registros.

Tabla 9: Comparativo resultados índices de Davies-Bouldin (*Datasets 100%)*

Dataset	Número de Clústeres	índice Davies-Bouldin	
		K-means	Jerárquico
Obsequios	3	0.5123821212846695	**0.5008800167913602**
Tipo de Cliente	23	**0.14286109922017623**	54.1595482603534
Demográfico	10	**0.49291602004140656**	6.183116635160513
Cosméticos	10	**0.48699949177820867**	2.093009038185543
Magistrales	3	**0.5210373538208933**	0.7423610224358916
Medicamentos	10	**0.3959949840944447**	0.4201378837101596

En los resultados de la Tabla 9, se observa que en los resultados de los índices de Davies-Bouldin de los *datasets* Tipo de Cliente y Demográfico con el modelo Jerárquico no estuvieron acotados entre el intervalo de 0 a 1, lo cual indica que no tienen una adecuada separación inter-grupos y cohesión intra-grupos. En referencia a los demás resultados, señala un mejor rendimiento con los modelos k-means, a excepción del dataset Obsequio que mostró un mejor resultado con el modelo jerárquico.

b. *Datasets* de Tipo 2, con porcentajes que se encuentran alrededor de los 127.000 registros.

Tabla 10: Comparativo resultados índices de Davies-Bouldin (*Datasets* 15%, 35%, 36% y 41%)

Dataset	Porcentaje	Número de Clústeres	índice Davies-Bouldin	
			K-means	Jerárquico
Obsequios	41%	3	**0.5050880241717788**	0.5737273649633 04
Tipo de Cliente	15%	23	**0.14191976731922185**	0.1517240844231968
Demográfico	15%	10	**0.47761694843589203**	0.48379800345681856
Cosméticos	35%	10	**0.4553914764915647**	0.47237467805455224
Magistrales	36%	3	**0.5166227865955737**	0.5700511384158627

Establecidos los resultados del índice Davies-Bouldin en la Tabla 10 sobre los modelos de los *datasets* con los porcentajes estipulados entre 15% al 41%, ratifica que los modelos *k-means* tienen mejor desempeño, con valores más cercanos a cero (0), por ejemplo en el resultado del *dataset* de Obsequios se obtuvo un valor de 0.5050880241717788 en el modelo k-means y en el modelo jerárquico un valor de 0.573727364963304, teniendo una mayor cercanía a cero (0) el modelo k-means, con una variación de 0,068639340791526 entre los modelos.

5.2. Análisis de resultados de los modelos obtenidos.

Con la finalidad de realizar el análisis de resultados de los clústeres, se propusieron varias visualizaciones de los modelos obtenidos con el gráfico de radar.

El gráfico del radar, también conocido como gráfico de araña o gráfico de estrella, muestra datos multivariados en forma de un gráfico bidimensional de variables cuantitativas representadas en ejes que se originan en el centro [19].

En los siguientes análisis se presentan y grafican los centroides de cada clúster de los *datasets,* en los modelos *k-means* y jerárquico con el 100% de los registros y otro con los porcentajes descritos anteriormente, con el propósito de analizar el comportamiento de los atributos en cada uno de los clústeres y definir similitudes, diferencias y comportamiento de los atributos en cada clúster.

5.2.1. *dataset* Obsequios

El *dataset* Obsequios comprende los registros de las ventas en el que se describen los productos bonificados (obsequios) y los descuentos otorgados a los clientes, según las unidades de productos compradas, entre mayor sea la compra mayor las unidades obsequiadas, según la política asignada al tipo de cliente. Entre los atributos más destacados se encuentran el valor de descuento asignado en la moneda pesos, las unidades obsequiadas y el descuento en porcentaje otorgado. Todo esto con la finalidad de examinar la afectación de los obsequios en los atributos demográficos, los meses de compra y en las categorías y marcas de los productos.

En la siguiente tabla se analizará los centroides del atributo Cantidades Obsequiadas contra los atributos más destacados:

Tabla 11: Centroides Cantidad Obsequios vs. atributos elegidos – modelo k-means 100%

Clústeres	Cant_Obsequio	Categoría	Marca	Ciudad	Mes_Compra
Clúster 0	-0,040149245	2,40874589	1,994920487	**0,449615521**	-0,098225175
Clúster 1	2,349874093	**0,188034014**	-0,367862402	**0,147004447**	2,320751168
Clúster 2	0,690275152	**0,403220095**	1,372941915	2,403380031	0,777474007

Comparado los centroides de los 3 clústeres con los atributos más significativos, se observa que el atributo Cantidad Obsequios, los puntos de cada clúster están bien diferenciados de los demás y al igual que en la mayoría de cada atributo en cada

clúster, con una pequeña cercanía en los puntos de los atributos Categoría en el clúster 1 y 2 y en el atributo Ciudad en los clústeres 0 y 1.

En el análisis con gráfico de radar, resalta las variaciones entre los clústeres generados, como en el modelo *k-means*, en el clúster 0, el porcentaje de descuento es alto para la mayoría de las marcas y categorías de los productos, y en el clúster 1, el valor de descuento y las cantidades obsequiadas son altas para la mayoría de los meses de compra.

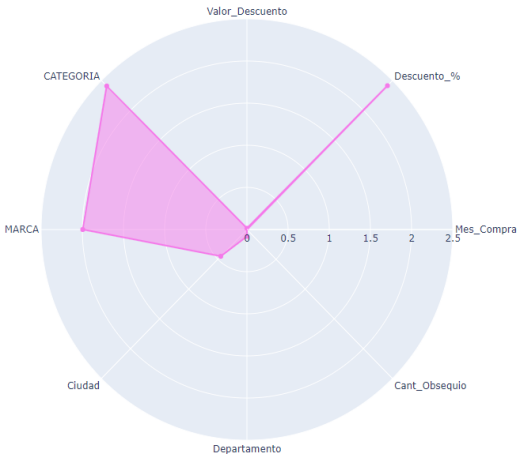


Imagen 66: Modelo *K-means* 100% – Clúster 0

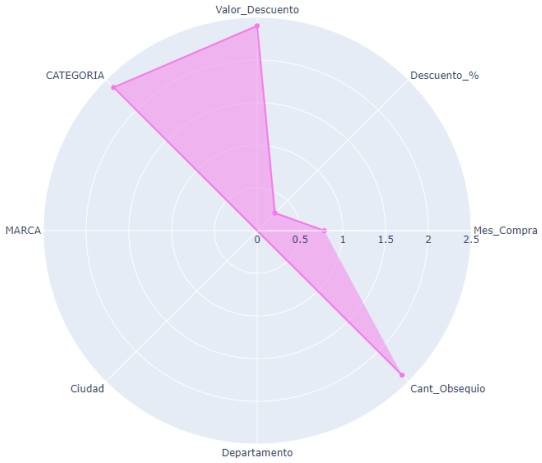


Imagen 67: Modelo Jerárquico 100% – Clúster 0

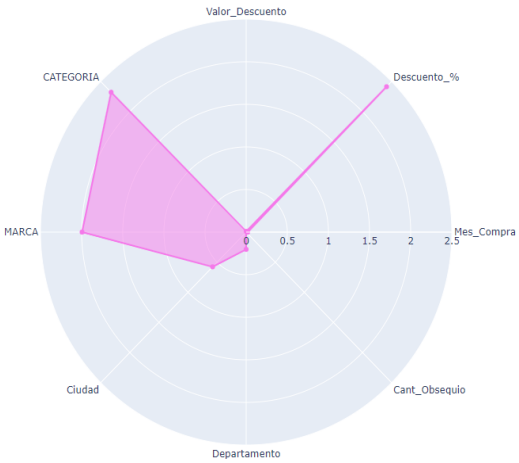


Imagen 68: Modelo *K-means* 41% – Clúster 0

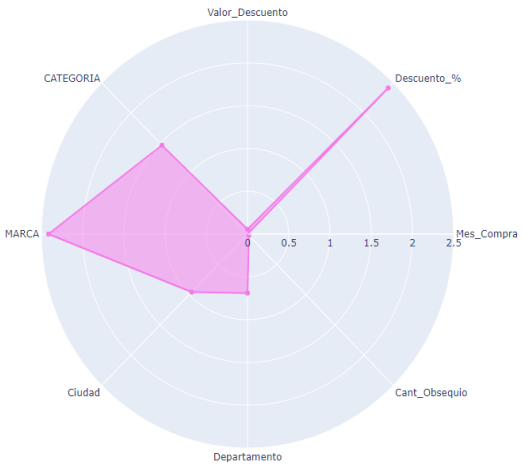


Imagen 69: Modelo Jerárquico 41% – Clúster 0

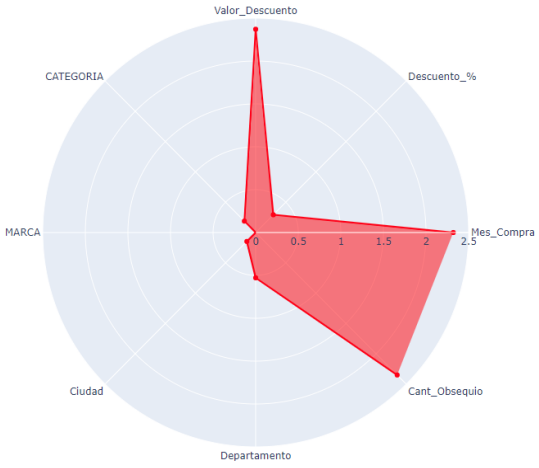


Imagen 70: Modelo *K-means* 100% – Clúster 1

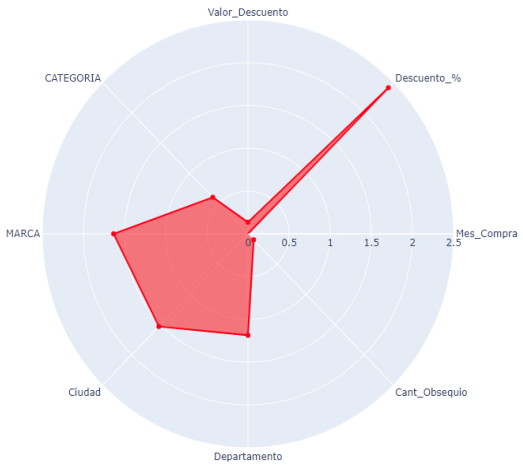


Imagen 71: Modelo Jerárquico 100% – Clúster 1

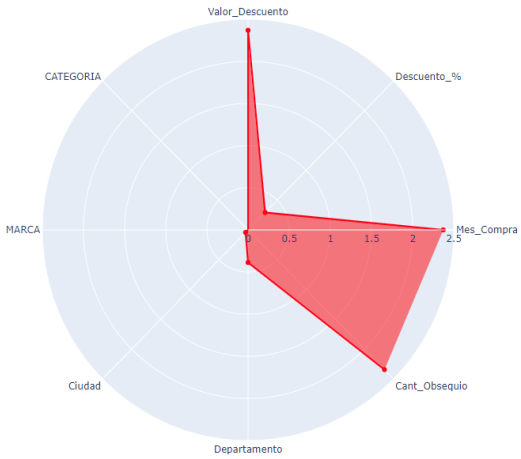


Imagen 72: Modelo *K-means* 41% – Clúster 1

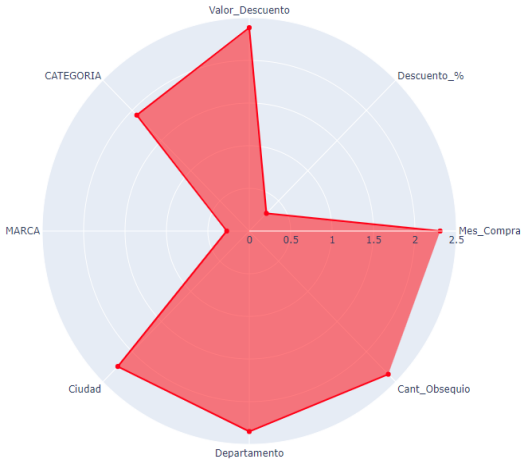


Imagen 73: Modelo Jerárquico 41% – Clúster 1

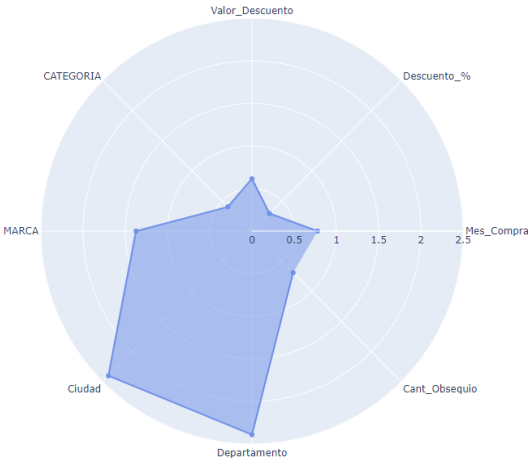


Imagen 74: Modelo *K-means* 100% – Clúster 2

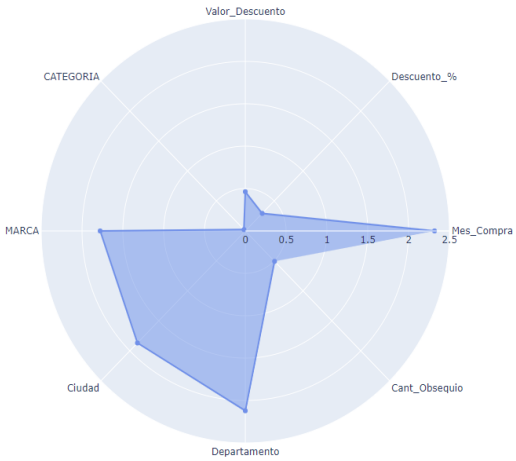


Imagen 75: Modelo Jerárquico 100% – Clúster 2

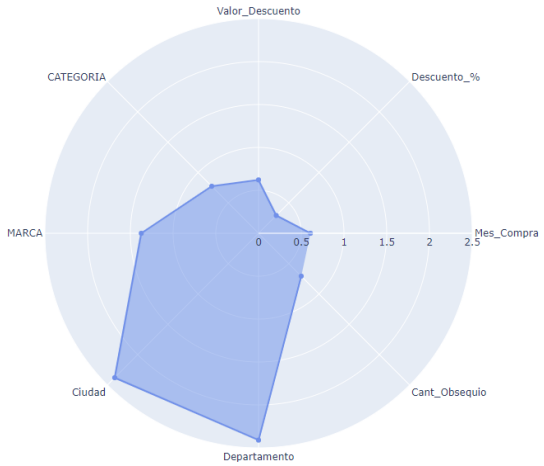


Imagen 76: Modelo *K-means* 41% – Clúster 2

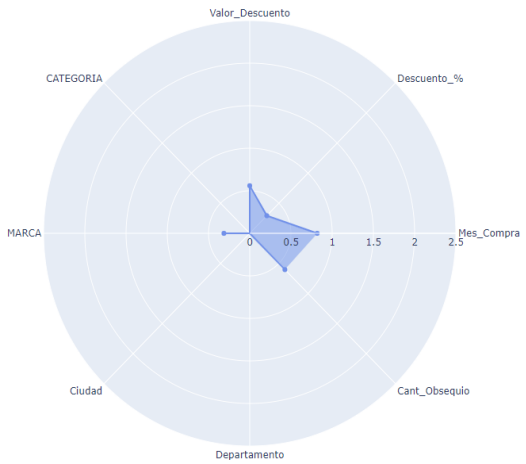


Imagen 77: Modelo Jerárquico 41% – Clúster 2

Del mismo modo se destaca las similitudes de los modelos k-means al 100% y al 41%, y cierta similitud con el modelo jerárquico al 41%, pero con el modelo jerárquico al 100%, se tiene una mayor diferencia, lo que se puede evidenciar con el índice de Davies-Bouldin, a continuación:

Tabla 12: Comparativo resultados índices de Davies-Bouldin datasets Obsequios

Datasets	Número de Clústeres	índice Davies-Bouldin	
		K-means	Jerárquico
Obsequios 100%	3	0.5123821212846695	**0.5008800167913602**
Obsequios 41%	3	**0.5050880241717788**	0.573727364963304

Aunque el mejor índice fue para el modelo jerárquico al 100%, los demás modelos están muy cerca a este índice.

En conclusión, el *dataset* Obsequios tiene 3 clústeres bien diferenciados, que permitiría al departamento de mercadeo realizar estrategias definidas a cada clúster, en cuanto a la selección del modelo y dadas las similitudes del índice de Davies-Bouldin, pero con diferencias notables en los centroides de cada atributo y los resultados de la tabla 9 y 10, se selecciona el modelo k-means.

5.2.2. *dataset* Tipo de Cliente

El *dataset* Tipo de Cliente contiene atributos propios y otorgados a cada cliente, como la condición de pago, que corresponde a los días de crédito otorgado al cliente que generalmente son 30, 60, 90 y 120 días o pago de contado, otro atributo es el cupo de crédito otorgado al cliente, y el grupo de descuento a que tienen derecho, según política comercial pactada.

En la siguiente tabla se analizará los centroides del atributo Condición de Pago, Cupo de Crédito y Grupo de Descuento pactado contra los atributos más destacados:

Tabla 14: Centroides *dataset* tipo de clientes con los atributos elegidos – modelo k-means 100%

Clústeres	Condicion_Pago	Cupo_Crédito	Grupo_Dscto	Valor_Venta	ZONA
Clúster 0	0,582681051	0,334408016	-0,014943616	0,342343107	0,843602
Clúster 1	**2,613777699**	3,068915693	**1,814940716**	2,721541729	0,226365
Clúster 2	2,599725216	1,435629732	1,505476004	1,490262675	0,942789
Clúster 3	**2,613777699**	4,293880164	**1,814940716**	2,591461365	0,226365
Clúster 4	0,966433598	2,048111968	4,170822519	1,314250487	0,727006
Clúster 5	0,796268216	0,823147497	0,839369434	1,175328679	0,598209
Clúster 6	**0,490064276**	0,578154603	-0,310006243	0,244962264	0,261028
Clúster 7	0,571222141	1,027308242	0,16592678	0,394593012	0,226365
Clúster 8	**0,490064276**	1,231468987	-0,52278109	0,429750877	0,226365
Clúster 9	0,552419327	0,251381101	1,779983891	0,391040733	1,976037
Clúster 10	**0,490064276**	0,455658156	0,371454586	0,438124353	1,076839
Clúster 11	**2,613777699**	1,843951223	**1,814940716**	4,365873462	0,226365
Clúster 12	0,531978709	0,292576818	1,514633786	0,456643477	1,582597
Clúster 13	1,236968686	0,618986752	0,519641924	0,631223538	0,488426
Clúster 14	**2,613777699**	1,639790478	0,03522518	1,869329607	1,256467
Clúster 15	-0,101215701	0,378047167	0,994794367	0,501800788	4,946308
Clúster 16	-1,064446901	0,212248632	1,560239186	0,513795353	0,833915
Clúster 17	2,071046527	0,705868534	1,264385103	1,066183342	2,082433
Clúster 18	0,777599772	0,271983392	0,843510564	0,507631975	1,047002
Clúster 19	0,56917514	0,414826007	0,277687442	0,337203878	0,610753
Clúster 20	0,004712042	0,230560836	1,638756842	0,354550814	0,97372
Clúster 21	**0,490064276**	0,313052579	0,990335051	0,463309165	1,345971
Clúster 22	**0,490064276**	0,530043421	-0,069333857	0,39879532	0,275072

Presentado los centroides de los 23 clústeres del *dataset*, se observan unos pocos centros con el mismo valor, en los atributos Condición de Pago y Grupo de Descuento, pero con variaciones en los demás atributos.

En el análisis del gráfico de radar, se observan clústeres con atributos altos y predominantes, al igual que clústeres con atributos bajos, entre ellos está el clúster 4, indicando que los clientes con Cupos en los Créditos altos tienen Ventas medias, y en el caso del clúster 15, que, a pesar de tener presencia en todas las zonas con algunos clientes, no indica que tenga ventas altas, ni medias, sino que son bajas.

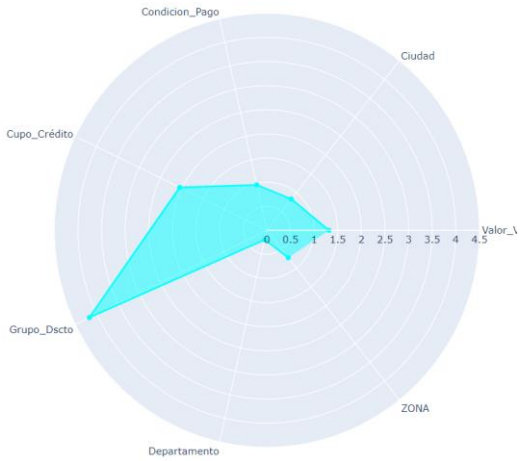


Imagen 78: Modelo *K-means* 100% – Clúster 4

Imagen 79: Modelo Jerárquico 100% – Clúster 4

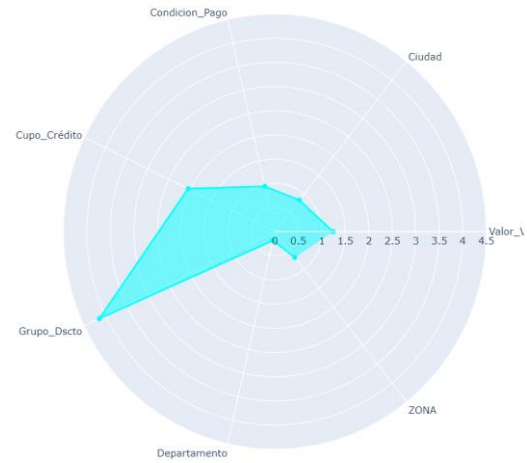


Imagen 80: Modelo *K-means* 15% – Clúster 4

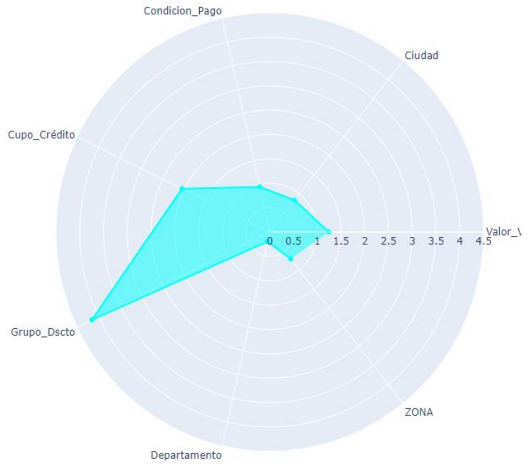


Imagen 81: Modelo Jerárquico 15% – Clúster 4

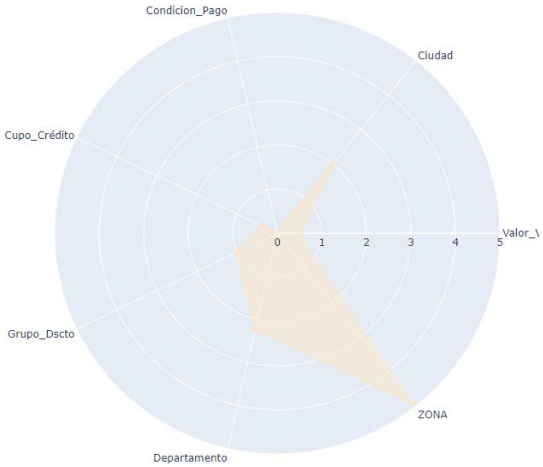


Imagen 82: Modelo *K-means* 100% – Clúster 15

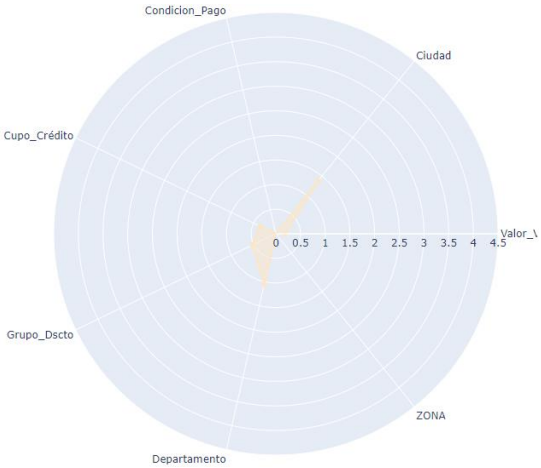


Imagen 83: Modelo Jerárquico 100% – Clúster 15

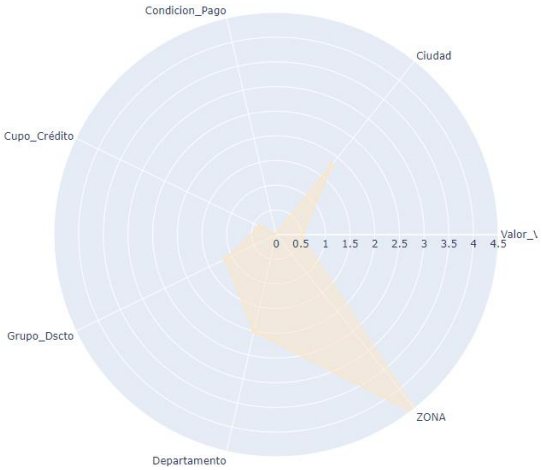


Imagen 84: Modelo *K-means* 15% – Clúster 15

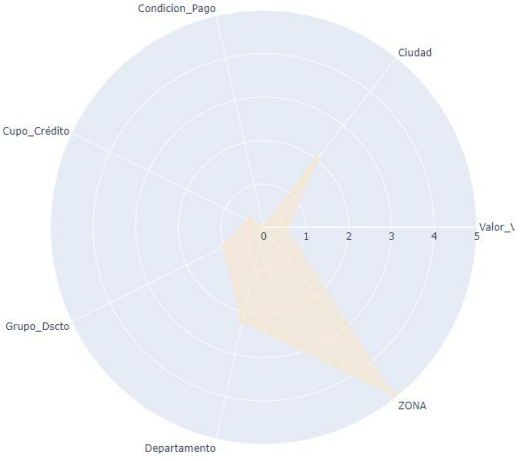


Imagen 85: Modelo Jerárquico 15% – Clúster 15

En los gráficos de radar también se perciben variaciones y similitudes entre los clústeres, como en los clústeres 1, 3, 11, 14 y los clústeres 6, 8, 10, 21 y 22, ilustrados en el Anexo N.1, que, a pesar de ser visualmente semejantes, tienen variaciones en los centroides que definen cada clúster. Y respecto a los modelos podemos observar considerar el índice de Davies-Bouldin para seleccionar el modelo.

Tabla 15: Comparativo resultados índices de Davies-Bouldin datasets Tipo de Cliente

Datasets	Número de Clústeres	índice Davies-Bouldin	
		K-means	Jerárquico
Tipo de Cliente 100%	23	**0.14286109922017623**	54.1595482603534
Tipo de Cliente 15%	23	**0.14191976731922185**	0.1517240844231968

En los resultados del índice de Davies-Bouldin, determina que el modelo *k-means* es el más óptimo, debido a que sus resultados son más cercanos a cero (0), aunque en el modelo jerárquico al 15% obtuvo un valor cercano, sin embargo, el resultado del modelo jerárquico al 100% quedo totalmente por fuera del rango del índice de Davies-Bouldin, debido a que debe ser entre 0 y 1.

Para concluir, se selecciona el modelo *k-means* dados los índices de la tabla 15, y con referencia a las similitudes entre algunos clústeres, se va a revisar en detalle con el departamento de mercadeo y ventas si se aplican estrategias de mercadeo en conjunto o por separado.

5.2.3. *dataset* Demográfico

Este *dataset* muestra el comportamiento de las ventas en valores y unidades en las diferentes ciudades, departamentos y zonas del país, al igual que las ventas en la República de Panamá.

En la tabla 16 se presentan los centroides de los 10 clústeres con los 5 atributos seleccionados.

Tabla 16: Centroides *dataset* Demográfico con todos los atributos– modelo k-means 100%

Clusters	Ciudad	Departamento	ZONA	Valor_Venta	Cant_Total_Venta
Clúster 0	2,755846	2,875775	2,954773	0,362958	0,326656
Clúster 1	0,379458	0,352925	0,354319	0,631614	0,644827
Clúster 2	1,087468	1,327346	0,858495	1,234562	1,293099
Clúster 3	0,81906	0,660681	0,728039	0,50966	0,511208
Clúster 4	-0,64531	-0,51968	-0,29193	3,596937	3,435424
Clúster 5	1,721118	1,521259	1,241917	0,432413	0,425468
Clúster 6	0,610626	0,435489	0,457247	0,850211	0,859698
Clúster 7	0,293143	0,213	0,427096	0,077249	-0,02257
Clúster 8	2,522351	2,482905	2,755527	0,385209	0,360186

En el comparativo de los centroides, se observa que todos los centroides están bien diferenciados, todos con valores diferentes.

Por otro lado, en el análisis con los gráficos de radar, destaca al clúster 4 en donde el valor de las ventas y las cantidades vendidas son altas, pero que se centra en una sola ciudad, Bogotá, la cual tiene el mayor movimiento en la operación de PHARMADERM y SKINDRUG.

En la mayoría de los otros clústeres, ilustrados en el Anexo N. 2, se puede determinar que existe una proporción entre los atributos demográficos y las ventas, y en otros que las ventas son bajas pero que se tiene presencia en ciertas áreas geográficas.

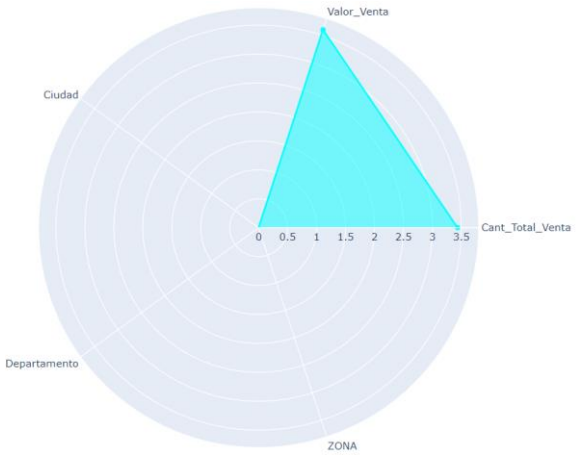


Imagen 86: Modelo *K-means* 100% – Clúster 4

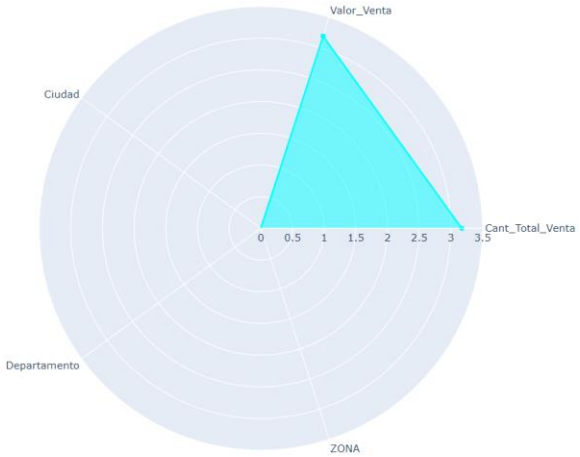


Imagen 87: Modelo Jerárquico 100% – Clúster 4

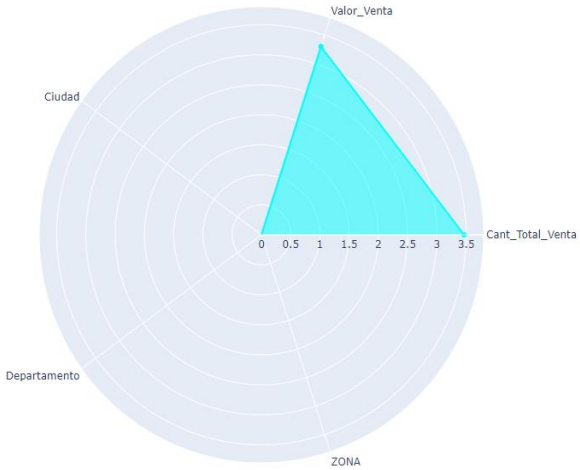


Imagen 88: Modelo *K-means* 15% – Clúster 4

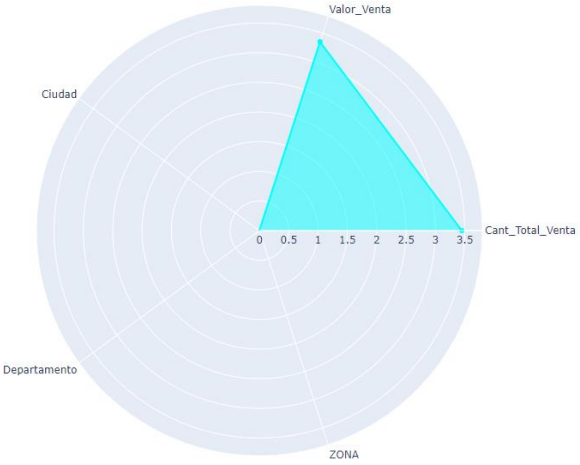


Imagen 89: Modelo Jerárquico 15% – Clúster 4

De igual manera, en los modelos generados se observa una similitud entre los modelos k-means al 15% y el modelo jerárquico al 15%, a diferencia con el modelo jerárquico al 100%, que es el que muestra mayor variación con los demás *datasets*, como se puede también observar en la tabla a continuación.

Tabla 17: Comparativo resultados índices de Davies-Bouldin Demográfico

Dataset	Número de Clústeres	índice Davies-Bouldin	
		K-means	Jerárquico
Demográfico 100%	10	**0.49291602004140656**	6.183116635160513

Demográfico 15%	10	**0.47761694843589203**	0.48379800345681856

Dados los resultados de los índices de Davies-Bouldin, el modelo *k-means* presenta mejor desempeño, al igual que el modelo jerárquico al 15%, igual que el *dataset* anterior, el modelo jerárquico al 100% el índice se sale del intervalo de 0 a 1.

En definitiva, se selecciona el algoritmo *k-means*, debido a los resultados obtenidos en el índice de Davies-Bouldin y a la variación que presentó en los clústeres vistos en la tabla 16, adicionalmente se analizara con el departamento de mercadeo y ventas, si se realizara una estrategia en conjunto para los clústeres 3 y 6 o se hará por separado.

5.2.4. *dataset* Cosméticos

El dataset de Cosméticos contiene los registros de las ventas de todos los productos dermo-cosméticos, que son de venta libre, como protectores solares, limpiadores faciales, hidratantes corporales y faciales, entre otros. El análisis en este *dataset* tiene el objetivo de explorar nuevas ideas para este mercadeo, que supera el 50% de las ventas.

En la tabla 18, aunque todos los centroides tienen valores diferentes, existen unos pocos que tienen alguna proximidad, como en los atributos del mes de compra, la cantidad vendida y el valor de las ventas.

Tabla 18: Centroides *dataset* Cosméticos con los atributos seleccionados– modelo k-means 100%

Clústeres	CANAL	MARCA	Mes_Compra	Cant_Total_Venta	Valor_Venta	Ciudad
Clúster 0	-1,261576322	0,96298644	**0,370844702**	**0,344065721**	**0,379878279**	2,581909465
Clúster 1	1,344881538	0,431361586	1,039565604	0,63156889	0,619036333	-0,001014407
Clúster 2	0,439262182	0,92413353	0,423320179	0,427651904	0,436004437	0,900627929
Clúster 3	1,774598829	1,599186625	1,487527334	1,306867558	1,249035361	2,110413472
Clúster 4	1,926693847	0,787260599	3,42489791	3,32311341	3,517054895	-0,755422578
Clúster 5	1,325549095	0,354924105	1,91100237	0,861725361	0,826522937	0,482350105
Clúster 6	1,410105267	-0,119304264	0,496259521	-0,097338975	0,007398944	1,739301152
Clúster 7	-0,223084091	0,751587678	**0,35697439**	**0,374262306**	**0,399220227**	1,157543999
Clúster 8	2,092615685	3,714612937	-0,262355766	2,324459676	2,063392419	1,701008322
Clúster 9	1,170953968	0,593250764	0,751963757	0,50362415	0,502456167	0,083282541

En los diagramas de radar se observa, en el clúster 4, ventas altas en la mayor parte de los meses, pero demográficamente en una sola área, y en el clúster 6, los atributos CANAL y Categoría tiene valores medio-bajos cubriendo la mitad de los clientes demográficamente, todas estas variaciones y anomalías aportarían al departamento de mercadeo a entender el comportamiento de estos productos dermo-cosméticos.

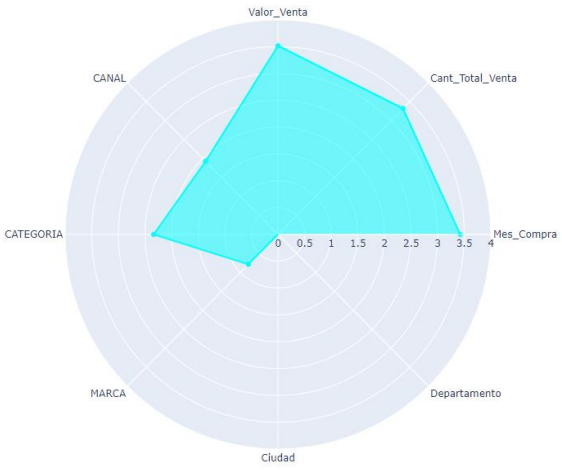


Imagen 90: Modelo *K-means* 100% – Clúster 4

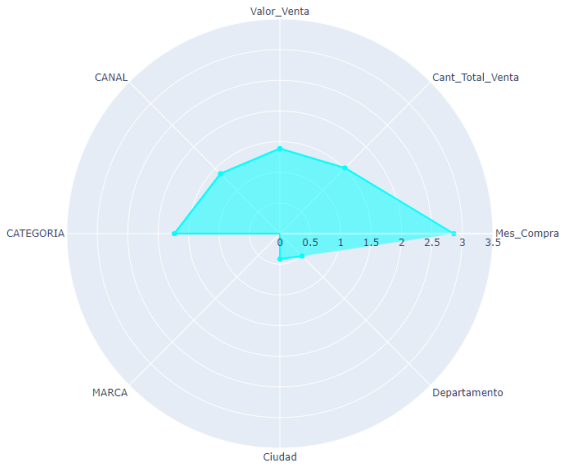


Imagen 91: Modelo Jerárquico 100% – Clúster 4

Imagen 92: Modelo *K-means* 35% – Clúster 4

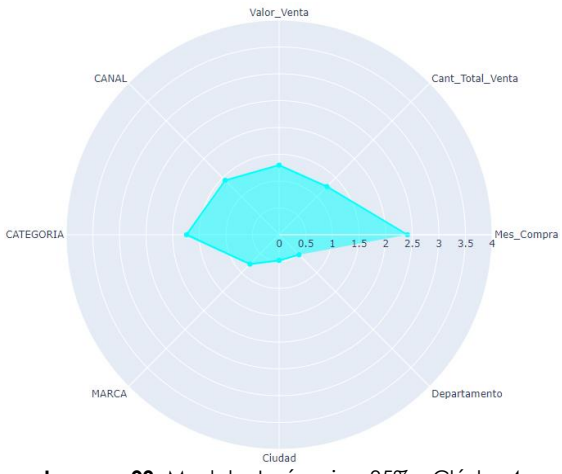


Imagen 93: Modelo Jerárquico 35% – Clúster 4

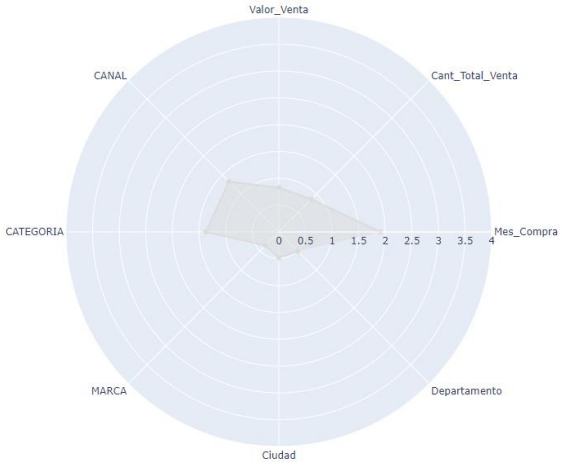


Imagen 94: Modelo *K-means* 100% – Clúster 5

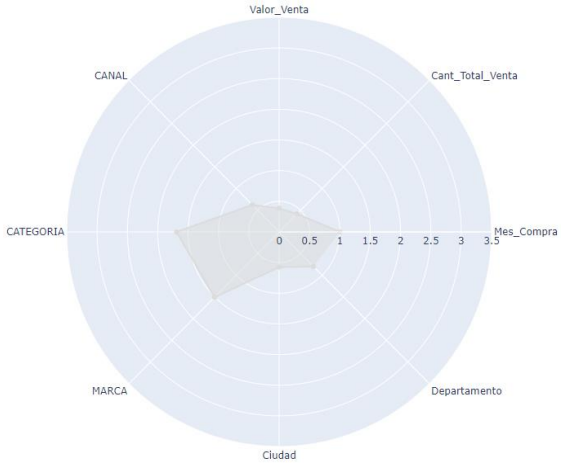


Imagen 95: Modelo Jerárquico 100% – Clúster 5

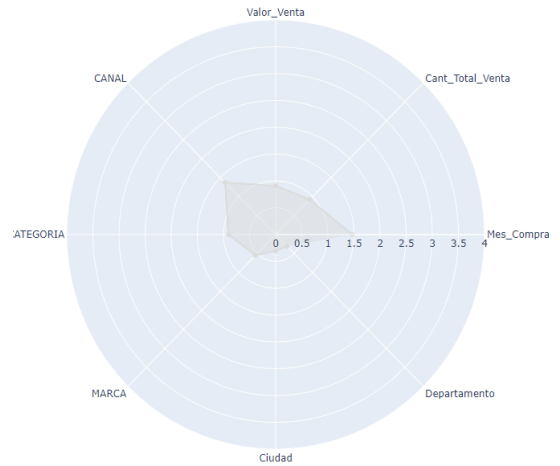


Imagen 96: Modelo *K-means* 35% – Clúster 5

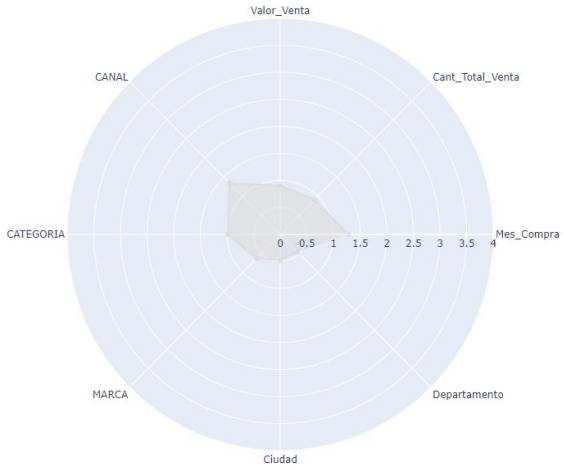


Imagen 97: Modelo Jerárquico 35% – Clúster 5

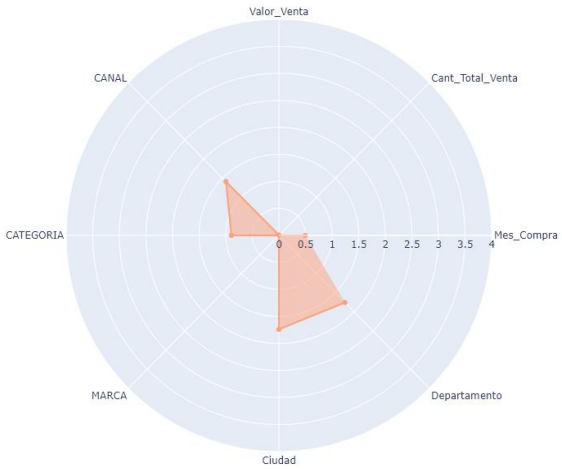


Imagen 98: Modelo *K-means* 100% – Clúster 6

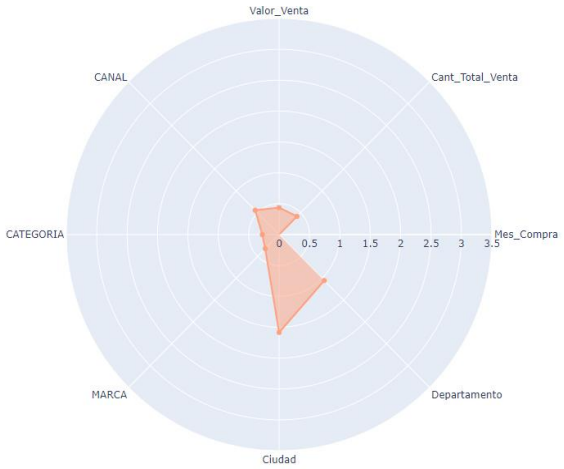


Imagen 99: Modelo Jerárquico 100% – Clúster 6

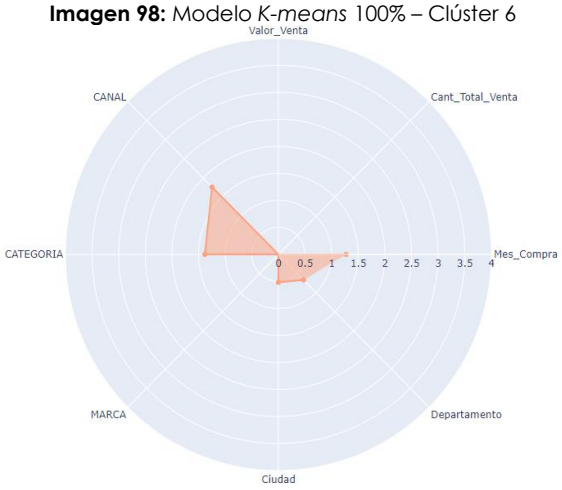


Imagen 100: Modelo *K-means* 35% – Clúster 6

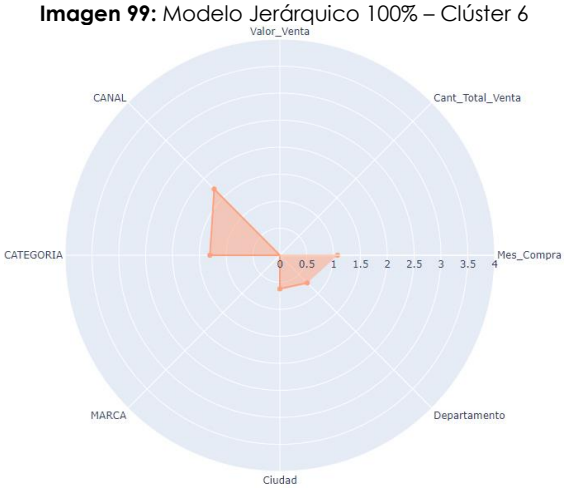


Imagen 101: Modelo Jerárquico 35% – Clúster 6

Además, al igual que en la tabla de los centroides, podemos observar la variedad entre todos los clústeres, ilustrados en el Anexo N.3, y pese a que el clúster 4 tenga la

misma forma geométrica que el clúster 5, el clúster 4 tiene los valores más altos que lo diferencia.

En relación con las similitudes y diferencias de los modelos, existe mayores similitudes entre los modelos k-means y el modelo jerárquico al 35%, por el contrario, el modelo jerárquico es muy diferente a los demás, como también se puede apreciar en el resultado de índice de Davies-Bouldin a continuación:

Tabla 19: Comparativo resultados índices de Davies-Bouldin Cosméticos

Dataset	Número de Clústeres	índice Davies-Bouldin	
		K-means	Jerárquico
Cosméticos 100%	10	**0.48699949177820867**	2.093009038185543
Cosméticos 35%	10	**0.4553914764915647**	0.47237467805455224

Igualmente, que los modelos anteriores, el modelo *k-means* determina mejor resultados.

En resumen, conforme a la tabla de los centroides y al diagrama de radar, señalan clústeres bien diferenciados, útiles para generar estrategias de marketing y presentado el resultado de Davies-Bouldin del modelo jerárquico con el *dataset* al 100%, se observa que el rango se sale del índice especificado, por lo que se selecciona el modelo k-means.

5.2.5. *dataset* Magistrales

El dataset de Magistrales dispone de los movimientos y registros de los productos que son formulados por el Dermatólogo, con componentes específicos para cada paciente. Los atributos considerados para este *dataset* son algunos demográficos y los relacionados con las ventas, con la finalidad de ampliar la cobertura, con planes dirigidos a los clientes apoyada de la segmentación.

En la siguiente tabla se analizará los centroides de todos los atributos del *dataset*:

Tabla 20: Centroides *dataset* Magistrales modelo k-means 100%

Clusters	Cant_Total_Venta	Valor_Venta	Ciudad	Departamento	ZONA
Clúster 0	-0,001443161	0,00078375	-0,38749	-0,26411	-0,37186
Clúster 1	0,635947797	0,632904506	**1,456776**	1,082948	**1,983421**
Clúster 2	2,365495364	2,366311745	**1,930714**	2,181162	**1,388438**

En donde los atributos de ventas y el departamento están bien diferenciados y distanciados, con una pequeña cercanía en los atributos señalados de la ciudad y la zona.

En los gráficos de radar se observa mejor la diferencia entre los clústeres, una alta diferenciación entre cada clúster. En el clúster 0 indica que se deben tomar acciones con los clientes para aumentar las cifras en las ventas. En el clúster 1 tiene una presencia media y en el clúster 2 tiene una presencia alta en el mercado, lo cual daría partida al departamento de mercadeo para analizar estrategias.

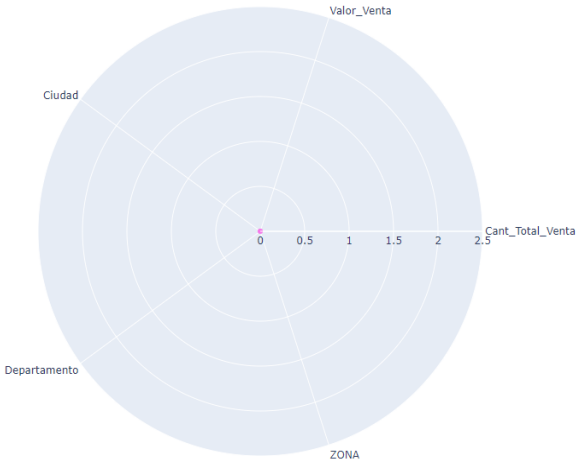


Imagen 102: Modelo *K-means* 100% – Clúster 0

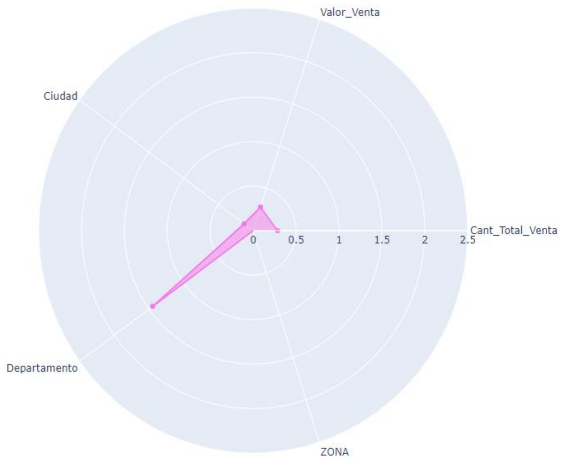


Imagen 103: Modelo Jerárquico 100% – Clúster 0

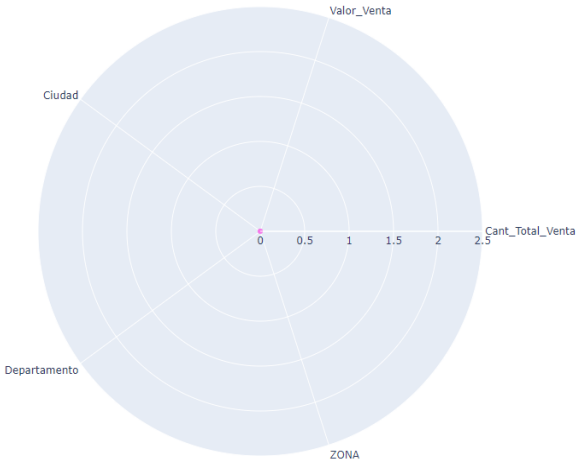


Imagen 104: Modelo *K-means* 36% – Clúster 0

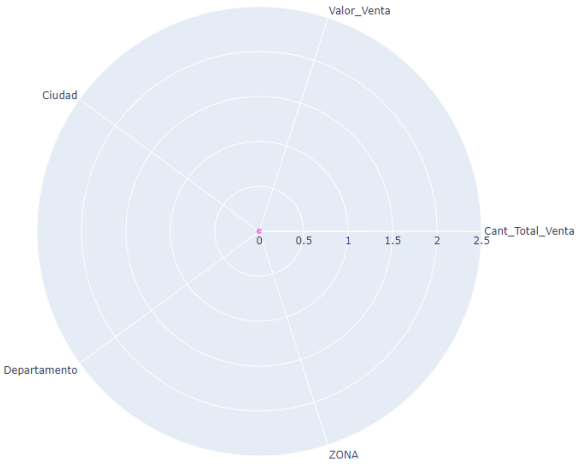


Imagen 105: Modelo Jerárquico 36% – Clúster 0

Imagen 106: Modelo *K-means* 100% – Clúster 1

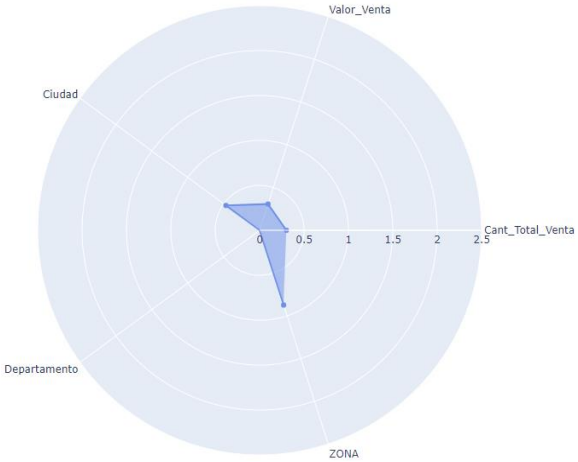


Imagen 107: Modelo Jerárquico 100% – Clúster 1

Imagen 108: Modelo *K-means* 36% – Clúster 1

Imagen 109: Modelo Jerárquico 36% – Clúster 1

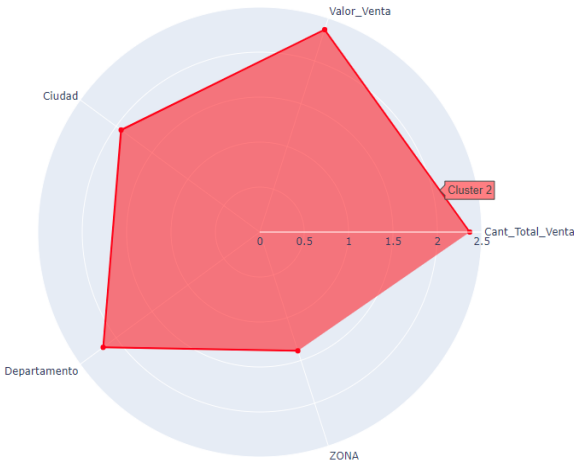


Imagen 110: Modelo *K-means* 100% – Clúster 2

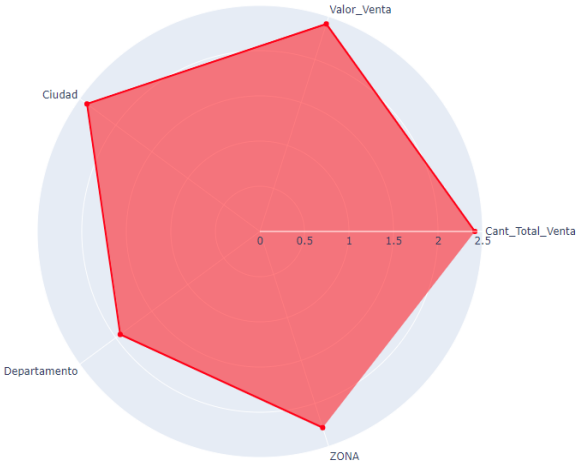


Imagen 111: Modelo Jerárquico 100% – Clúster 2

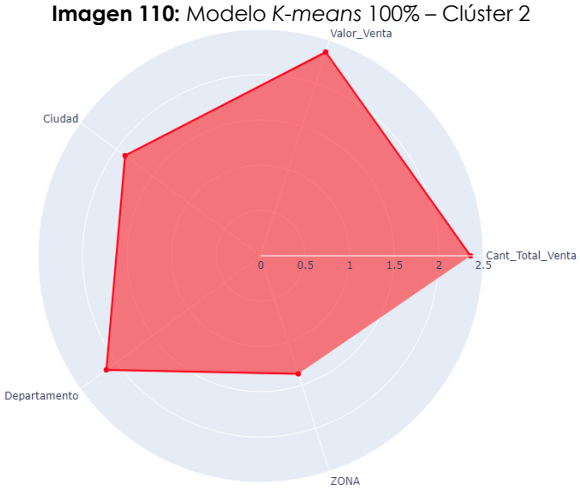


Imagen 112: Modelo *K-means* 36% – Clúster 2

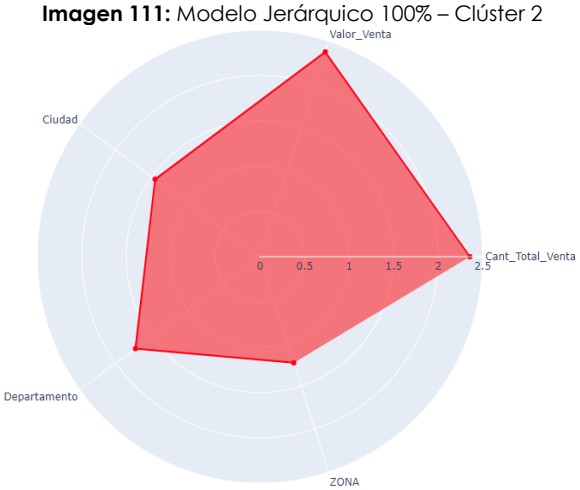


Imagen 113: Modelo Jerárquico 36% – Clúster 2

En cuanto a las similitudes y desigualdad entre los modelos generados, observamos una gran similitud entre los modelos *k-means* al 100% y al 36% con el modelo jerárquico al 36%, en cambio, con el modelo jerárquico al 100%, tiene discrepancias.

En la tabla con los resultados de los índices de Davies-Bouldin presenta el mismo comportamiento.

Tabla 21: Comparativo resultados índices de Davies-Bouldin Magistrales

Dataset	Número de Clústeres	índice Davies-Bouldin	
		K-means	Jerárquico
Magistrales 100%	3	**0.5210373538208933**	0.7423610224358916
Magistrales 36%	3	**0.5166227865955737**	0.5700511384158627

Para concluir, se selecciona el modelo *k-means* por sus dos resultados en el índice Davies-Bouldin ante el modelo jerárquico, el cual cuenta con clústeres bien diferenciados que aportarían utilidad al departamento de mercadeo y ventas.

5.2.6. *dataset* Medicamentos.

Mencionado anteriormente, este *dataset* contiene los registros de las ventas y devoluciones de los productos destinados para el tratamiento de una patología médica, el cual contiene atributos como la categoría y la marca de los medicamentos, los cuales agrupan todas las diversidades de productos.

En la siguiente tabla se presentan los centroides de los atributos más representativos:

Tabla 22: Centroides *dataset* Medicamentos modelo k-means 100%

Clústeres	Mes_Compra	Cant_Total_Venta	Valor_Venta	CANAL	CATEGORIA	Ciudad
Clúster 0	**0,761901519**	**0,628084416**	**0,653607019**	-0,23098	1,620399	0,497721
Clúster 1	1,353737134	0,906980949	0,959934343	0,930218	0,851447	0,268786
Clúster 2	0,89315995	0,67885628	0,701951727	0,615316	1,818816	0,341157
Clúster 3	2,084317392	2,094257445	1,868934965	1,805688	0,359995	1,588056
Clúster 4	-1,54681836	3,262517692	3,359947717	2,446423	-0,46196	2,759512
Cluster 5	1,537029566	-0,657523301	-0,715840761	2,509792	-0,71416	2,947781
Cluster 6	**0,796841735**	**0,593812261**	**0,629183725**	-0,80377	2,344416	0,669191
Cluster 7	2,289321528	1,317786523	1,266963458	1,037944	0,756963	0,347122
Cluster 8	1,178265079	0,787641879	0,795308547	0,907259	1,184417	0,274604
Cluster 9	0,652244457	0,387585857	0,48000926	0,782104	2,239666	0,30607

En los centroides de los clústeres, a nivel general se observa clústeres bien diferenciados, con unas cercanías en los centroides de los clústeres 0 y 6 en los atributos de ventas y el mes de compra.

En el análisis de los gráficos de radar, a continuación, manifiesta el mismo comportamiento anterior, diferenciando el clúster 0 del 6, con los valores de los centroides más altos.

Por otro parte, se identifican comportamientos en cada clúster, como en el clúster 4 las ventas son altas en la mayoría de las ciudades y en los canales de ventas, pero no existe rotación en la marca y categoría de los productos, a diferencia del clúster 6, en

donde ocurre lo contrario, mayor presencia en mayoría de las marcas, pero las ventas son bajas. Las demás ilustraciones de los clústeres se encuentran en el Anexo N.4.

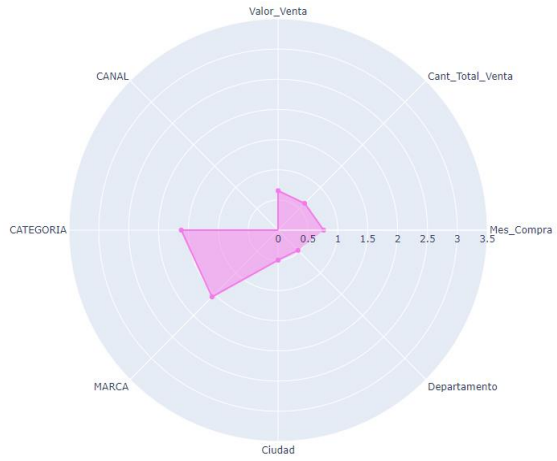


Imagen 114: Modelo *K-means* – Clúster 0

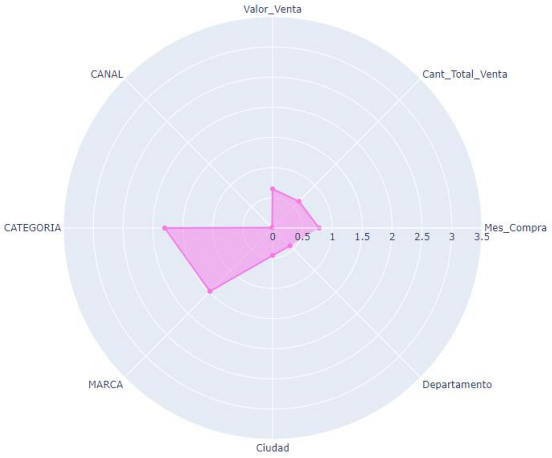


Imagen 115: Modelo Jerárquico – Clúster 0

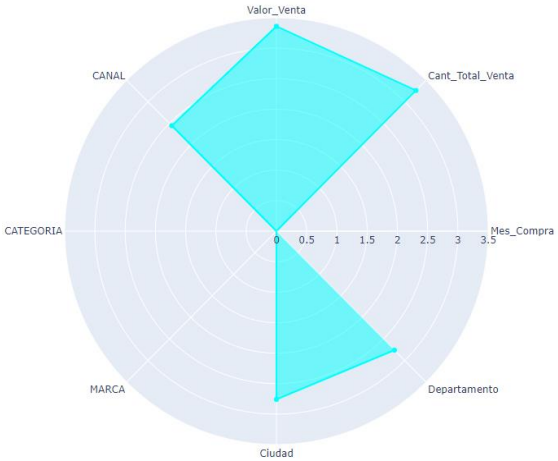


Imagen 116: Modelo *K-means* – Clúster 4

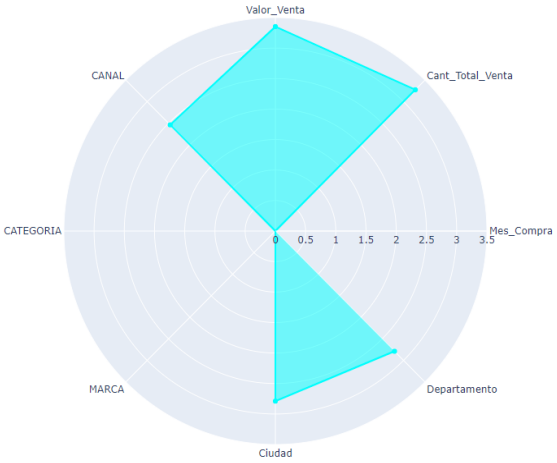


Imagen 117: Modelo Jerárquico – Clúster 4

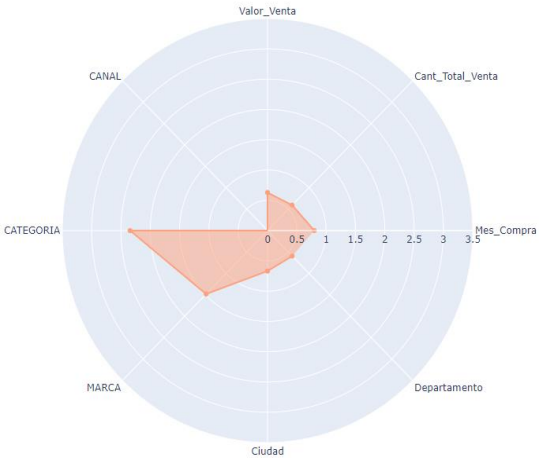


Imagen 118: Modelo *K-means* – Clúster 6

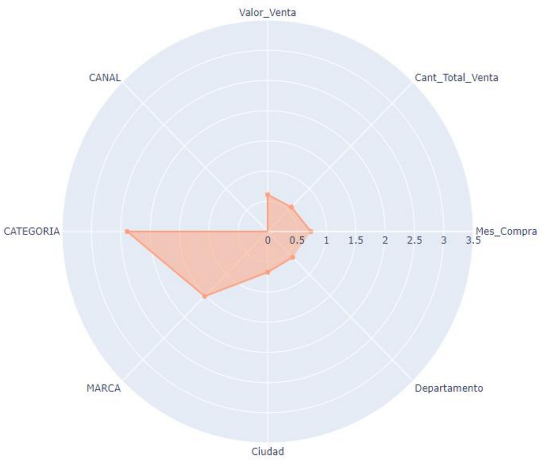


Imagen 119: Modelo Jerárquico – Clúster 6

En relación con las similitudes y diferencias de los modelos k-means y jerárquico, los dos modelos son muy similares en todos los clústeres, aunque en la tabla a continuación, indica que el modelo k-means tiene un mejor resultado en el índice de Davies-Bouldin, aunque sigue estando muy cercanos los dos índices.

Tabla 23: Comparativo resultados índices de Davies-Bouldin Medicamentos

Dataset	Número de Clústeres	índice Davies-Bouldin	
		K-means	Jerárquico
Medicamentos 100%	10	**0.3959949840944447**	0.4201378837101596

Finalmente, se selecciona el modelo k-means, dado el índice de Davies-Bouldin con 10 clústeres bien diferenciados, pero para el caso de los clústeres 0 y 6, se analizará con el departamento de mercadeo y ventas, si se unirán los clientes para dar una estrategia unificada.

5.3. Análisis de los resultados con el departamento de Mercadeo y Ventas.

Presentados los resultados anteriores, el departamento de Mercadeo y Ventas brinda y acoge las siguientes propuestas y estrategias de marketing.

a. dataset Obsequios:

- En el Clúster 0, identificar las marcas y categorías de productos que generan un porcentaje alto en los descuentos, pero no genera Valores en los descuentos.
- En el Clúster 1, se debe crear propuestas con descuentos especiales para los meses con menores compras.
- Verificar las propuestas comerciales actuales por ciudad y departamento del Clúster 2, dado que, a mayor descuento y unidades obsequiadas, aumenta el número de unidades vendidas y respectivamente las ventas.

b. dataset Tipo de Cliente:

- En los clústeres 0, 7, 8 y 22, a no tener un grupo de descuento definido, se otorgarán ofertas Dúo Pack.
- En los clústeres 6 y 19, por las deficientes ventas, se debe verificar los clientes que pertenecen a estos clústeres, y verificar las políticas actuales de mercadeo.
- En cuento a los clústeres 12, 16, 18, 20 y 21 que tienen asignado 50% de cupo con respecto al 100% de sus Ventas, se analizara con el departamento de cartera el aumento del cupo de crédito.
- Con los clústeres 1, 3 y 11, que presentan las ventas más altas, pero solo en una zona demográfica, se ofrecerá incentivos como descuentos, si se presentan aperturas de nuevos puntos de ventas en nuevas ciudades o departamentos.

c. dataset Demográfico

- En el clúster 0, 4 y 8, se obtiene alta presencia demográfica, pero con baja participación en las ventas, por tal motivo la estrategia es dar a conocer los productos por medio de obsequios (muestras reducidas).
- En el caso de los clústeres 4 y 9, que disponen de ventas altas, no se tendrán por el momento, para realizar ningún cambio de mercadeo.
- Por último, los clústeres con ventas bajas y presencia bajan demográficamente, se evaluará las ciudades y departamentos a los que pertenecen para dar una mejor apreciación.

d. dataset Cosméticos

- Iniciando, en los clústeres 1, 5, 6, 7, 8 y 9 que la participación de los productos dermo-cosméticos es baja, se realizaran lanzamientos presenciales en ciudades centrales en cada departamento.
- En los clústeres 4 y 8 que las ventas son altas, no se realizara ninguna acción por el momento.
- Finalmente, en los clústeres con baja presencia en marcas como el 2 y 5, se realizarán promociones u ofertas Dúo Pack y 2x1, para promover productos que no tengan buenas ventas.

e. dataset Magistrales

- En el clúster 0, hay que realizar una revisión exhaustiva de los clientes y de las políticas actuales de las fórmulas magistrales.
- En el clúster 1, se aumentaría la visita médica, enfocada en fórmulas magistrales, para ofertar el vademécum actual, y aplicaría de igual forma el clúster 0.
- En el clúster 2, por el momento no se indicaría ninguna propuesta de ventas.

f. dataset Medicamentos

- Valor de Venta y Cantidad Total de Ventas: En los clústeres 0, 1, 2, 5, 6, 8 y 9 las ventas son bajas, se analizará los clientes que pertenecen a esos clústeres.
- Los tributos de Categorías y Marcas se analizarán específicamente en cada clúster, determinando cuáles son las que no tienen movimiento o tienen muy poco movimiento, para generar una estrategia de mercadeo con visita médica para incentivar la prescripción médica.
- En el caso de la Ciudad y el Departamento se analizará las ventas por productos de cada clúster, para determinar el movimiento en cada área geográfica.

6. VISUALIZACIÓN DE LOS MODELOS

Los laboratorios PHARMADERM y SKINDRUG actualmente están ejecutando un proyecto en Análisis de Datos con Power BI, con la información de la base de datos del ERP Siesa Enterprise, que busca implementar visualizaciones interactivas, basados en los reportes de las áreas de mercadeo y ventas, cartera, contabilidad, costos, entre otros. Debido a que Power BI es una solución de análisis empresarial basado en la nube, que permite unir diferentes fuentes de datos, analizarlos y presentar un análisis de estos a través de informes y paneles [20]. Por otra parte, por política general, se solicita estandarizar procesos con el mismo objetivo de negocio en una sola herramienta tecnológica.

En consecuencia, la visualización de los modelos de clustering de este proyecto de grado y los próximos proyectos de analítica de datos con machine learning, se integrarán a Microsoft Power BI, para la visualización de los resultados.

6.1. Especificaciones de la herramienta

Para llevar a cabo la visualización se adquirió la Licencia Power BI Pro y la Herramienta Microsoft Power BI Desktop.

Tabla 24: Especificaciones del Software y el Licenciamiento de Power BI Pro [21].

Características	Descripción/Aplica
Versión Power BI Desktop	2.106.883.0 64-bit
Acceso de aplicaciones móviles	✓
Publicación de informes para compartir y colaborar	✓
Límite de tamaño del modelo	1 GB
Frecuencia de actualización	8/día
Conexión con más de 100 orígenes de datos	✓
Cree informes y visualizaciones con Power BI Desktop	✓
Integración de API y controles	✓
Objetos visuales de IA	✓
Seguridad y cifrado de datos	✓
Métricas de creación, consumo y publicación de contenido	✓
Almacenamiento máximo	10 GB/usuario

La característica que permite la visualización de los clustering desarrollados, es la de objetos visuales de IA (Inteligencia Artificial), debido a que ofrece análisis y visualizaciones avanzadas.

6.2. Objetos visuales de IA

Obtenidos los resultados del entrenamiento de los modelos elegidos, se procede a crear los objetos visuales en Power BI con Python (Py), en donde se ejecuta los scripts de Python en Power BI Desktop, ilustrado en la imagen 282.

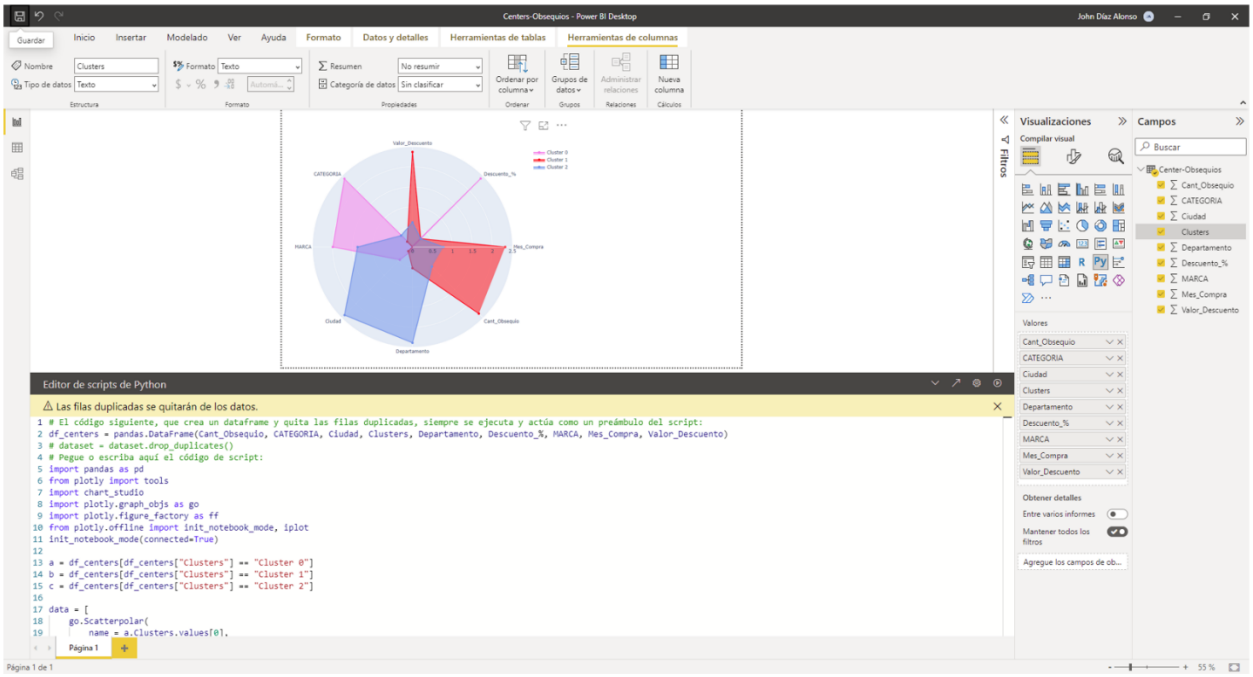


Imagen 120: Dashboard Power BI Desktop - Obsequios

Del mismo modo, se publica el informe en línea en el servicio de Power BI, en el área de trabajo definida, y al grupo de usuarios con los permisos establecidos, que, en este caso, la visualización obtenida de los modelos de clustering, están dirigidos al grupo de usuarios de Mercadeo y Ventas, como el Gerente de Ventas, el ejecutivo de Cuentas y jefe de Producto y Promoción.

6.2.1. Atributos e interacciones en la visualización.

Con el fin de orientar la visualización al grupo de mercadeo y ventas, los atributos seleccionados para el Dashboard son:

- Mes_Compra
- Cantidad_Total_Venta
- Descuento_%
- Valor_Venta
- Valor_Descuento
- CANAL
- CATEGORÍA
- MARCA
- Ciudad
- Condición_Pago
- Departamento
- Cant_Obsequiada
- Cliente

Asimismo, las interacciones que se podrán realizar en las visualizaciones de Power BI, serán:

- Seleccionar un solo clúster o trasponer varios clústeres de un dataset definido en una sola vista, con el objetivo de examinar independientemente las variaciones de los atributos en cada uno de los clústeres o comparar simultáneamente dos o más clústeres.

Por otro lado, se utilizarán las características de interacciones que ofrece Power BI, como:

- Efectuar filtros por atributo, con el propósito de verificar el comportamiento de una variable en cada uno de los clústeres.
- Resaltar una fracción de los valores de cada atributo, para observar el impacto que tiene sobre el valor total del atributo más significativo.
- Finalmente, la característica de no interactuar, la cual bloquea cualquier tipo de interacción.

7. CONCLUSIONES

— Efectuada la preparación de los datos, seleccionados los hiper-parámetros para el entrenamiento de los modelos y presentados los resultados obtenidos en la evaluación, se desarrolló y escogió el algoritmo ideal para la identificación de conglomerados de los clientes de los laboratorios, el modelo k-means, y respaldado del análisis de resultados del departamento de Mercadeo y Ventas, se generaron estrategias de *marketing*.

— La limitación presentada en las capacidades de cómputo de memoria y rendimiento que exigía el modelo jerárquico (*Hierarchical clustering*), fue subsanada con dos alternativas, una de ellas fue dividir los *datasets* en partes iguales o inferiores a 127.000 para poder ejecutar el algoritmo a cada una de ellas y después realizar un análisis de reagrupamiento de los clústeres generaros, la otra alternativa optada fue la de crear porcentajes de los *datasets* que permitiría realizar una comparación con el modelo k-means, para que no fuera descartado.

— Cabe destacar que el modelo k-means, con una menor capacidad de cómputo en memoria y procesamiento, dispone considerablemente de un mejor rendimiento frente al modelo jerárquico, del mismo modo presenta mejores resultados, con una adecuada separación inter-grupos y cohesión intra-grupos.

— Asimismo, se generan diferencias entre los *datasets* reagrupados, en particular en el modelo Jerárquico, los resultados obtenidos en el índice de Davies Bouldin no estuvieron acotados entre el intervalo de 0 a 1 al contrario con los *datasets* segmentados con porcentajes establecidos del 15%, 35%, 36% y 41% que sí estuvieron acotados y con mejores resultados, por lo cual quedo a consideración de la gerencia general, el pago de una instancia *cloud* que permita correr el algoritmo con el *dataset* al 100%, para validar nuevos indicadores.

— El departamento de ventas y mercadeo, en su análisis de resultados, identifico estrategias de *marketing,* a simple vista de algunos clústeres generados.
  - *dataset* Obsequios: Crear propuestas a los clientes con descuentos especiales para las marcas con menores compras.
  - *dataset* Tipo de Cliente: Los clientes que no tienen un grupo de descuento especificado, se otorgarán ofertas Dúo Pack.
  - *dataset* Demográfico: En las ciudades con alta presencia, pero con baja participación en unidades vendidas, se obsequiarán muestras reducidas.
  - *dataset* Cosméticos: Los clientes asociados a los clústeres que la participación de los productos dermo-cosméticos sea baja, se realizaran relanzamientos.
  - *dataset* Magistrales: Se asignará una visitadora médica exclusiva a los clientes con bajas compras, para renegociar los productos del vademécum actual.
  - dataset Medicamentos: Se dará prioridad en la visita médica a marcas de productos con baja rotación para incentivar la prescripción médica.

Sin embargo, el Gerente de Mercadeo y Ventas avanzará con la revisión con cada uno de los clústeres obtenidos, para fijar un plan de mercadeo dirigido a los clientes.

— Adicionalmente, se propuso con el departamento de Mercadeo y Ventas para un trabajo futuro, automatizar el modelo de *clustering k-means* con la finalidad de obtener resultados actualizados, y adicionalmente, explorar modelos predictivos con el objetivo de mejorar la planeación en la manufactura de los productos terminados.

— Finalmente, los datos accesibles fueron las tablas con los movimientos de las ventas y la información específica de los clientes, para cada laboratorio, en cuanto a la preparación de los datos se eliminó información correlacionada, se asignaron valores a los campos *NaN*, entre otros, para la construcción de los modelos se seleccionaron hiper-parámetros por medio de experimentos, y con respecto a la evaluación de la calidad de los modelos desarrollados, se realizó con la medida del índice *Davies-Bouldin*, y por último, la visualización de los resultados obtenidos se llevaron a cabo con *Microsoft Power BI*, debido a la política de estatalización de procesos tecnológicos y al licenciamiento otorgado.

8. REFERENCIAS BIBLIOGRÁFICAS

[1] A. Zuleta, "Colombia.AI – Comunidad de Machine Learning e Inteligencia Artificial de Colombia," *Cyxtera*. https://colombia.ai/ (accessed Apr. 14, 2021).

[2] M. Cleary, *Python Data Science*, vol. 53, no. 9. 2019.

[3] J. Palma and R. Marín, "Inteligencia Artificial," *Inteligencia Artificial*, no. December, pp. 1–158, 2008, doi: 10.13140/2.1.3720.0960.

[4] E. Martínez, "Inteligencia Artificial Aplicada al Análisis de Datos," *Topcart*, vol. XXX, pp. 32–35, 2015.

[5] M. Kirk, *Thoughtful Machine Learning*. 2015.

[6] S. Raschka and V. Mirjalili, *Aprendizaje automático con Python*, 2nd ed. Spain: Marcombo, 2019.

[7] Scikit Learn Developers, "Sklearn Cluster KMeans." https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans (accessed Mar. 02, 2022).

[8] Scikit Learn Developers, "Sklearn Cluster DBSCAN." https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html#sklearn.cluster.DBSCAN (accessed Mar. 02, 2022).

[9] Scikit Learn Developers, "Sklearn Cluster Agglomerative Clustering." https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html#sklearn.cluster.AgglomerativeClustering (accessed Mar. 02, 2022).

[10] J. Bobadilla, *Machine Learning y Deep Learning. Usando Python, Scikit y Keras*, 1ª. Madrid: RAMA, 2020.

[11] H. Gómez Ruiz, "Perfilando de nodos, solución para Garlanet," Universitat Oberta de Catalunya.

[12] Scikit Learn Developers, "Sklearn Metrics Davies Bouldin score." https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html (accessed Jun. 08, 2022).

[13] Scikit Learn Developers, "Nearest Neighbors." https://scikit-learn.org/stable/modules/neighbors.html#neighbors (accessed Jun. 13, 2022).

[14] Scikit Learn Developers, "Sklearn Neighbors NearestNeighbors." https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.NearestNeighbors.html#sklearn.neighbors.NearestNeighbors.radius_neighbors (accessed Jun. 13, 2022).

[15] D. Macías, S. Buitrago, and E. García, "Desarrollo de un modelo de clusterización de los taxistas para la rentabilización del segmento corporativo de Smart Taxi," Pontificia Universidad Javeriana, 2018.

[16] D. Vesga, "Modelo de planeación de inventarios para E-commerce, utilizando herramientas de inteligencia artificial para hacer pronósticos de demanda y clasificación de inventarios," Universidad de los Andes, 2020.

[17] IBM Corporation, "Conceptos básicos sobre modelado - Documentación de IBM." https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=modeling-overview (accessed Jun. 12, 2022).

[18] H.-P. Kriegel, E. Schubert, and A. Zimek, "The (black) art of runtime evaluation: Are we comparing algorithms or implementations?," *Springer-Verlag*, pp. 1–38, 2016.

[19] Plotly, "Radar charts in Python." https://plotly.com/python/radar-chart/ (accessed Jun. 23, 2022).
[20] Deloitte España, "¿Qué es Power BI?" https://www2.deloitte.com/es/es/pages/technology/articles/que-es-power-bi.html (accessed Jul. 11, 2022).
[21] Microsoft Corporation, "Comparación de productos y precios | Microsoft Power BI." https://powerbi.microsoft.com/es-es/pricing/ (accessed Jul. 12, 2022).

9. ANEXOS

ANEXO 1. Gráfico de Radar del *dataset* Tipo de Cliente

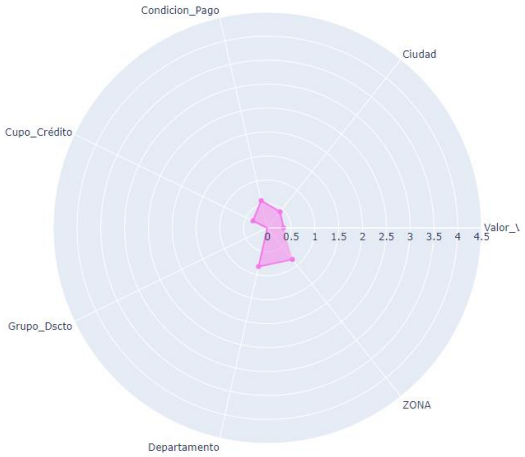


Imagen 121: Modelo *K-means* 100% – Clúster 0

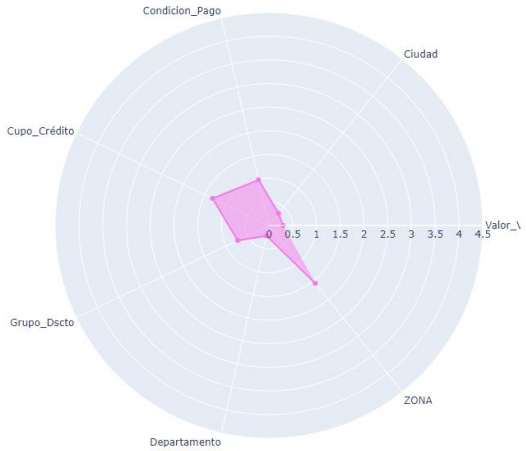


Imagen 122: Modelo Jerárquico 100% – Clúster 0

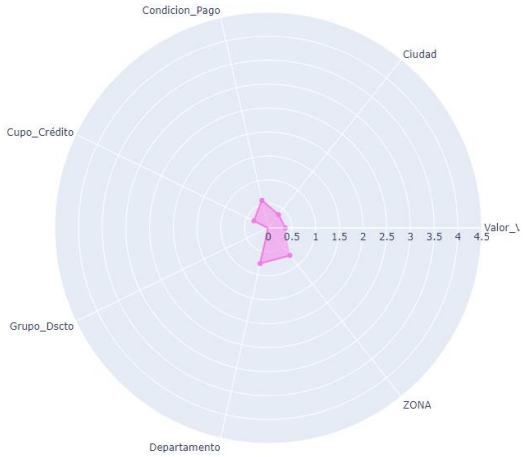


Imagen 123: Modelo *K-means* 15% – Clúster 0

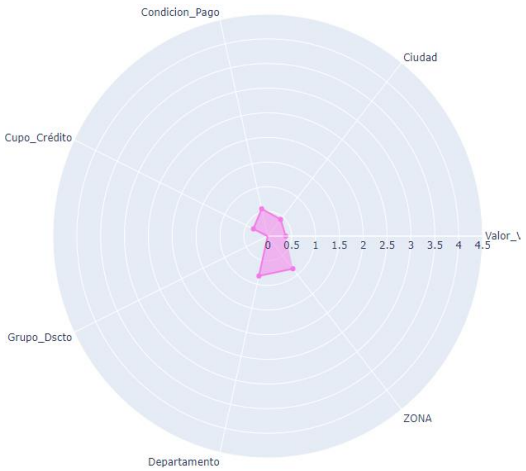
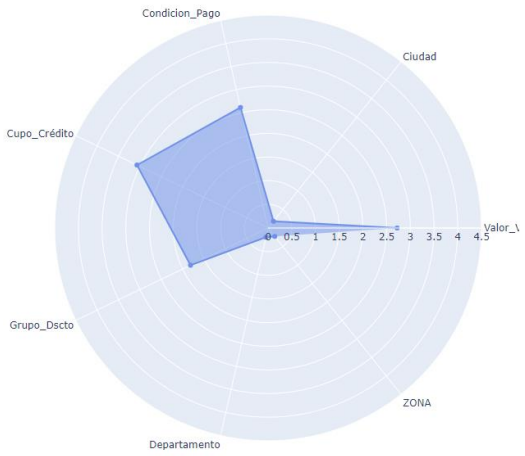


Imagen 124: Modelo Jerárquico 15% – Clúster 0



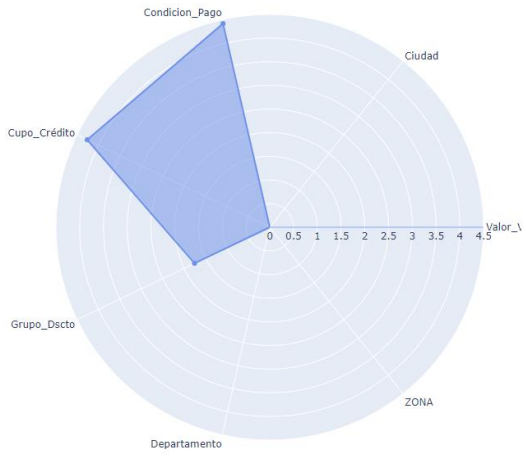


Imagen 125: Modelo *K-means* 100% – Clúster 1

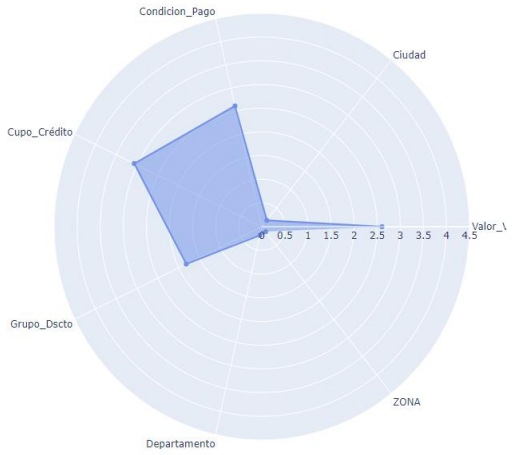


Imagen 126: Modelo Jerárquico 100% – Clúster 1

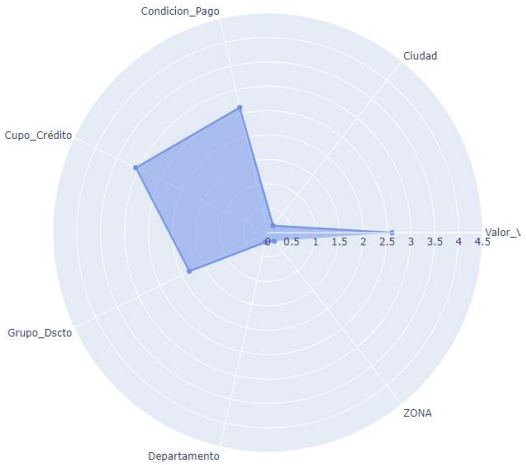


Imagen 127: Modelo *K-means* 15% – Clúster 1

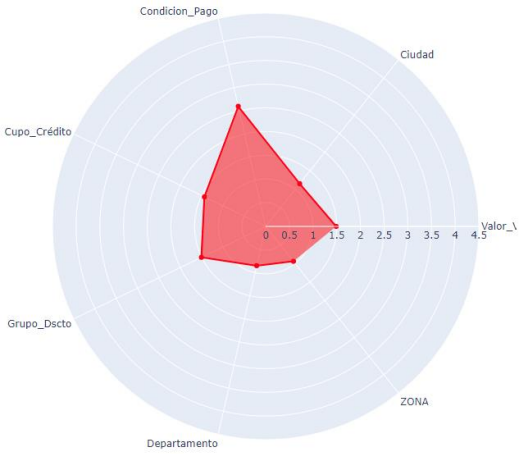


Imagen 128: Modelo Jerárquico 15% – Clúster 1

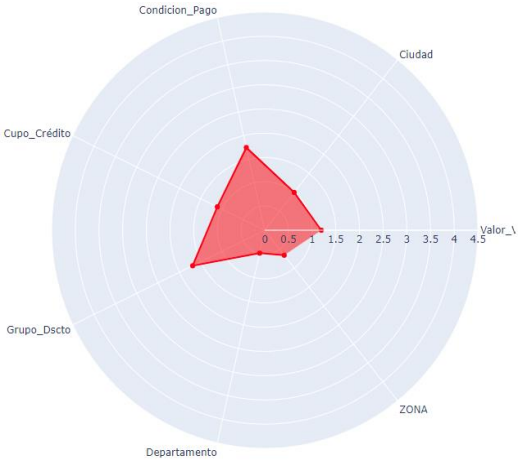


Imagen 129: Modelo *K-means* 100% – Clúster 2

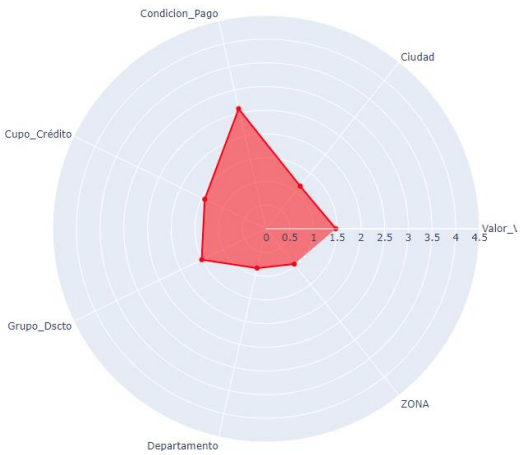
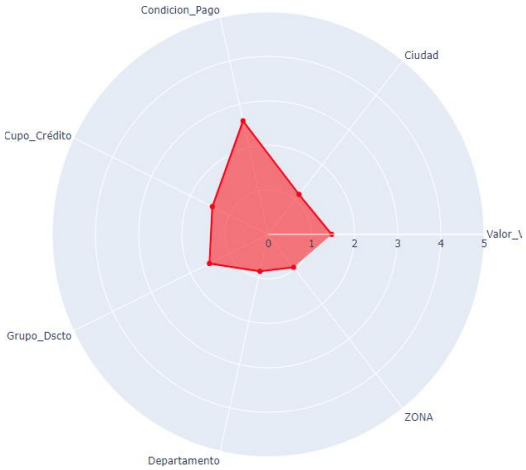


Imagen 130: Modelo Jerárquico 100% – Clúster 2

Imagen 131: Modelo *K-means* 15% – Clúster 2

Imagen 132: Modelo Jerárquico 15% – Clúster 2



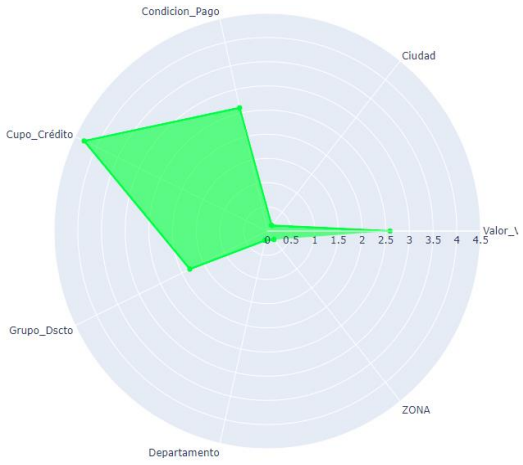


Imagen 133: Modelo *K-means* 100% – Clúster 3

Imagen 134: Modelo Jerárquico 100% – Clúster 3

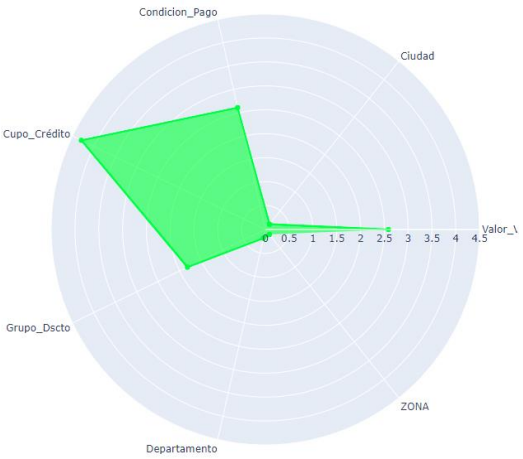


Imagen 135: Modelo *K-means* 15% – Clúster 3

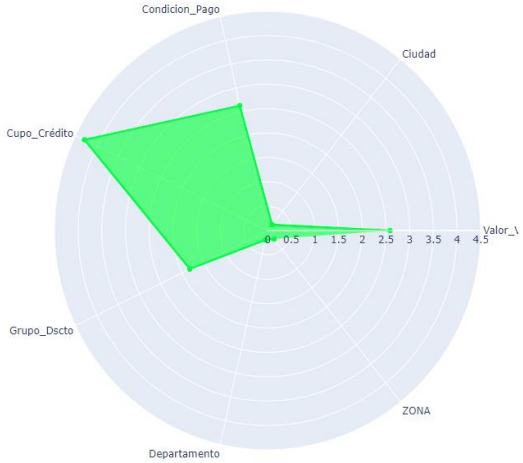


Imagen 136: Modelo Jerárquico 15% – Clúster 3

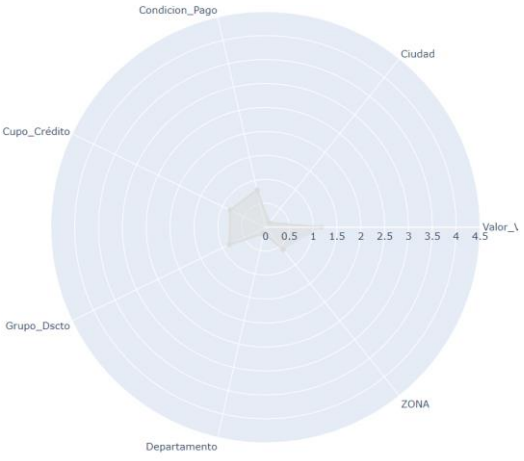


Imagen 137: Modelo *K-means* 100% – Clúster 5

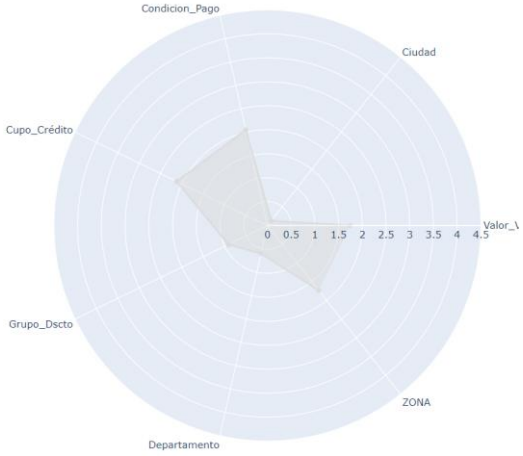


Imagen 138: Modelo Jerárquico 100% – Clúster 5

Imagen 139: Modelo *K-means* 15% – Clúster 5

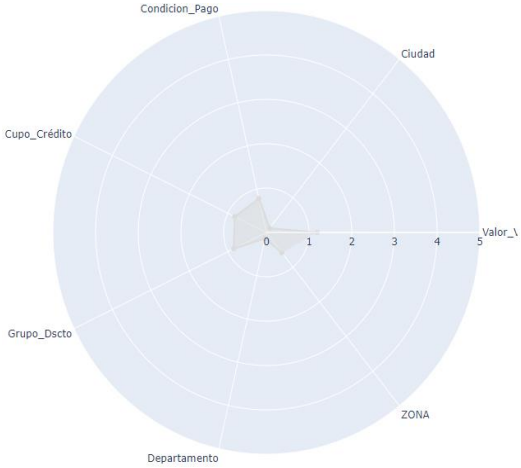


Imagen 140: Modelo Jerárquico 15% – Clúster 5

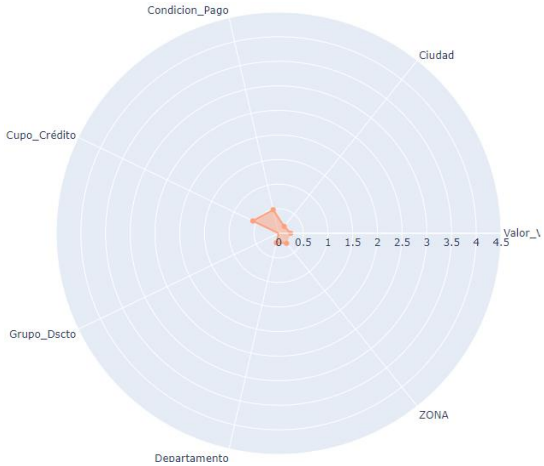


Imagen 141: Modelo *K-means* 100% – Clúster 6

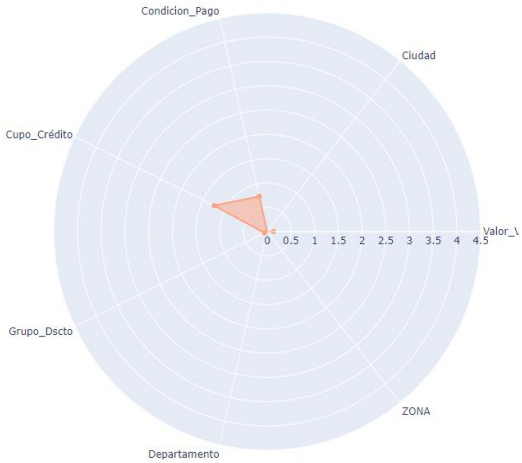


Imagen 142: Modelo Jerárquico 100% – Clúster 6

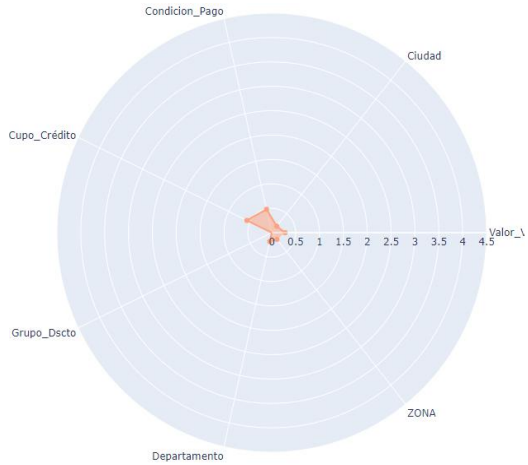


Imagen 143: Modelo *K-means* 15% – Clúster 6

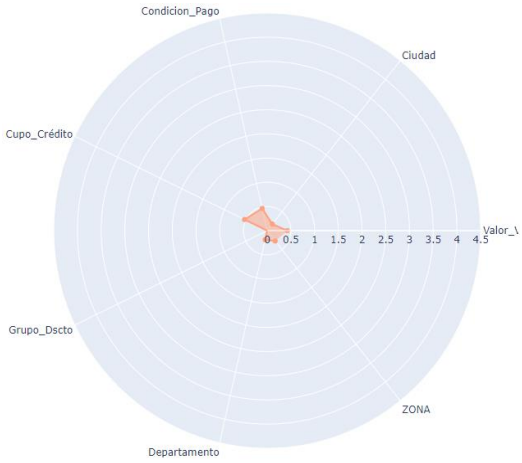


Imagen 144: Modelo Jerárquico 15% – Clúster 6

Imagen 145: Modelo *K-means* 100% – Clúster 7

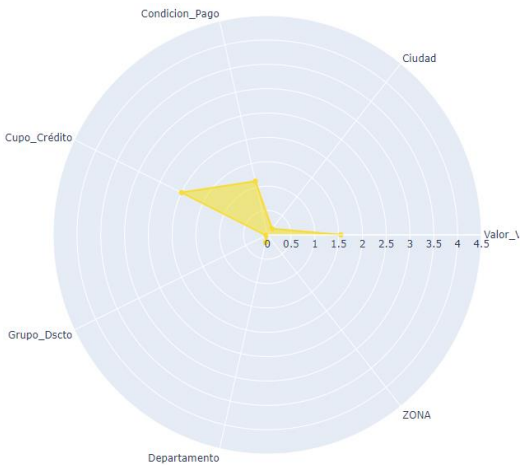


Imagen 146: Modelo Jerárquico 100% – Clúster 7

Imagen 147: Modelo *K-means* 15% – Clúster 7

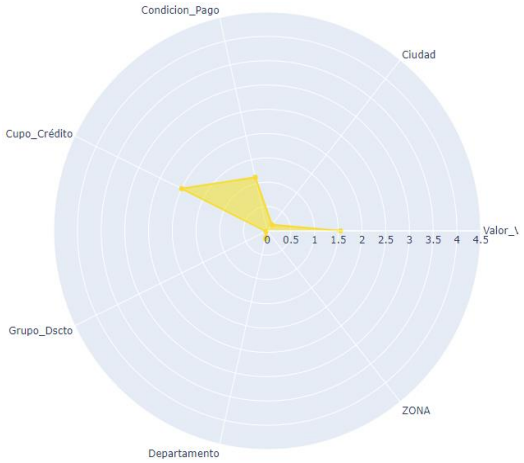


Imagen 148: Modelo Jerárquico 15% – Clúster 7

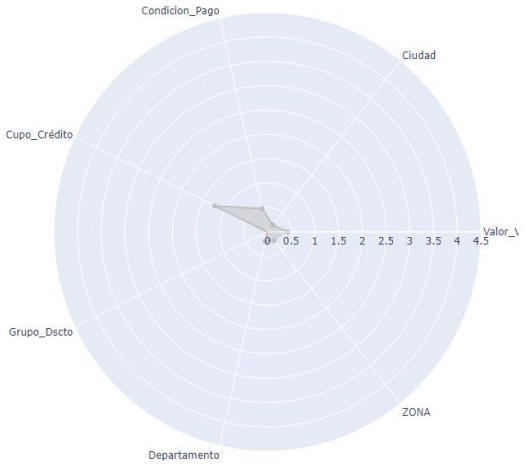


Imagen 149: Modelo *K-means* 100% – Clúster 8

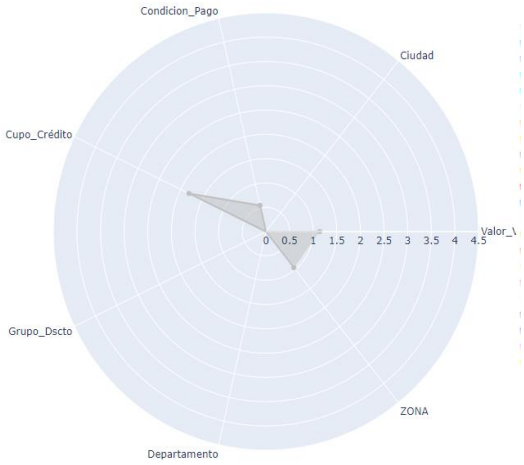


Imagen 150: Modelo Jerárquico 100% – Clúster 8

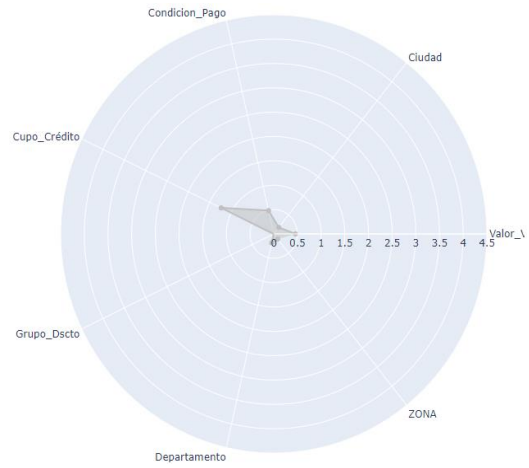


Imagen 151: Modelo *K-means* 15% – Clúster 8

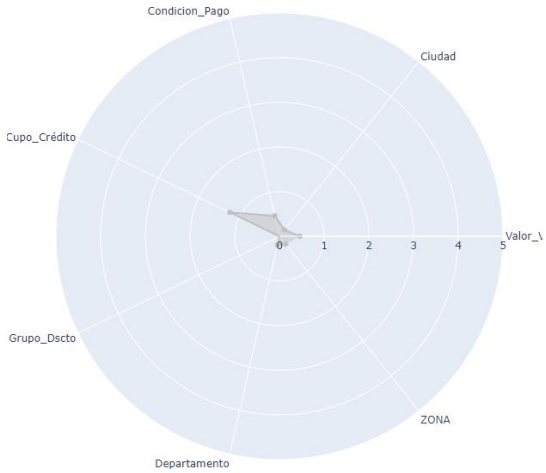


Imagen 152: Modelo Jerárquico 15% – Clúster 8

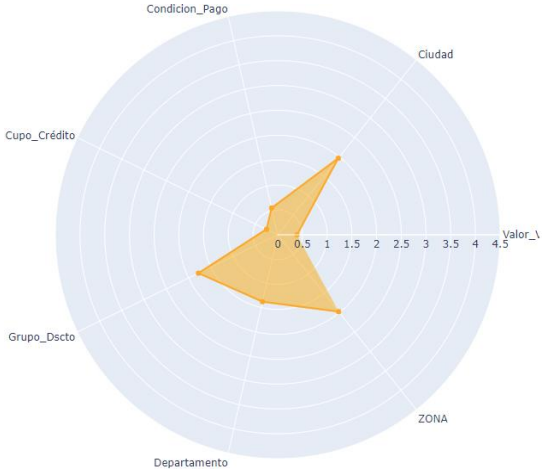


Imagen 153: Modelo *K-means* 100% – Clúster 9

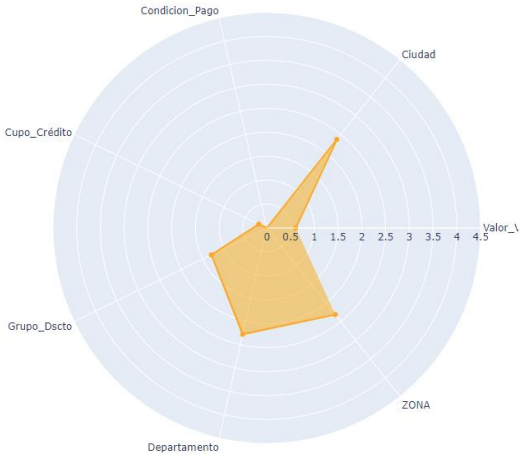


Imagen 154: Modelo Jerárquico 100% – Clúster 9

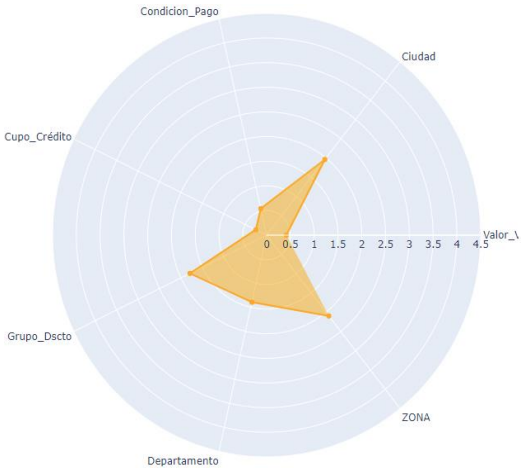


Imagen 155: Modelo *K-means* 15% – Clúster 9

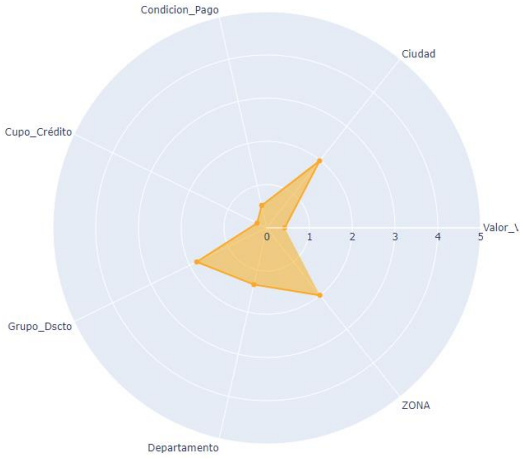


Imagen 156: Modelo Jerárquico 15% – Clúster 9

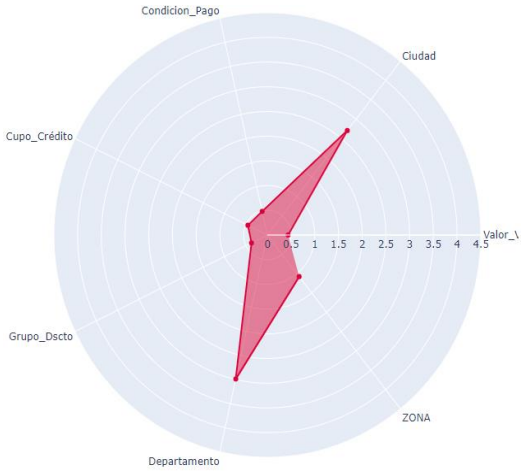


Imagen 157: Modelo *K-means* 100% – Clúster 10

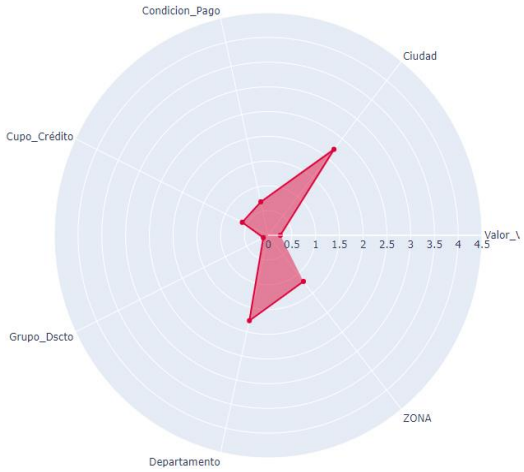


Imagen 158: Modelo Jerárquico 100% – Clúster 10

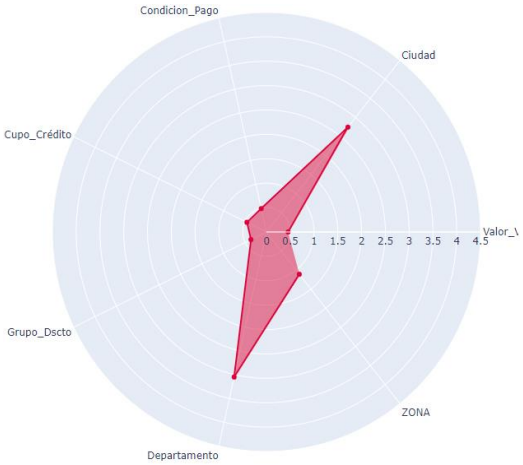


Imagen 159: Modelo *K-means* 15% – Clúster 10

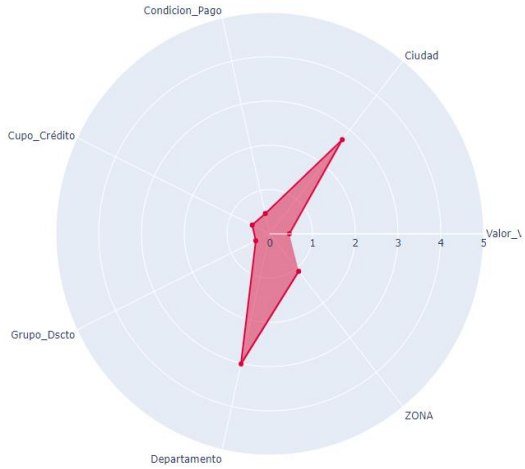


Imagen 160: Modelo Jerárquico 15% – Clúster 10

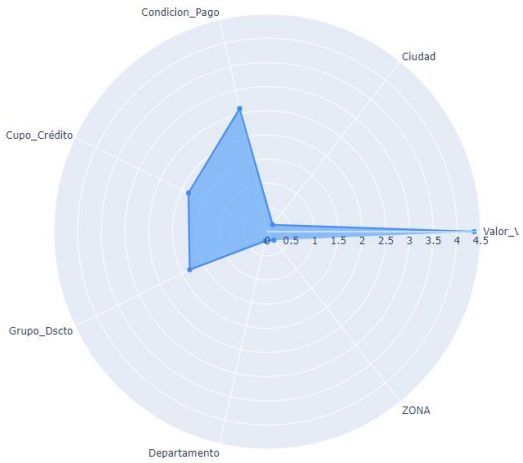


Imagen 161: Modelo *K-means* 100% – Clúster 11

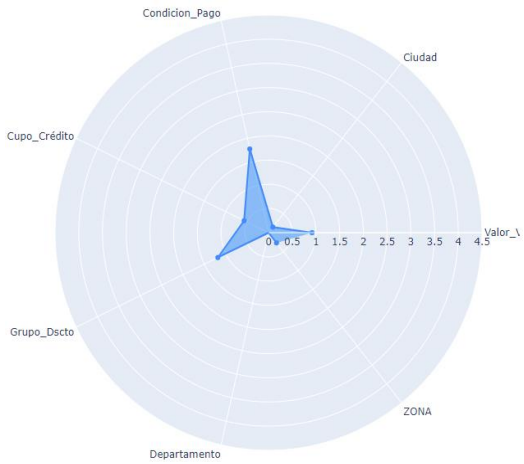


Imagen 162: Modelo Jerárquico 100% – Clúster 11

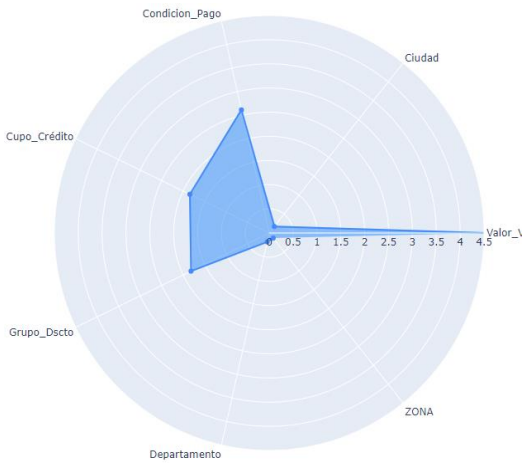


Imagen 163: Modelo *K-means* 15% – Clúster 11

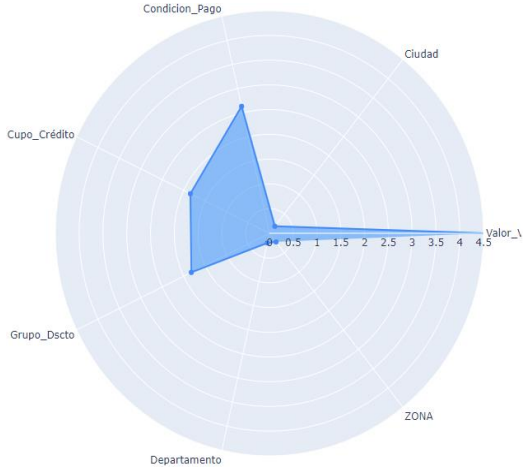


Imagen 164: Modelo Jerárquico 15% – Clúster 11

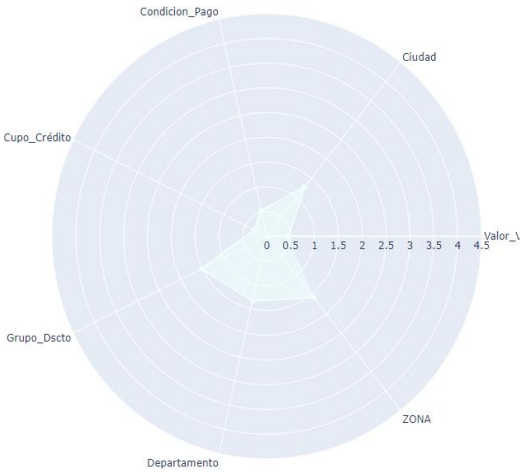


Imagen 165: Modelo *K-means* 100% – Clúster 12

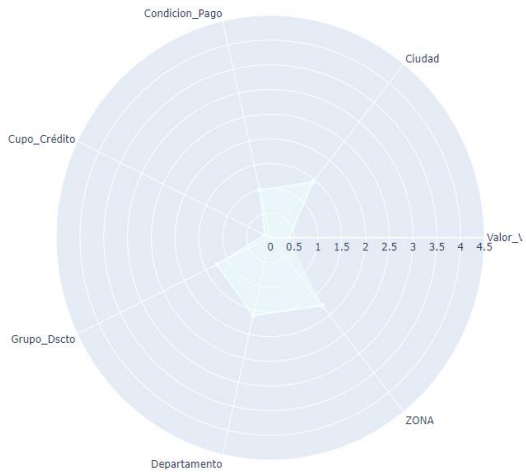


Imagen 166: Modelo Jerárquico 100% – Clúster 12

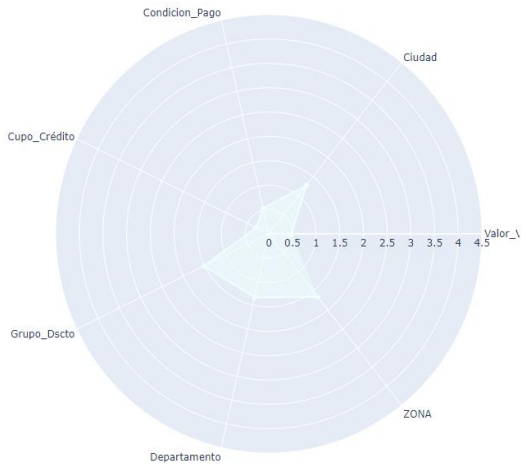


Imagen 167: Modelo *K-means* 15% – Clúster 12

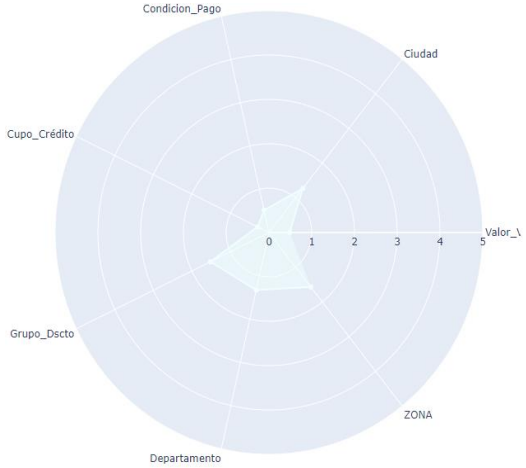


Imagen 168: Modelo Jerárquico 15% – Clúster 12

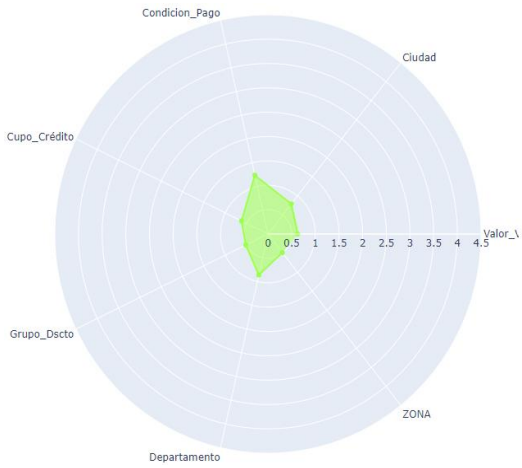


Imagen 169: Modelo *K-means* 100% – Clúster 13

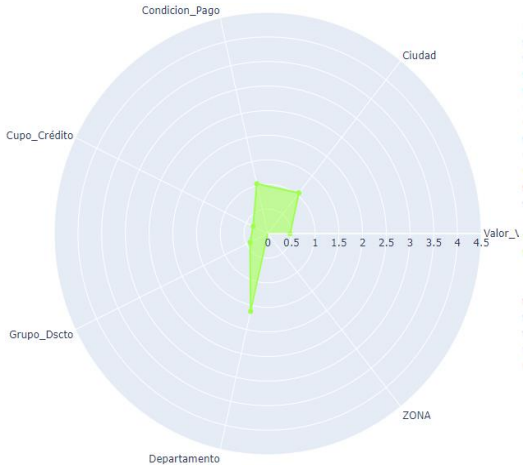


Imagen 170: Modelo Jerárquico 100% – Clúster 13

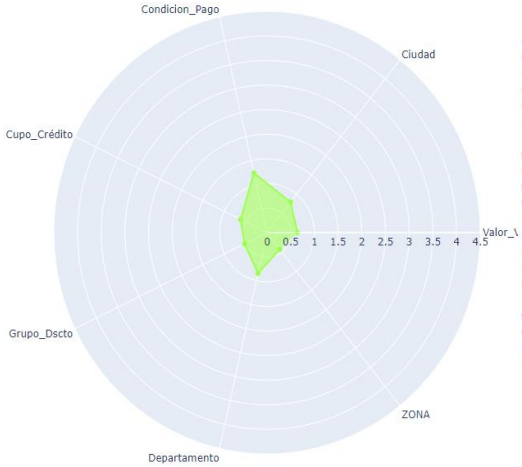


Imagen 171: Modelo *K-means* 15% – Clúster 13

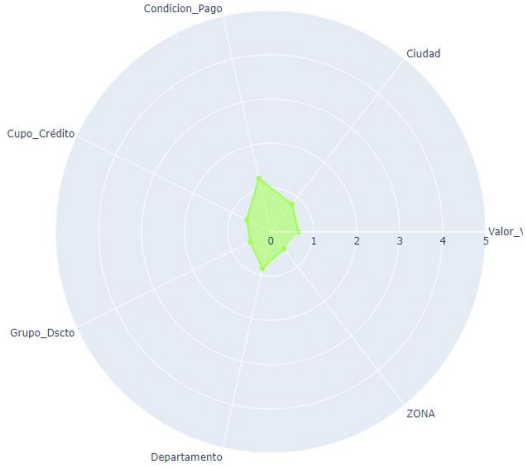


Imagen 172: Modelo Jerárquico 15% – Clúster 13

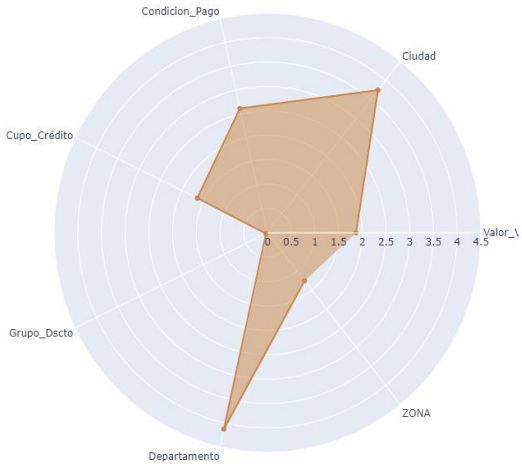


Imagen 173: Modelo *K-means* 100% – Clúster 14

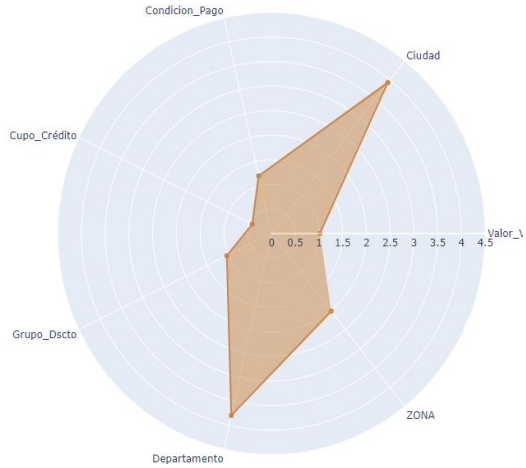


Imagen 174: Modelo Jerárquico 100% – Clúster 14

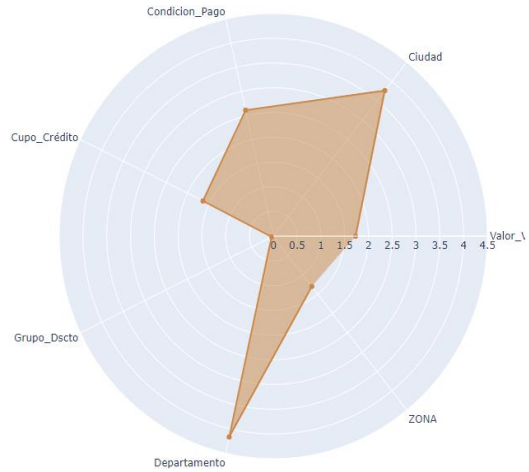


Imagen 175: Modelo *K-means* 15% – Clúster 14

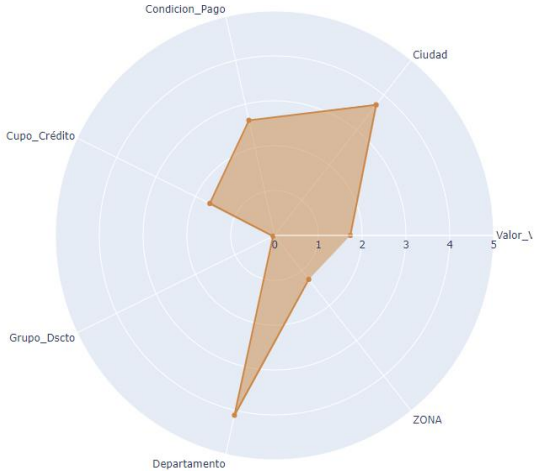


Imagen 176: Modelo Jerárquico 15% – Clúster 14

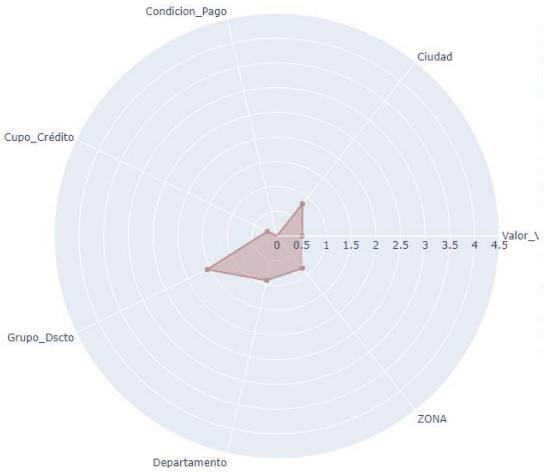


Imagen 177: Modelo *K-means* 100% – Clúster 16

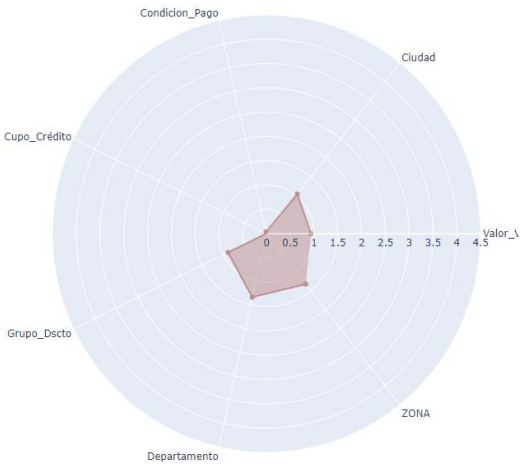


Imagen 178: Modelo Jerárquico 100% – Clúster 16

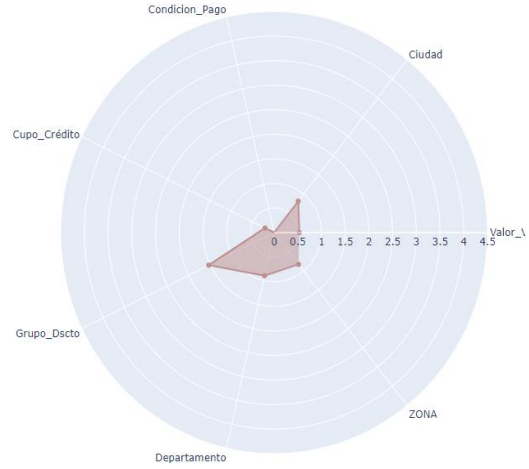


Imagen 179: Modelo *K-means* 15% – Clúster 16

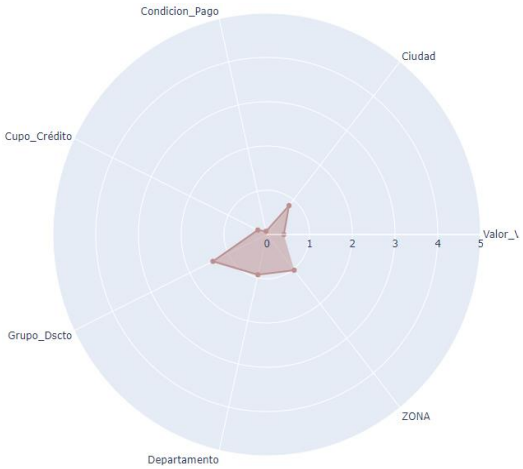


Imagen 180: Modelo Jerárquico 15% – Clúster 16

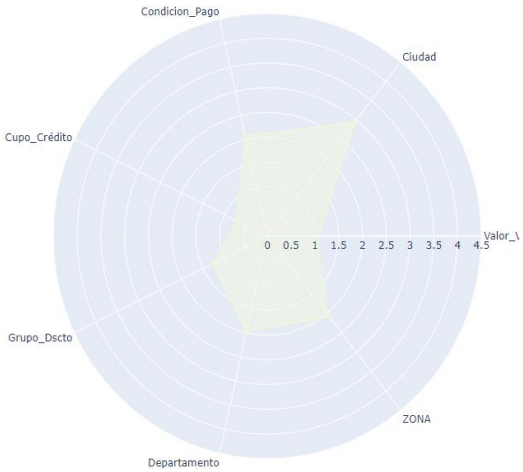


Imagen 181: Modelo *K-means* 100% – Clúster 17

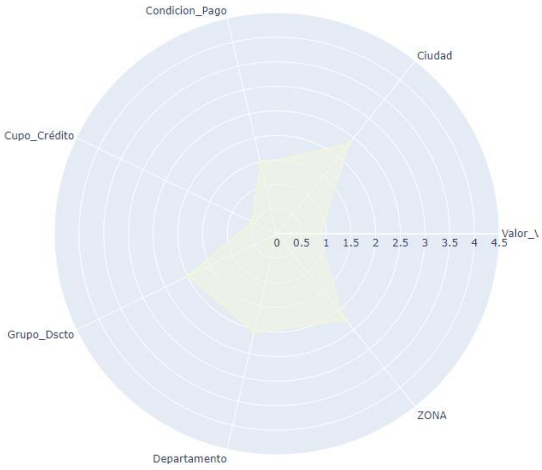


Imagen 182: Modelo Jerárquico 100% – Clúster 17

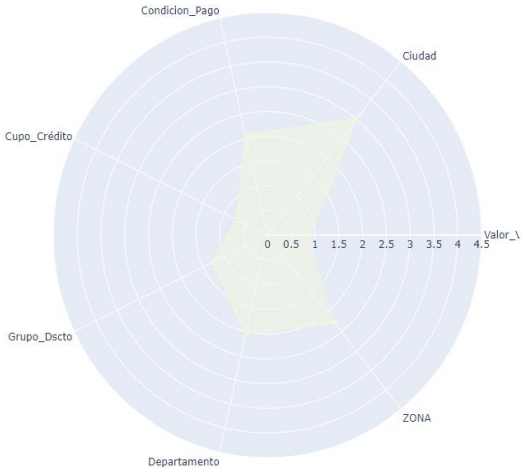


Imagen 183: Modelo *K-means* 15% – Clúster 17

Imagen 184: Modelo Jerárquico 15% – Clúster 17

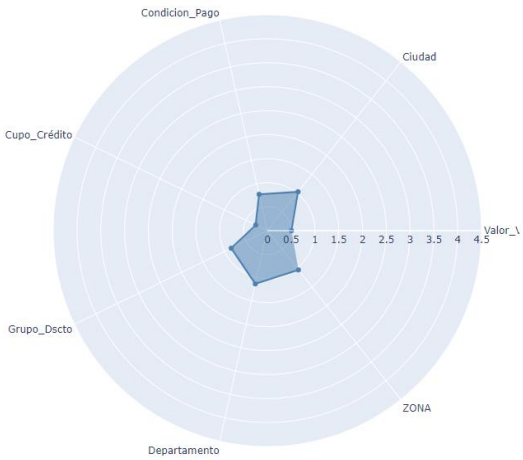


Imagen 185: Modelo *K-means* 100% – Clúster 18

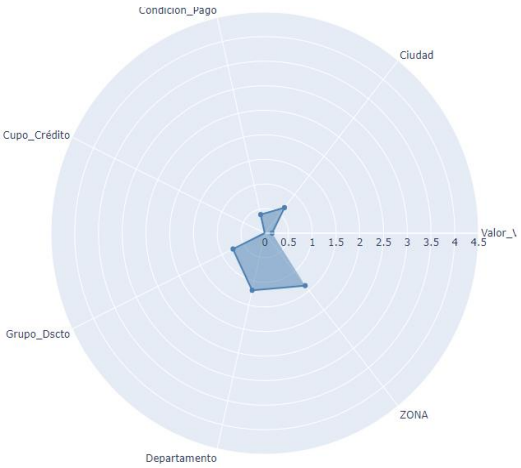


Imagen 186: Modelo Jerárquico 100% – Clúster 18

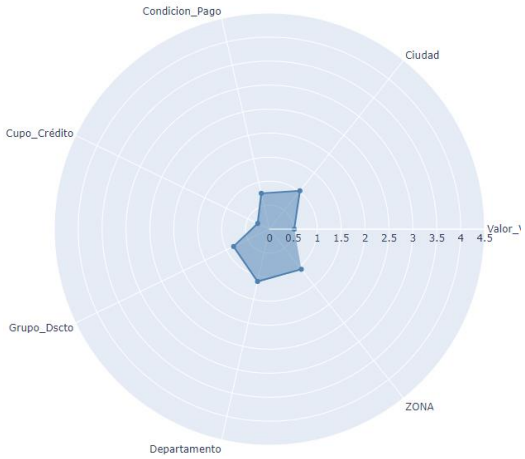


Imagen 187: Modelo *K-means* 15% – Clúster 18

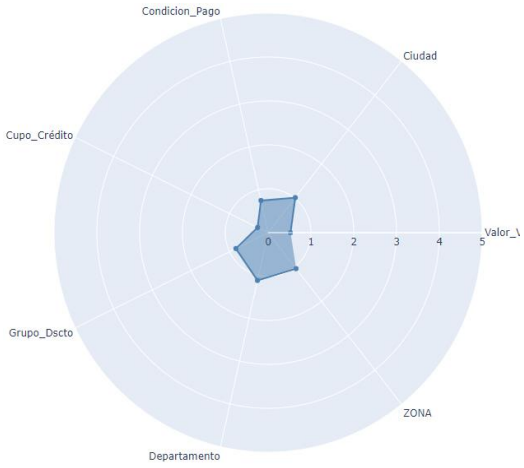


Imagen 188: Modelo Jerárquico 15% – Clúster 18

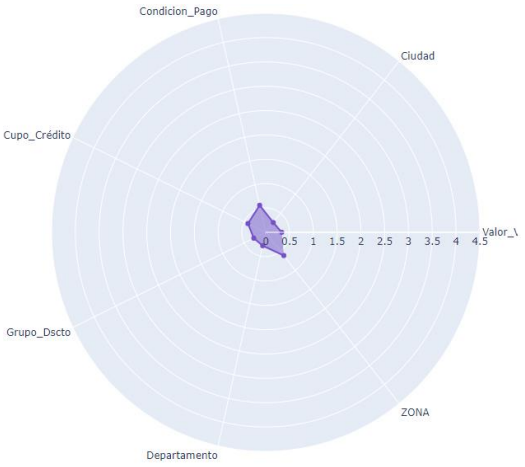


Imagen 189: Modelo *K-means* 100% – Clúster 19

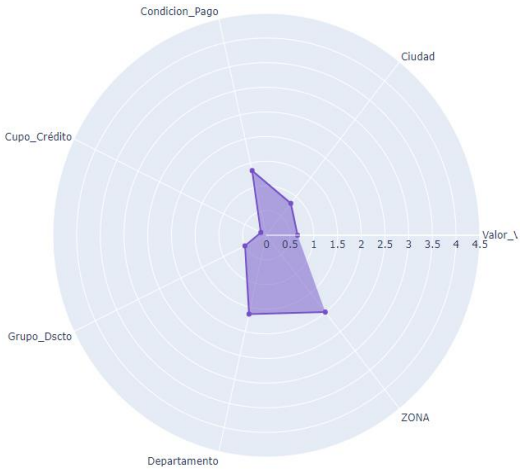


Imagen 190: Modelo Jerárquico 100% – Clúster 19

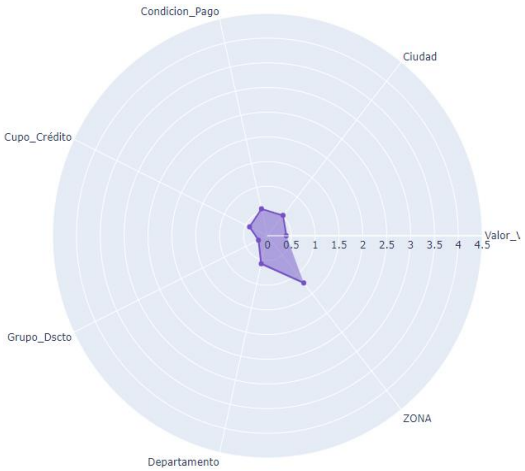


Imagen 191: Modelo *K-means* 15% – Clúster 19

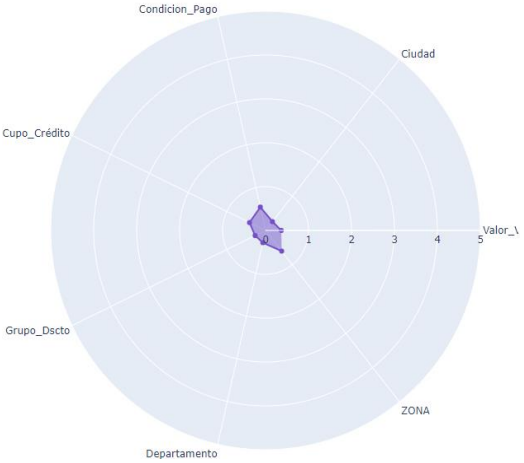


Imagen 192: Modelo Jerárquico 15% – Clúster 19

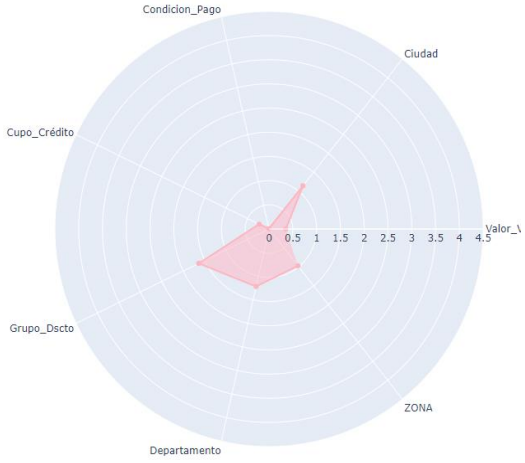


Imagen 193: Modelo *K-means* 100% – Clúster 20

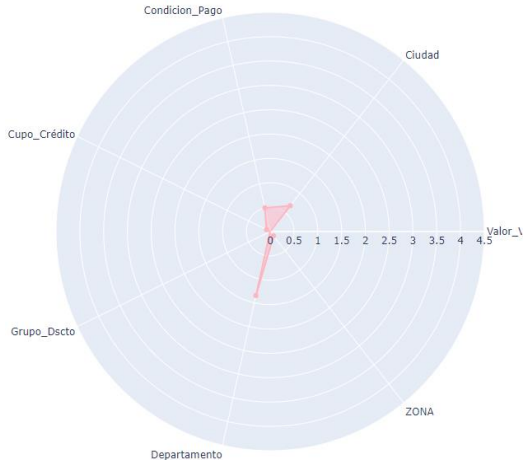


Imagen 194: Modelo Jerárquico 100% – Clúster 20

Imagen 195: Modelo *K-means* 15% – Clúster 20

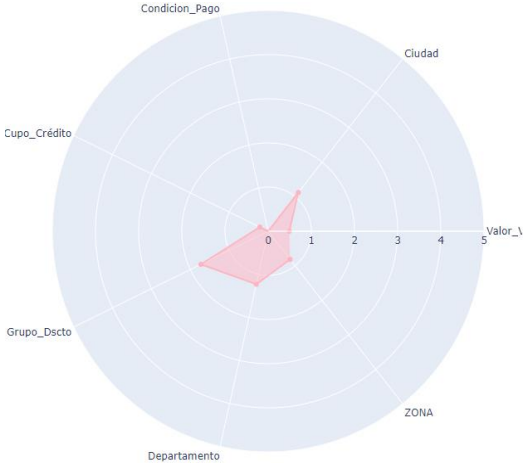


Imagen 196: Modelo Jerárquico 15% – Clúster 20

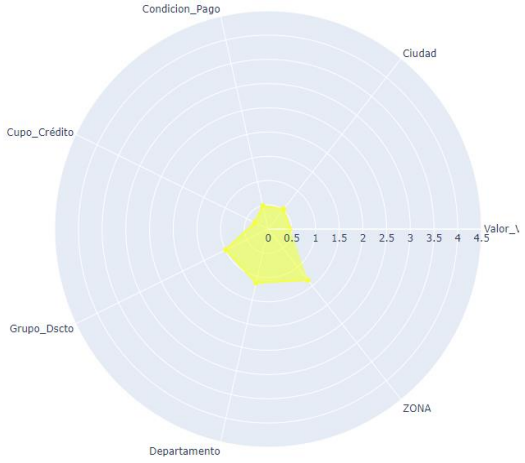


Imagen 197: Modelo *K-means* 100% – Clúster 21

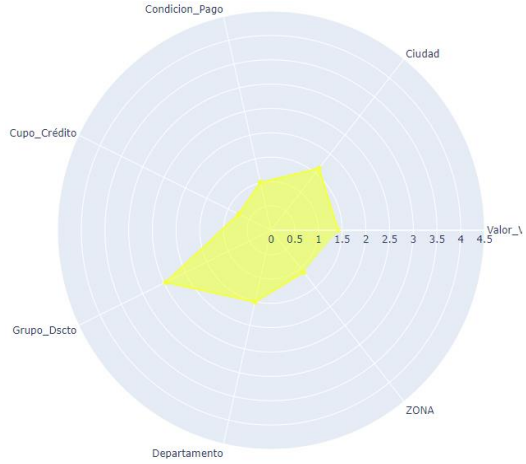


Imagen 198: Modelo Jerárquico 100% – Clúster 21

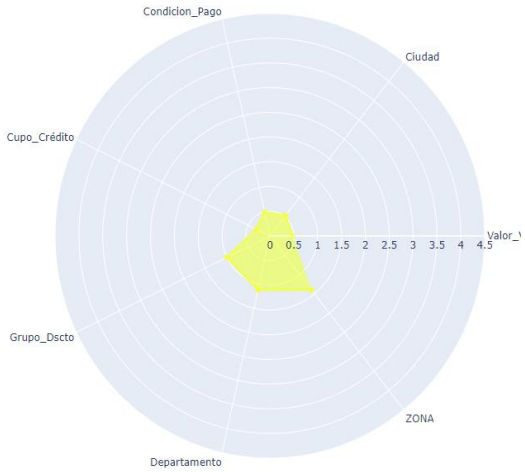


Imagen 199: Modelo *K-means* 15% – Clúster 21

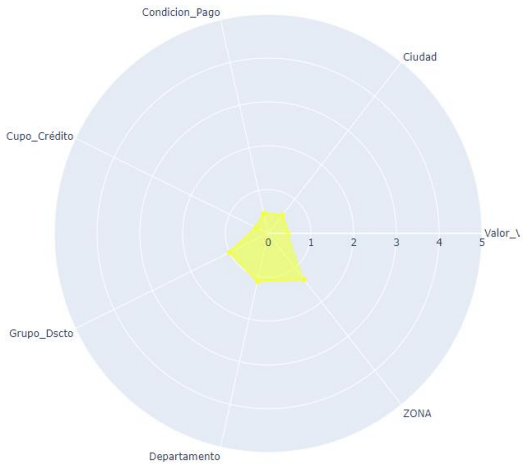


Imagen 200: Modelo Jerárquico 15% – Clúster 21

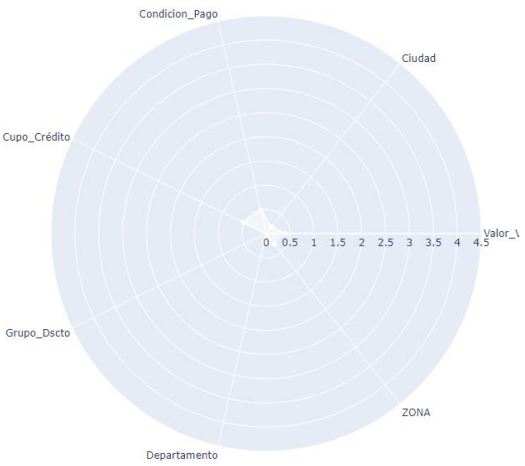


Imagen 201: Modelo *K-means* 100% – Clúster 22

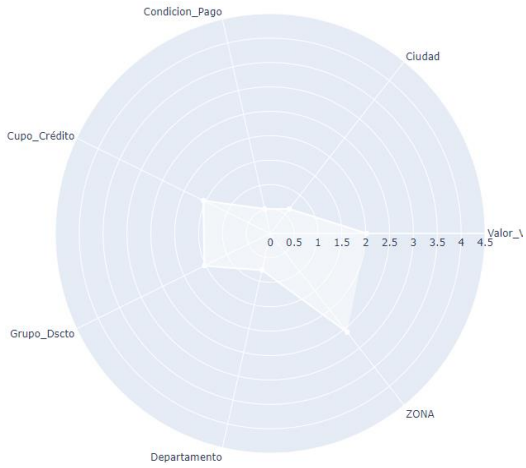


Imagen 202: Modelo Jerárquico 100% – Clúster 22

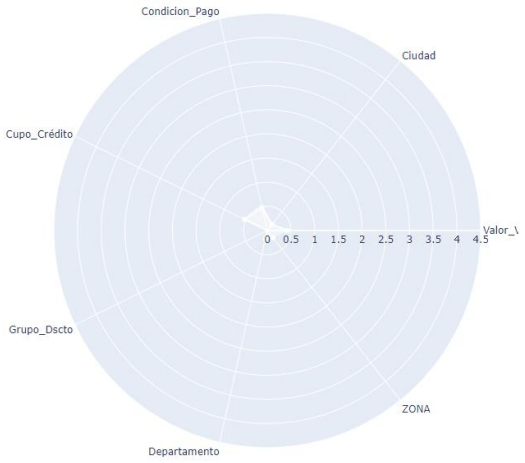


Imagen 203: Modelo *K-means* 15% – Clúster 22

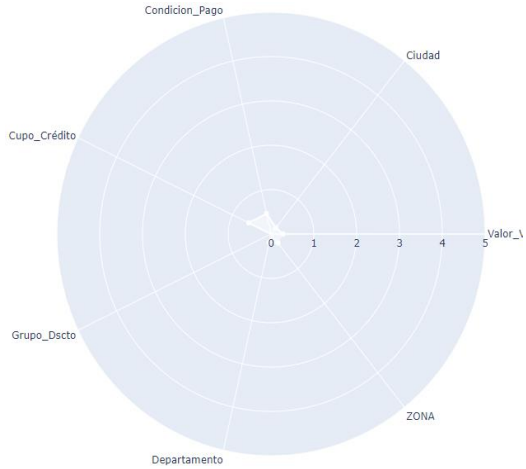


Imagen 204: Modelo Jerárquico 15% – Clúster 22

ANEXO 2. Gráfico de Radar del *dataset* Demográfico.

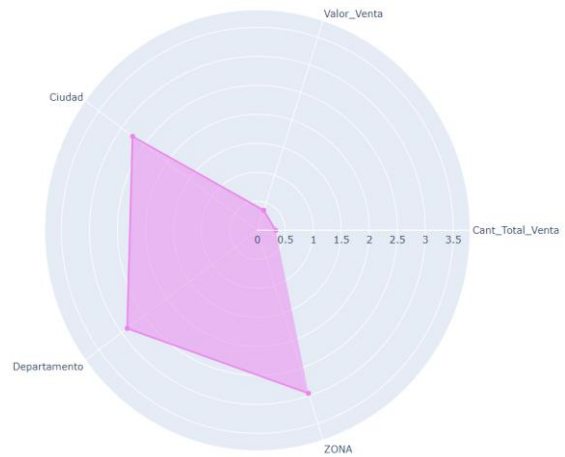


Imagen 205: Modelo *K-means* 100% – Clúster 0

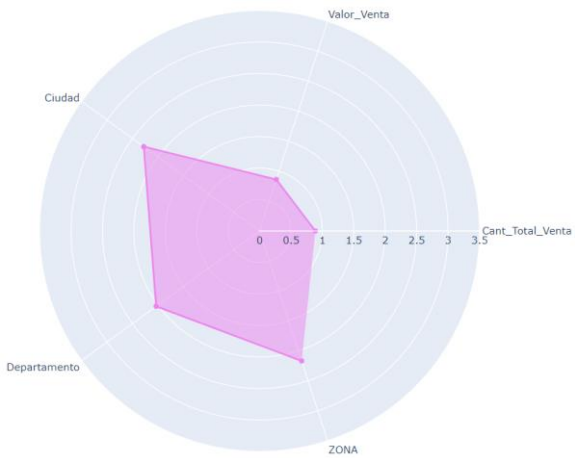


Imagen 206: Modelo Jerárquico 100% – Clúster 0

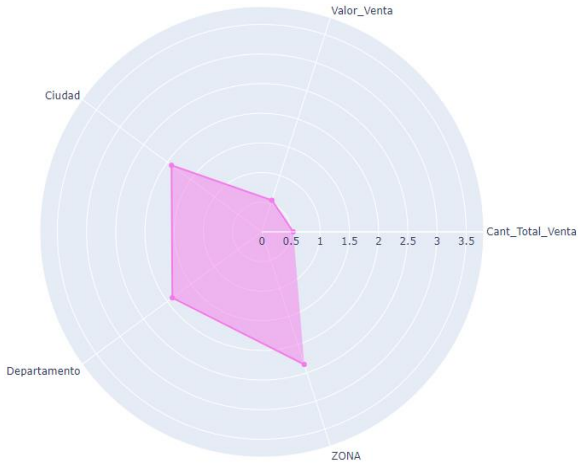


Imagen 207: Modelo *K-means* 15% – Clúster 0

Imagen 208: Modelo Jerárquico 15% – Clúster 0

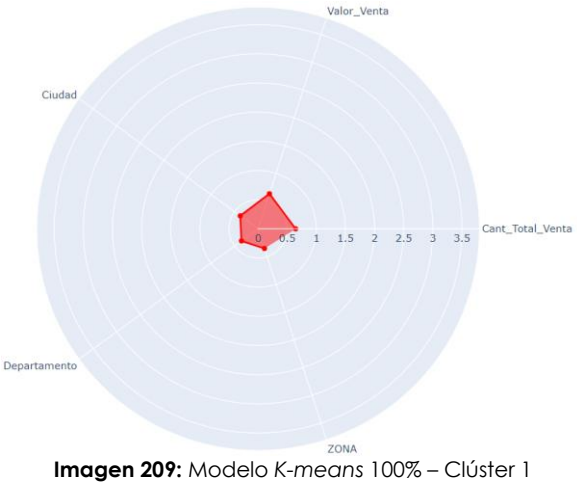


Imagen 209: Modelo *K-means* 100% – Clúster 1

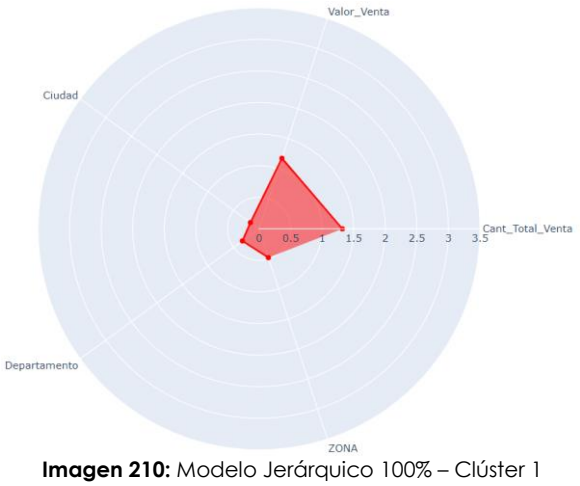


Imagen 210: Modelo Jerárquico 100% – Clúster 1

Imagen 211: Modelo *K-means* 15% – Clúster 1

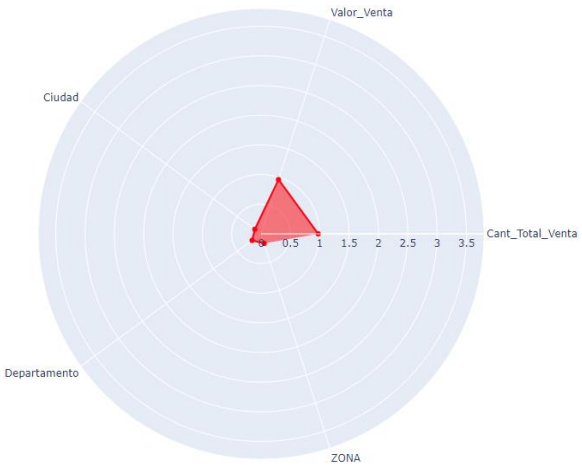


Imagen 212: Modelo Jerárquico 15% – Clúster 1

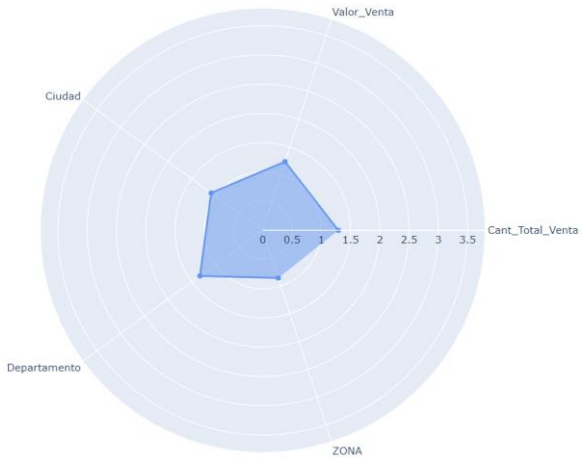


Imagen 213: Modelo *K-means* 100% – Clúster 2

Imagen 214: Modelo Jerárquico 100% – Clúster 2

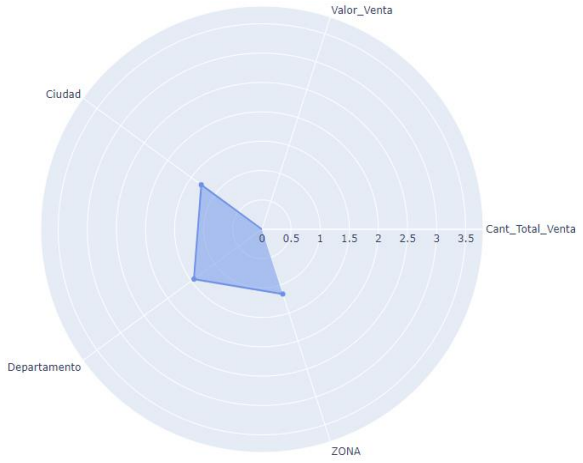


Imagen 215: Modelo *K-means* 15% – Clúster 2

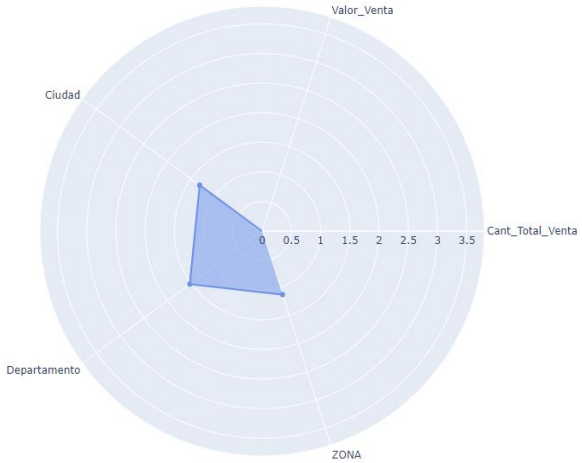


Imagen 216: Modelo Jerárquico 15% – Clúster 2

Imagen 217: Modelo *K-means* 100% – Clúster 3

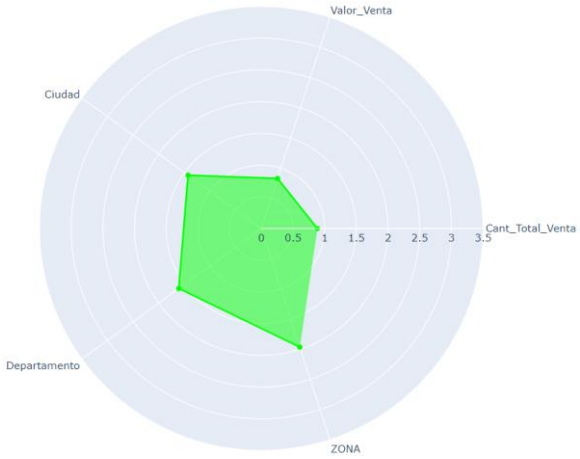


Imagen 218: Modelo Jerárquico 100% – Clúster 3

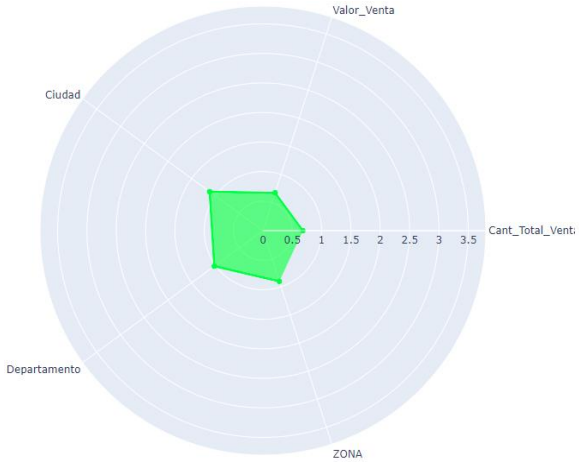


Imagen 219: Modelo *K-means* 15% – Clúster 3

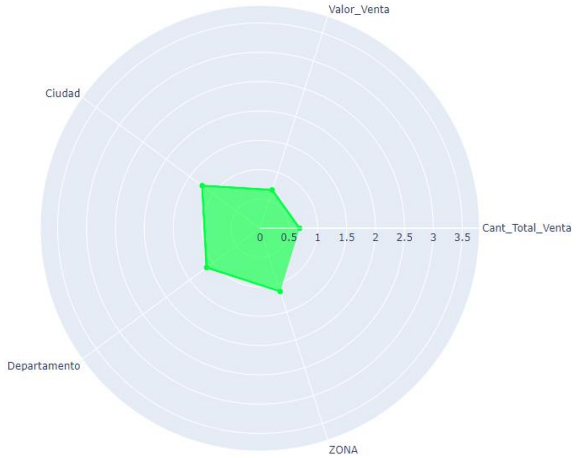


Imagen 220: Modelo Jerárquico 15% – Clúster 3

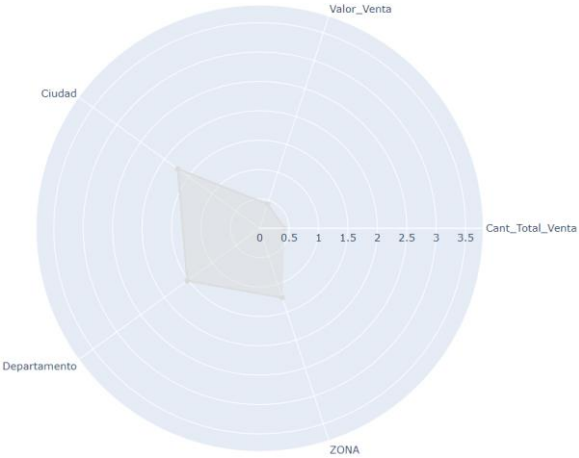


Imagen 221: Modelo *K-means* 100% – Clúster 5

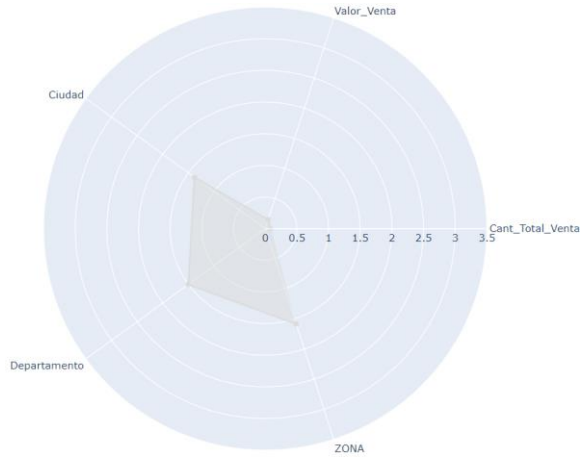


Imagen 222: Modelo Jerárquico 100% – Clúster 5

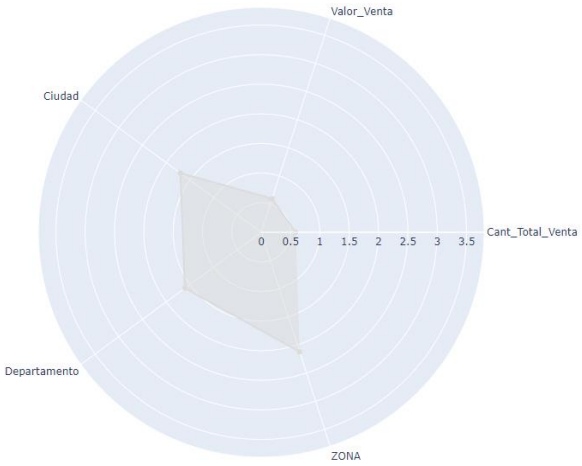


Imagen 223: Modelo *K-means* 15% – Clúster 5

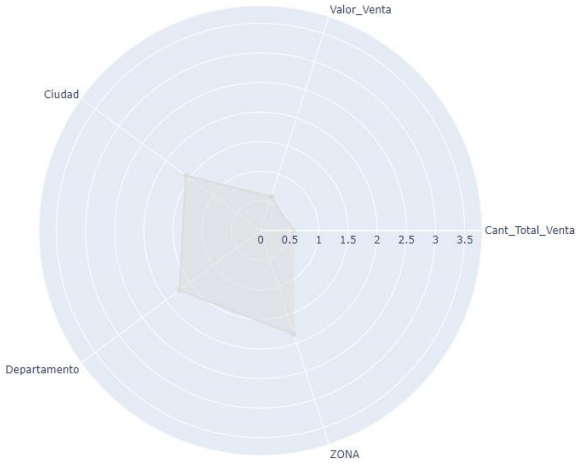


Imagen 224: Modelo Jerárquico 15% – Clúster 5

Imagen 225: Modelo *K-means* 100% – Clúster 6

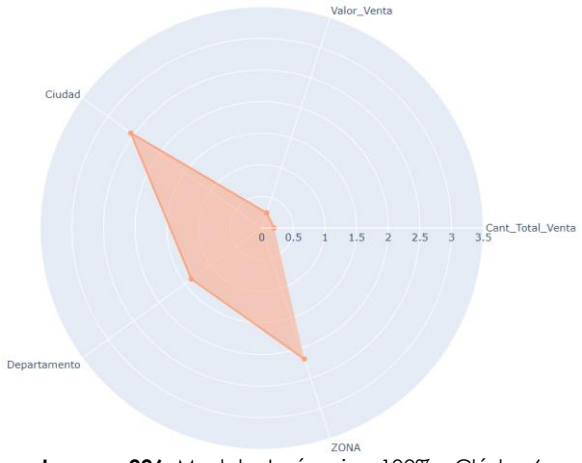


Imagen 226: Modelo Jerárquico 100% – Clúster 6

Imagen 227: Modelo *K-means* 15% – Clúster 6

Imagen 228: Modelo Jerárquico 15% – Clúster 6

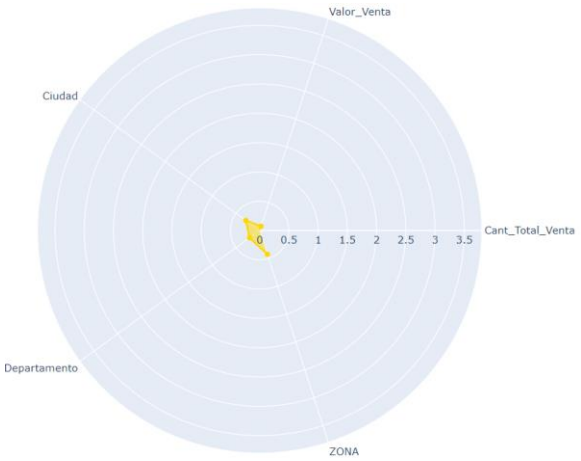


Imagen 229: Modelo *K-means* 100% – Clúster 7

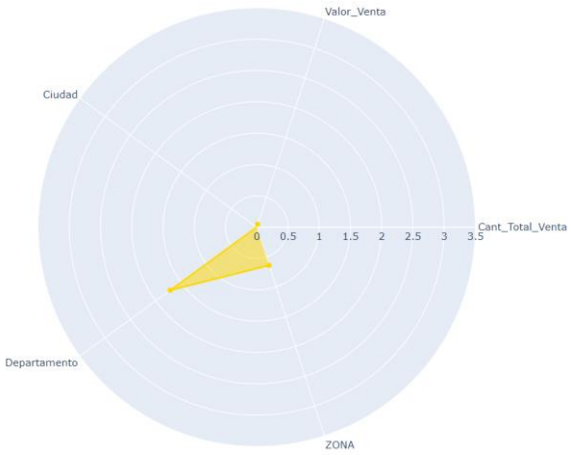


Imagen 230: Modelo Jerárquico 100% – Clúster 7

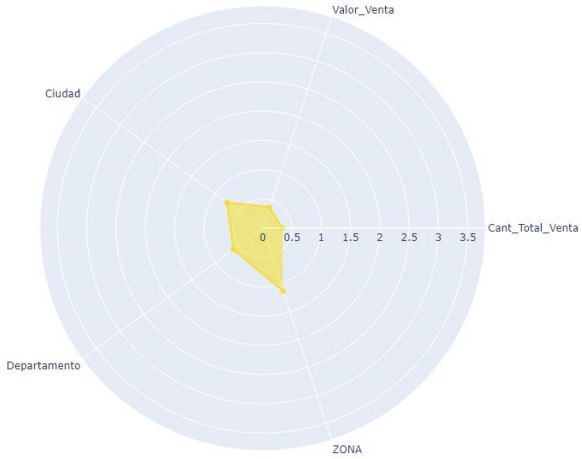


Imagen 231: Modelo *K-means* 15% – Clúster 7

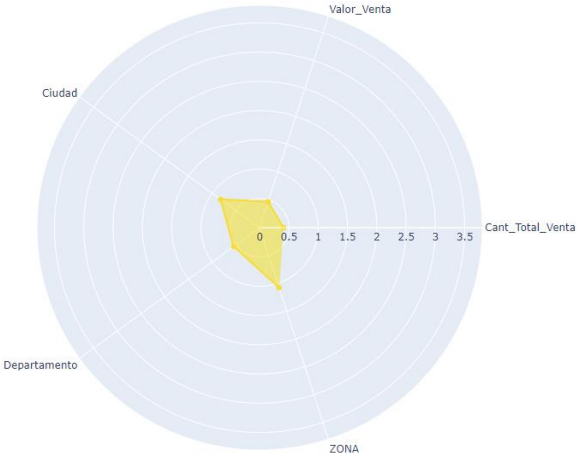


Imagen 232: Modelo Jerárquico 15% – Clúster 7

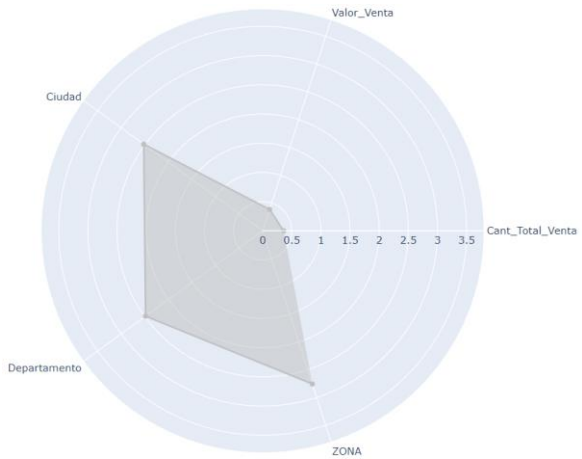


Imagen 233: Modelo *K-means* 100% – Clúster 8

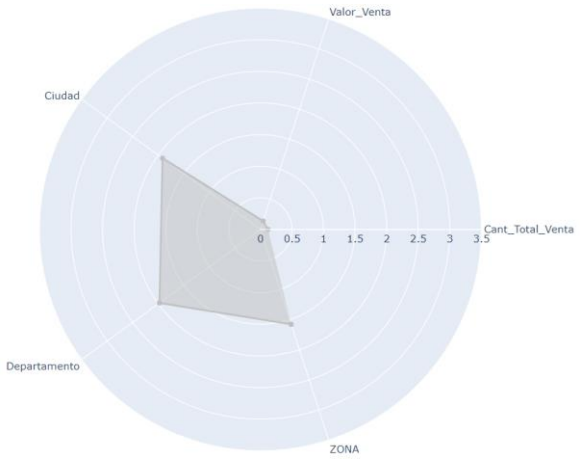


Imagen 234: Modelo Jerárquico 100% – Clúster 8

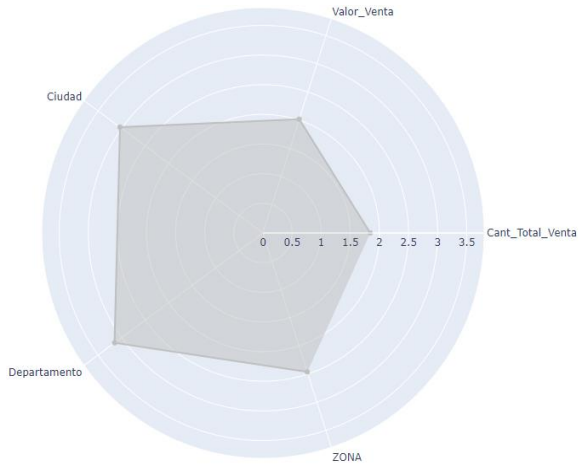


Imagen 235: Modelo *K-means* 15% – Clúster 8

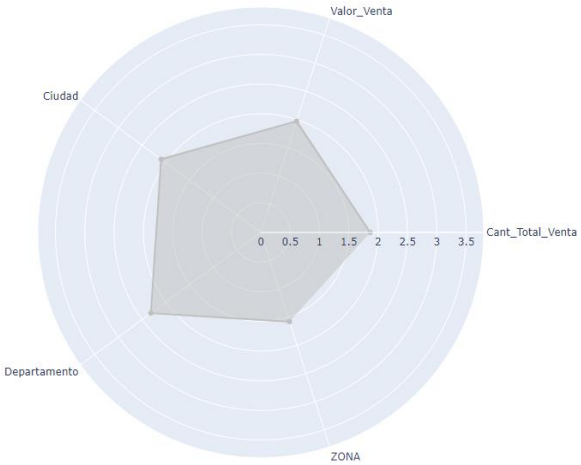


Imagen 236: Modelo Jerárquico 15% – Clúster 8

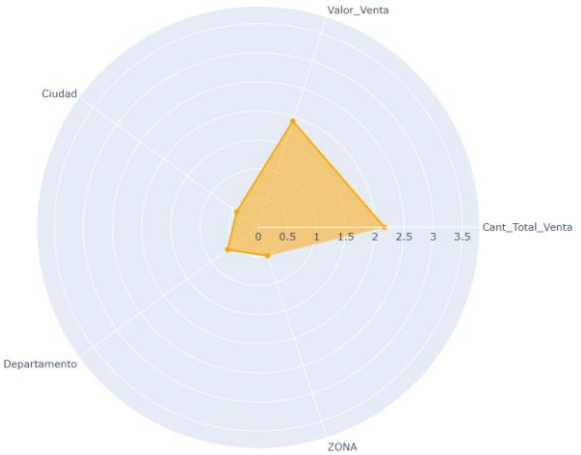


Imagen 237: Modelo *K-means* 100% – Clúster 9

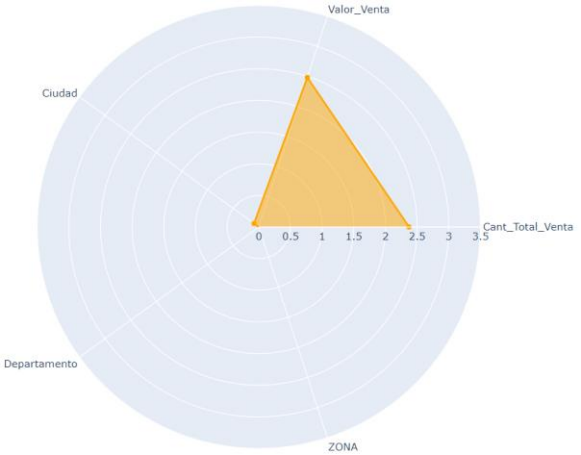


Imagen 238: Modelo Jerárquico 100% – Clúster 9

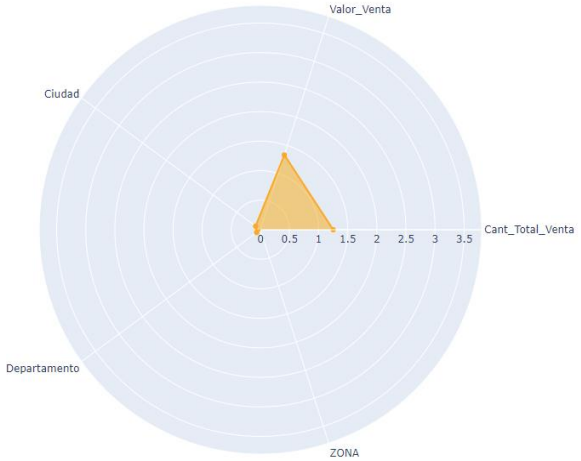


Imagen 239: Modelo *K-means* 15% – Clúster 9

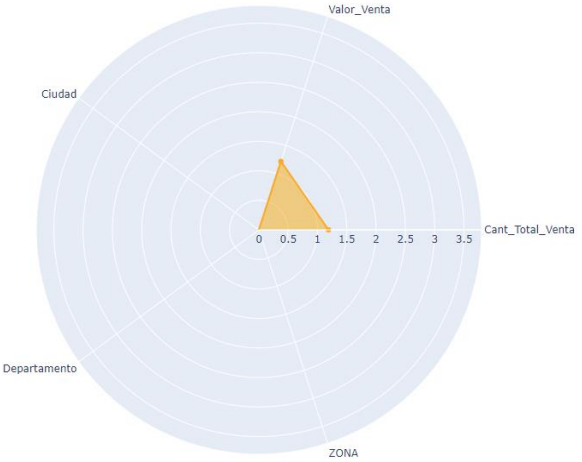


Imagen 240: Modelo Jerárquico 15% – Clúster 9

ANEXO 3. Gráfico de Radar del *dataset* Cosméticos

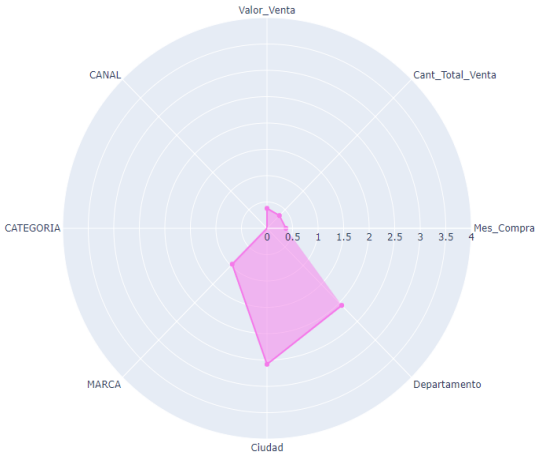


Imagen 241: Modelo *K-means* 100% – Clúster 0

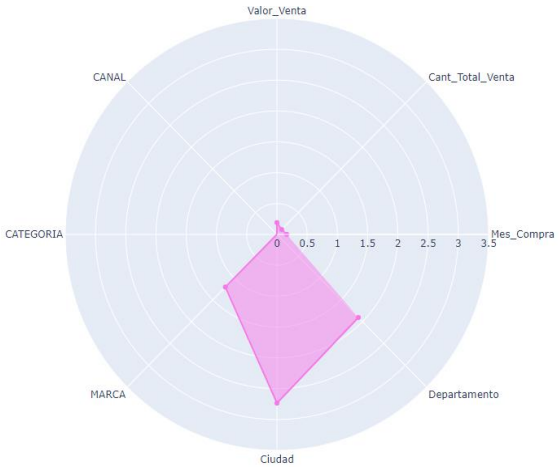


Imagen 242: Modelo Jerárquico 100% – Clúster 0

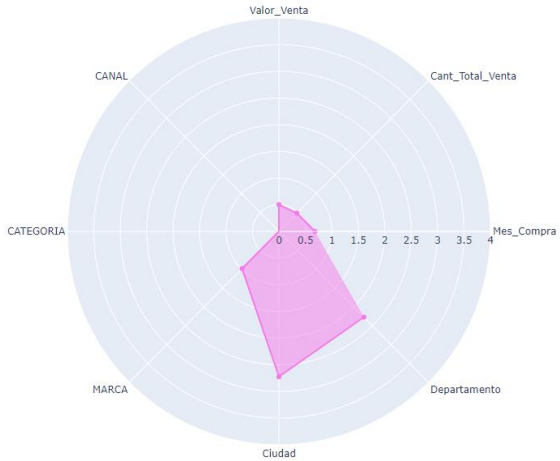


Imagen 243: Modelo *K-means* 35% – Clúster 0

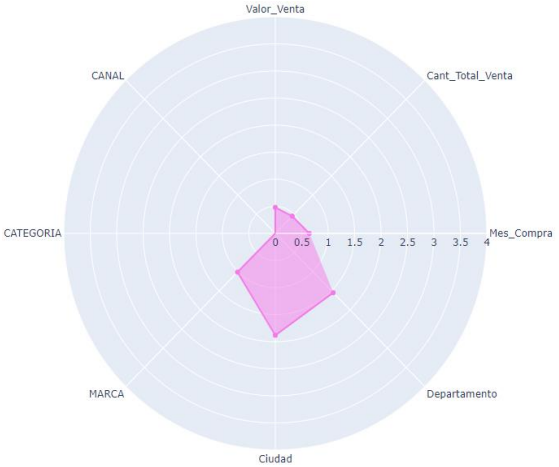


Imagen 244: Modelo Jerárquico 35% – Clúster 0

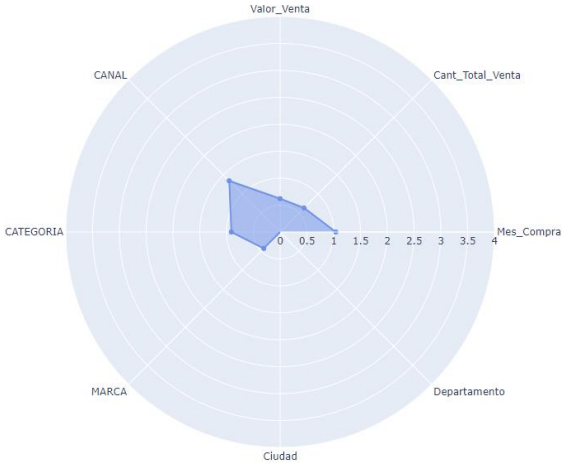


Imagen 245: Modelo *K-means* 100% – Clúster 1

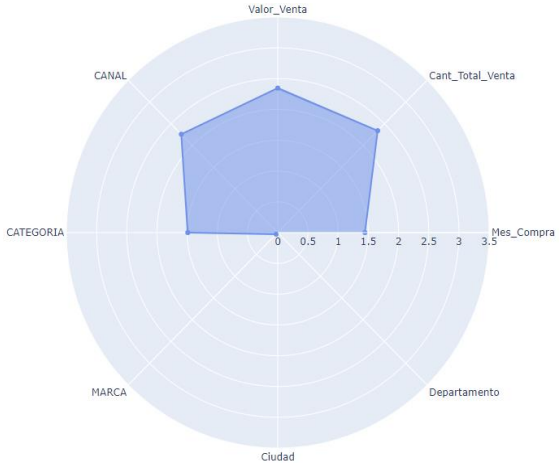


Imagen 246: Modelo Jerárquico 100% – Clúster 1

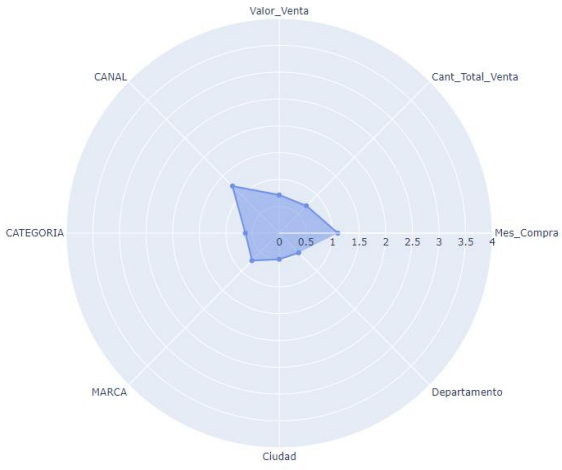


Imagen 247: Modelo *K-means* 35% – Clúster 1

Imagen 248: Modelo Jerárquico 35% – Clúster 1

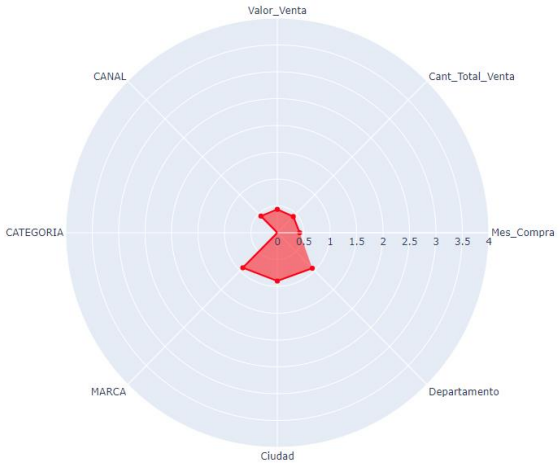


Imagen 249: Modelo *K-means* 100% – Clúster 2

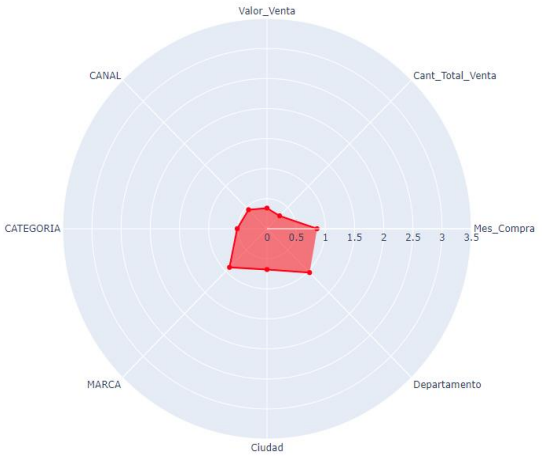


Imagen 250: Modelo Jerárquico 100% – Clúster 2

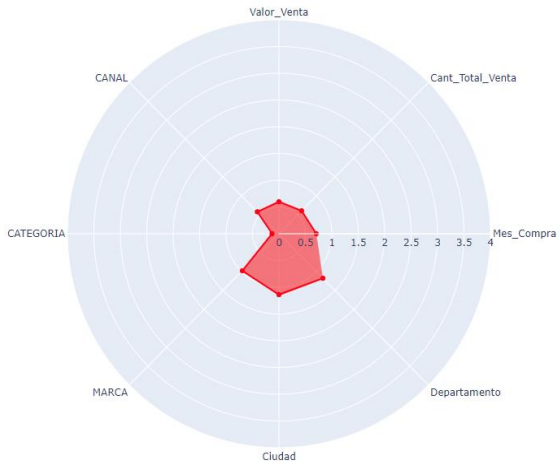


Imagen 251: Modelo *K-means* 35% – Clúster 2

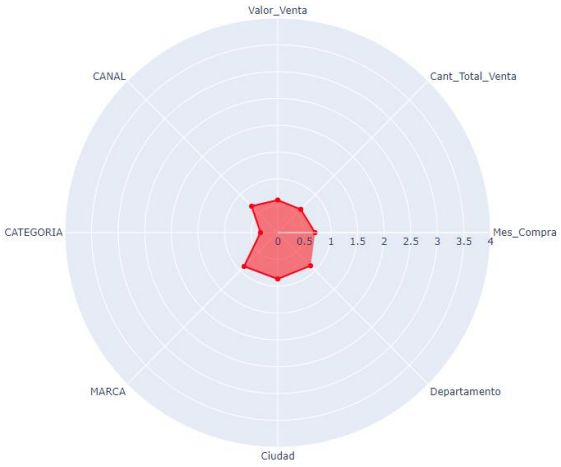


Imagen 252: Modelo Jerárquico 35% – Clúster 2

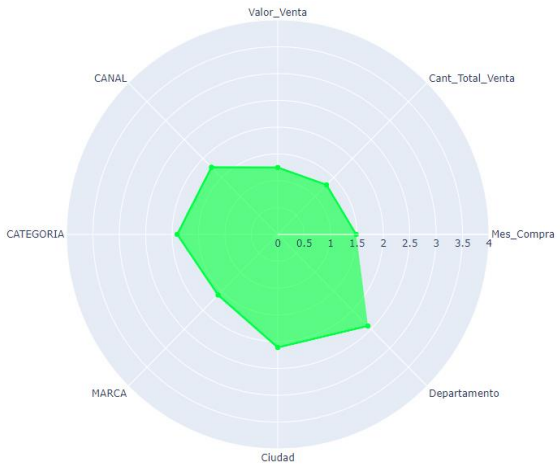


Imagen 253: Modelo *K-means* 100% – Clúster 3

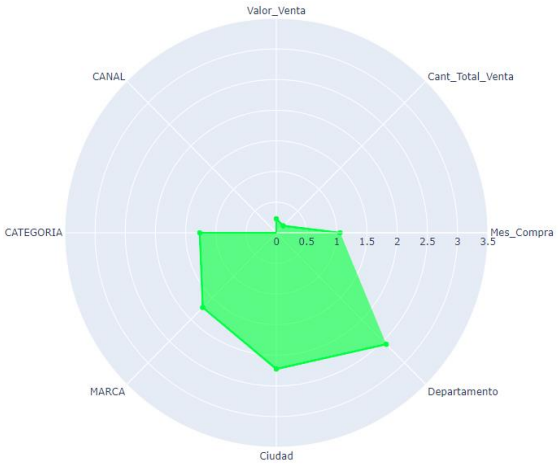


Imagen 254: Modelo Jerárquico 100% – Clúster 3

Imagen 255: Modelo *K-means* 35% – Clúster 3

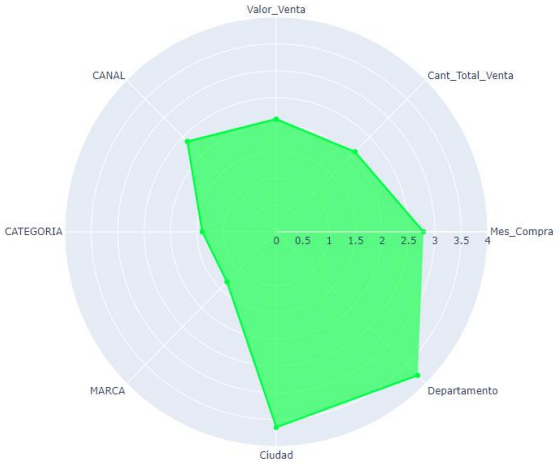


Imagen 256: Modelo Jerárquico 35% – Clúster 3

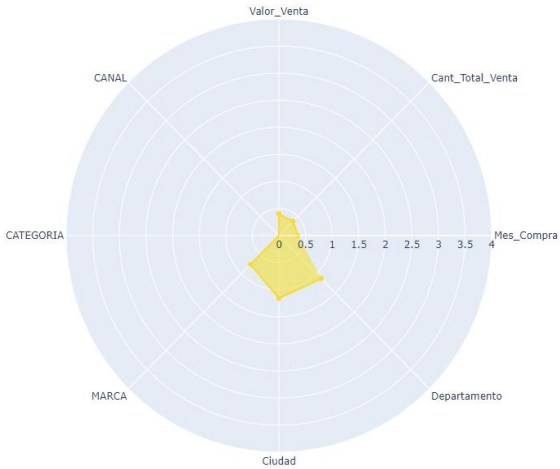


Imagen 257: Modelo *K-means* 100% – Clúster 7

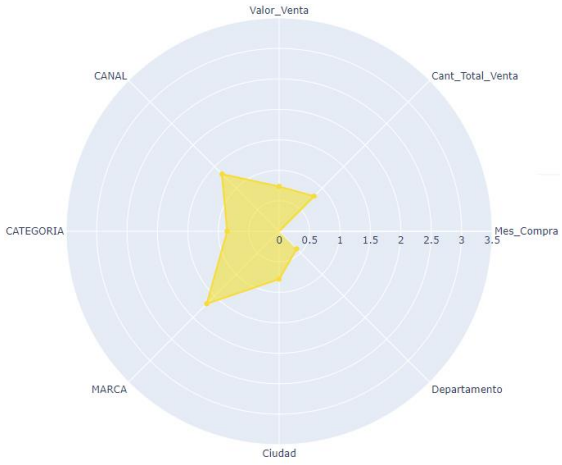


Imagen 258: Modelo Jerárquico 100% – Clúster 7

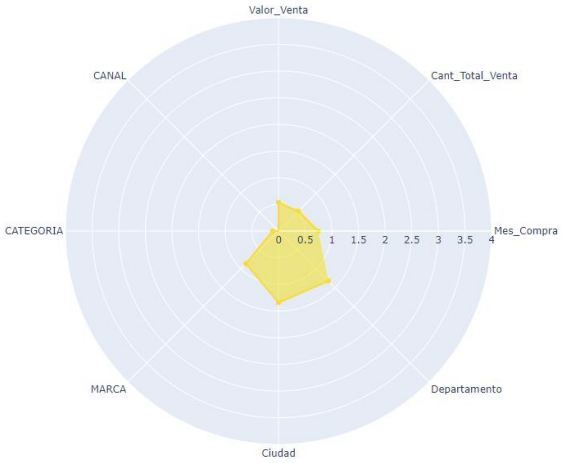


Imagen 259: Modelo *K-means* 35% – Clúster 7

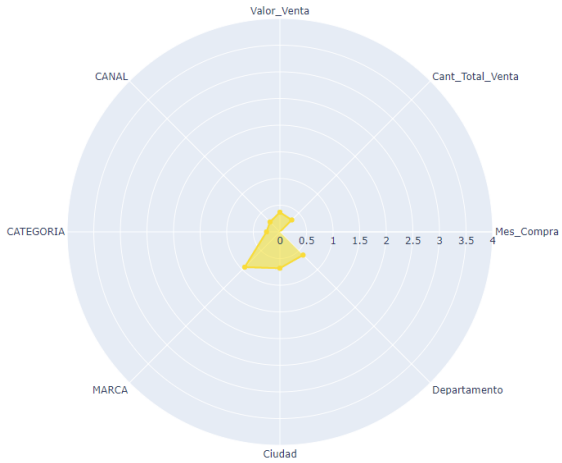


Imagen 260: Modelo Jerárquico 35% – Clúster 7

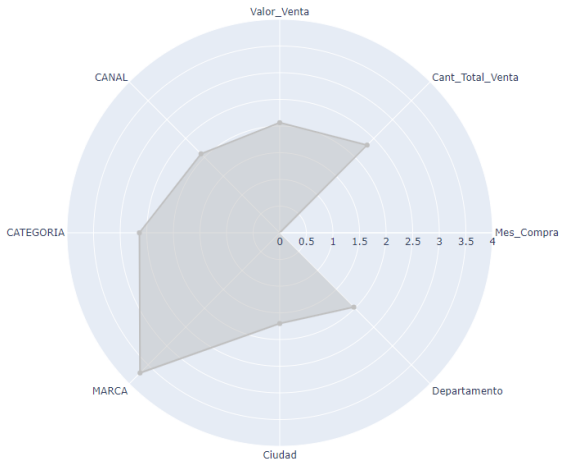


Imagen 261: Modelo *K-means* 100% – Clúster 8

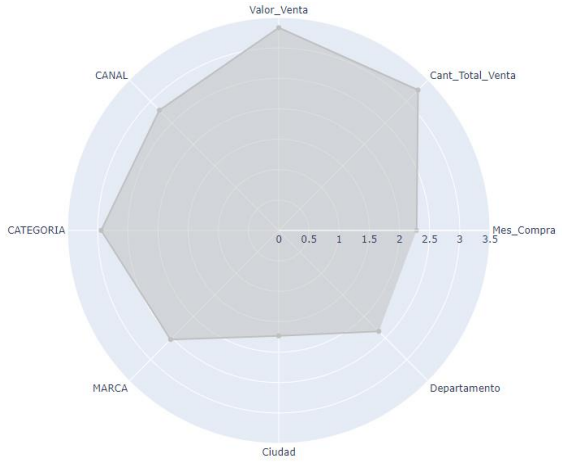


Imagen 262: Modelo Jerárquico 100% – Clúster 8

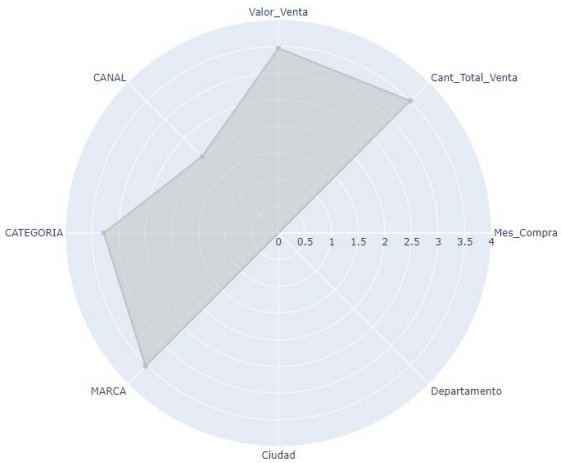


Imagen 263: Modelo *K-means* 35% – Clúster 8

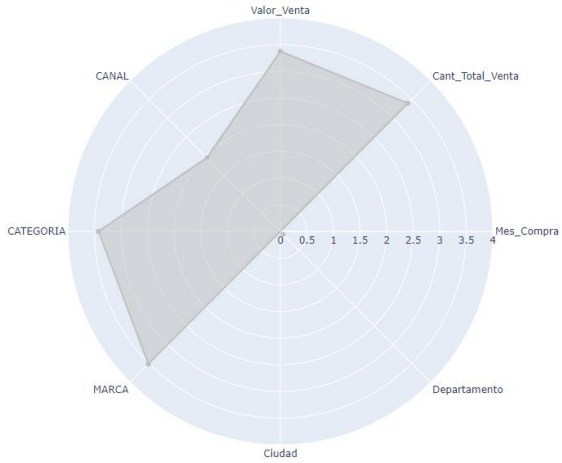


Imagen 264: Modelo Jerárquico 35% – Clúster 8

Imagen 265: Modelo *K-means* 100% – Clúster 9

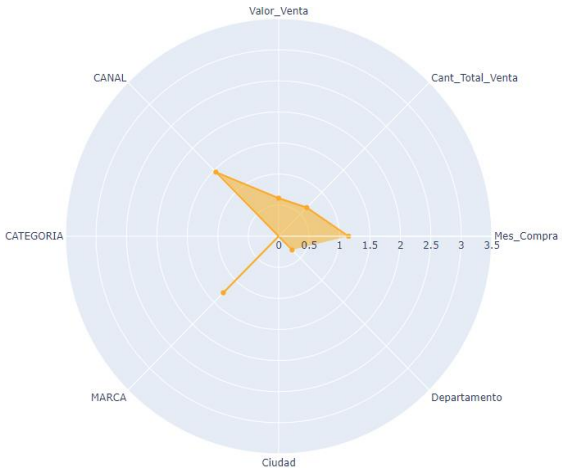


Imagen 266: Modelo Jerárquico 100% – Clúster 9

Imagen 267: Modelo *K-means* 35% – Clúster 9

Imagen 268: Modelo Jerárquico 35% – Clúster 9

ANEXO 4. Gráfico de Radar del *dataset* Medicamentos.

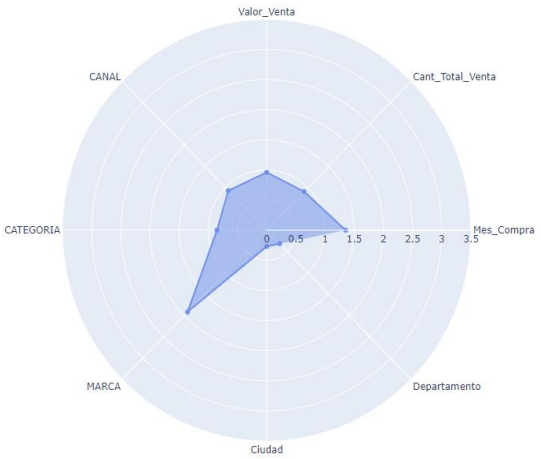


Imagen 269: Modelo *K-means* – Clúster 1

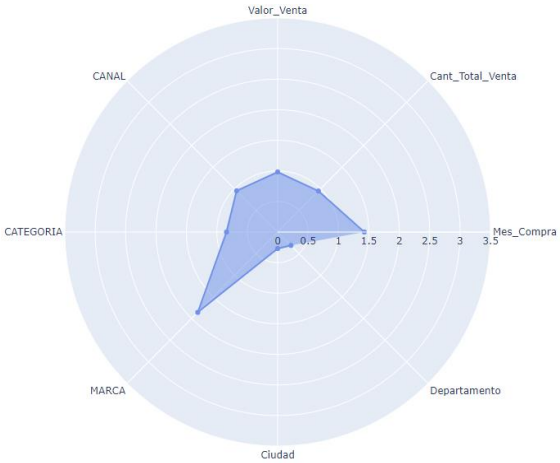


Imagen 270: Modelo Jerárquico – Clúster 1

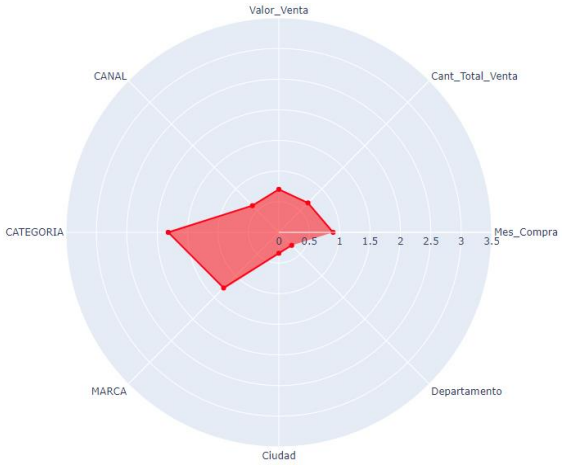


Imagen 271: Modelo *K-means* – Clúster 2

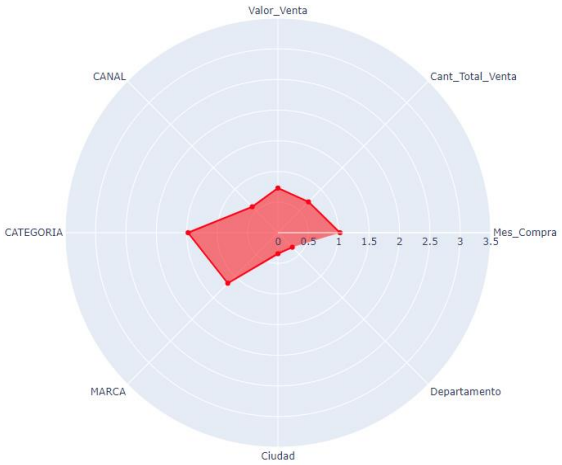


Imagen 272: Modelo Jerárquico – Clúster 2

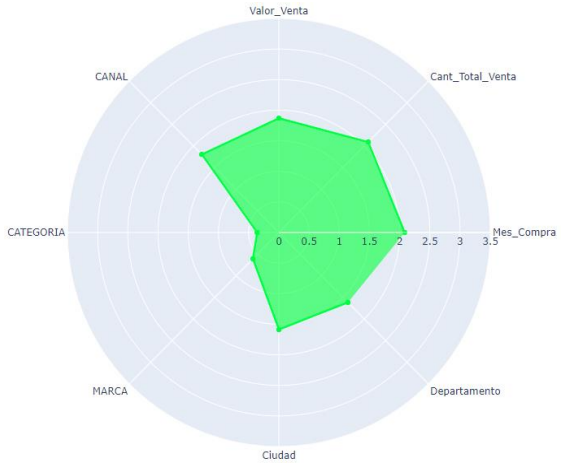


Imagen 273: Modelo *K-means* – Clúster 3

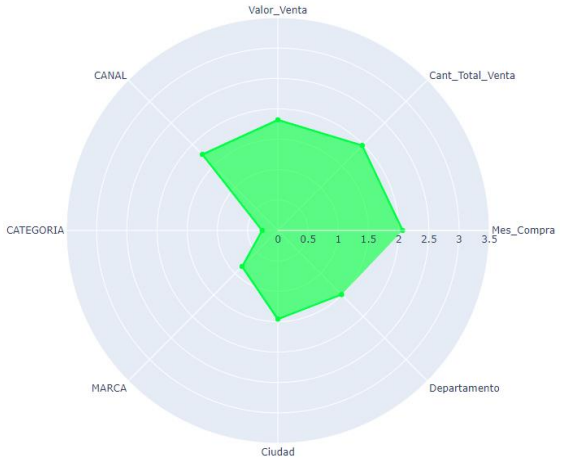


Imagen 274: Modelo Jerárquico – Clúster 3

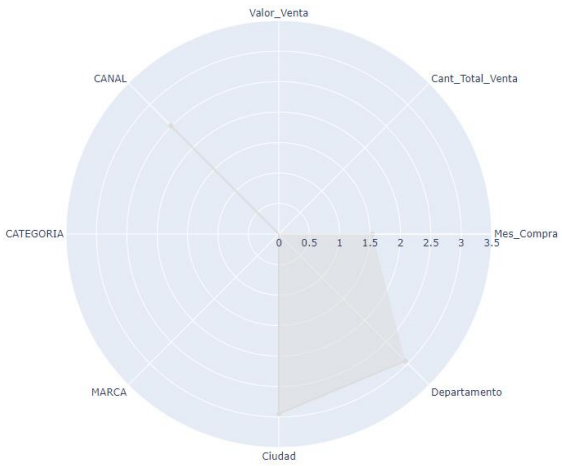


Imagen 275: Modelo *K-means* – Clúster 5

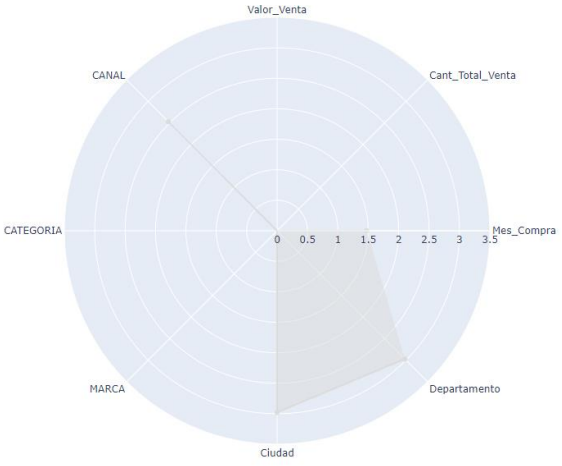


Imagen 276: Modelo Jerárquico – Clúster 5

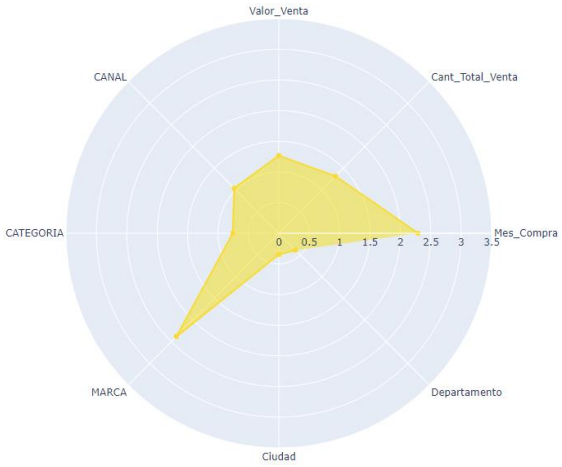


Imagen 277: Modelo *K-means* – Clúster 7

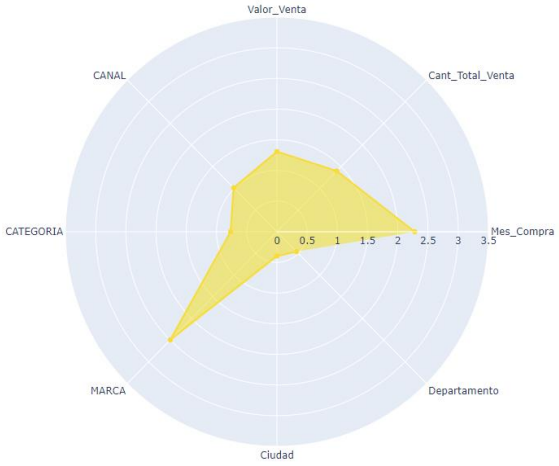


Imagen 278: Modelo Jerárquico – Clúster 7

Imagen 279: Modelo *K-means* – Clúster 8

Imagen 280: Modelo Jerárquico – Clúster 8

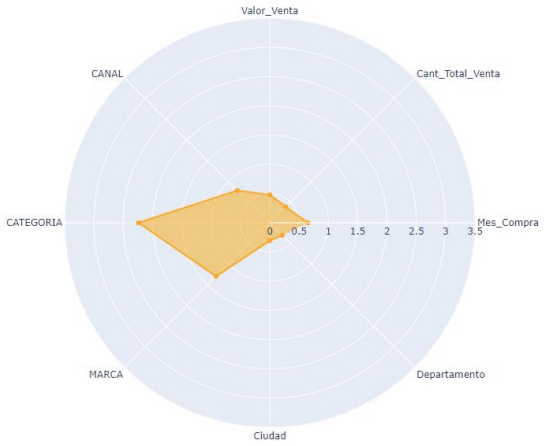


Imagen 281: Modelo *K-means* – Clúster 9

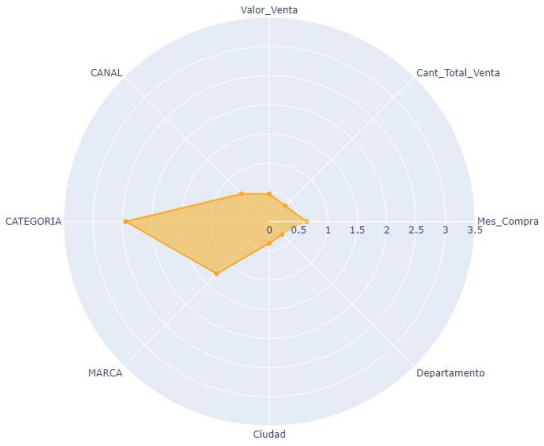


Imagen 282: Modelo Jerárquico – Clúster 9

Anexo 1. LICENCIA DE AUTORIZACIÓN PARA LA PUBLICACIÓN DE OBRAS EN VITELA REPOSITORIO INSTITUCIONAL DE LA UNIVERSIDAD JAVERIANA CALI

Señores
Biblioteca General
Pontificia Universidad Javeriana Cali
Cuidad

Por medio del presente documento otorgo (otorgamos) a la Pontificia Universidad Javeriana Cali para que, en perfeccionamiento de la siguiente licencia de uso parcial[1], pueda ejercer sobre mi (nuestra) obra las facultades que se indican a continuación, teniendo en cuenta que, en cualquier caso, la finalidad perseguida será facilitar, difundir y promover el aprendizaje, la enseñanza y la investigación.

En consecuencia, autorizo (autorizamos) a la Pontificia Universidad Javeriana Cali, a los usuarios de la Biblioteca General, así como a los usuarios de las redes, bases de datos y demás sitios web con los que la Universidad tenga perfeccionado un convenio o con los que establezcan redes de colaboración, son:

AUTORIZO (AUTORIZAMOS)
1. La conservación de los ejemplares necesarios
2. La consulta en línea
3. La reproducción y/o transformación por cualquier formato conocido o por conocer. Igualmente, la edición, o cualquier otra forma de reproducción, incluyendo la posibilidad de trasladarla al sistema o entorno digital.
4. La difusión y comunicación pública por cualquier medio físico o digital, así como su disposición en Internet a través de índices, buscadores y otros medios conocidos o por conocer.
5. La publicación en bases de datos y en sitios web, sean éstos onerosos o gratuitos, existiendo con ellos previo acuerdo desarrollado con la Pontificia Universidad Javeriana para efectos de cumplir los fines predichos.
6. La inclusión en el repositorio institucional de la Pontificia Universidad Javeriana Cali.
7. La inclusión en cualquier otro formato digital, físico o soporte como multimedia, colecciones, recopilaciones o, en general, servir de base para cualquier otra obra derivada.

[1] *De conformidad con lo establecido en el artículo 30 de la Ley 23 de 1982 y el artículo 11 de la Decisión Andina 351 de 1993, "Los derechos morales sobre el trabajo son propiedad de los autores", los cuales son irrenunciables, imprescriptibles, inembargables e inalienables. En consecuencia, la Pontificia Universidad Javeriana está en la obligación de respetarlos y hacerlos respetar, para lo cual tomará las medidas correspondientes para garantizar su observancia.*

La presente licencia parcial se otorga a título gratuito por el máximo tiempo legal colombiano, con el propósito de que en dicho lapso mi (nuestra) obra sea utilizada en las condiciones aquí concertadas y para los fines indicados, respetando siempre la titularidad de los derechos patrimoniales y morales correspondientes, de acuerdo con los usos honrados, de manera proporcional y justificada a la finalidad perseguida, sin ánimo de lucro ni de comercialización.

Si autorizo la publicación: ____X____

No autorizo la publicación: _____

De manera complementaria, garantizo (garantizamos) en mi (nuestra) calidad de autor (es) exclusivo (s), que la obra en cuestión, es producto de mi (nuestra) plena autoría, de mi (nuestro) esfuerzo personal intelectual, como consecuencia de mi (nuestra) creación original particular y, por tanto, soy (somos) el (los) único (s) titular (es) de la misma. Además, aseguro (aseguramos) que no contiene citas, ni transcripciones de otras obras protegidas, por fuera de los límites autorizados por la ley, según los usos honrados, y en proporción a los fines previstos; ni tampoco contempla declaraciones difamatorias contra terceros; respetando el derecho a la imagen, intimidad, buen nombre y demás derechos constitucionales. Adicionalmente, manifiesto (manifestamos) que no se incluyeron expresiones contrarias al orden público ni a las buenas costumbres. En consecuencia, la responsabilidad directa en la elaboración, presentación, investigación y, en general, contenidos de la obra es de mí (nuestro) competencia exclusiva, eximiendo de toda responsabilidad a la Pontifica Universidad Javeriana Cali por tales aspectos.

Si el documento ha sido apoyado o financiado por alguna organización, con excepción de la Pontificia Universidad Javeriana Cali, el (los) autor (es) garantiza (n) que se ha cumplido con las obligaciones requeridas por el respectivo acuerdo.

En constancia a lo anterior,

Título de la obra

SEGMENTACIÓN AUTOMATICA DE LOS CLIENTES DE PHARMADERM Y SKINDRUG, UTILIZANDO TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

NOMBRE COMPLETO	No. Documento Identidad	FIRMA
JOHN DÍAZ ALONSO	80.155. 659	

CORREO ELECTRÓNICO: john.diaz.alonso@outlook.com

FECHA: 12/09/2022