Segmentación de clientes y optimización de su fidelización mediante aprendizaje computacional

Marcos Alcocer Gil

Área 1

Tutor/a de TF
Santiago Rojo Muñoz

Profesor/a responsable de la asignatura
Albert Solé Ribalta

Junio de 2024

Universitat Oberta de Catalunya

Agradecimientos

A Santiago Rojo Muñoz, por su generosidad.

A Rocío Albuixech García, por su incondicionalidad y su ejemplo.

Ficha del Trabajo Final

Título del trabajo:	Segmentación de clientes y optimización de su fidelización mediante aprendizaje computacional
Nombre del autor/a:	Marcos Alcocer Gil
Nombre del Tutor/a de TF:	Santiago Rojo Muñoz
Nombre del/de la PRA:	Albert Solé Ribalta
Fecha de entrega:	06/2024
Titulación o programa:	Máster Universitario en Ciencia de Datos
Área del Trabajo Final:	Área 1
Idioma del trabajo:	Castellano
Palabras clave	Segmentación, fidelización, *machine learning*

Resumen del Trabajo

En un entorno cambiante como el del mercado actual, la capacidad de las empresas de entender los deseos y necesidades de su público resulta crucial para su fidelización y la implementación de una estrategia de venta eficaz. Este trabajo final de máster explora la aplicación de técnicas de aprendizaje computacional, tanto supervisado como no supervisado, en la segmentación y retención de clientes. Además, se analiza la maximización del valor del tique con el objetivo de mejorar la identificación de oportunidades de venta y la personalización de las ofertas. Partiendo de un conjunto de datos de venta en línea de información comercial, de riesgo y financiera de empresas de Colombia, se analiza la relevancia de las diferentes variables para la segmentación empleando técnicas estadísticas y de aprendizaje automático, incluyendo algoritmos de agrupamiento jerárquicos y no jerárquico, y redes neuronales. A continuación, y a través del entrenamiento de modelos predictivos, se identifican posibilidades de compra y se definen estrategias de venta personalizadas. Los resultados obtenidos muestran cómo el empleo de técnicas basadas en datos permite identificar criterios óptimos de segmentación para mejorar la tasa de conversión y la retención. Este trabajo proporciona un sustento teórico de la aplicación del aprendizaje computacional a las técnicas de mercadotecnia, pero también pretende servir como guía práctica de aplicación de estas herramientas en un entorno real para toda clase de negocios.

Abstract

In the changing environment of the current market, the ability of companies to understand the desires and needs of their audience is crucial for customer loyalty and the implementation of an effective sales strategy. This master's thesis explores the application of machine learning techniques, both supervised and unsupervised, in customer segmentation and retention. Additionally, it analyzes the maximization of ticket value with the aim of improving the identification of sales opportunities and the personalization of offers. Using a dataset of online sales and commercial, risk, and financial information from companies in Colombia, the relevance of different variables for segmentation is analyzed using statistical and machine learning techniques, including hierarchical and non-hierarchical clustering algorithms, and neural networks. Subsequently, through the training of predictive models, purchasing possibilities are identified and personalized sales strategies are defined. The results obtained show how the use of data-driven techniques allows for the identification of optimal segmentation criteria to improve conversion rates and retention. This work provides a theoretical foundation for the application of machine learning to marketing techniques, but it also aims to serve as a practical guide for applying these tools in a real environment for all types of businesses.

Índex

Lista de figuras

Lista de tablas

1. Introducción

1.1. Contexto

El presente Trabajo Final de Máster (en adelante, «TFM») analiza un conjunto de datos proveniente de una **empresa dedicada a la información comercial y la inteligencia de negocios** y que está especializada en la recopilación, organización y venta de datos de empresas radicadas en Colombia. Respecto a los datos con que trabajan estas empresas, podemos encontrar información financiera, el historial crediticio, detalles específicos sobre la gestión y propiedad de los negocios, así como otra información relevante para una variedad de propósitos empresariales.

Los informes proporcionados por esta clase de empresas pueden ser así utilizados con objetivos tales como:

- Evaluación crediticia: Ya que, teniendo acceso al historial financiero y el estado de las cuentas de los negocios, las empresas pueden valorar la conveniencia de ofrecerles crédito de acuerdo con su solvencia.
- Mercadotecnia: Al ofrecer un conocimiento más profundo del mercado que permite identificar a posibles clientes o ayudar a personalizar la oferta de los servicios.
- Conocimiento de los socios: Informándose sobre el comportamiento y grado de fiabilidad de sus potenciales proveedores y clientes, las empresas pueden decidir si un posible socio comercial estará a la altura de sus estándares.
- Analítica de datos: Para el descubrimiento de patrones relevantes y perspectivas no evidentes sobre el comportamiento del mercado.

La base usuaria de la empresa llega a ésta a través de diferentes canales y, previa cumplimentación de un formulario de registro, tienen acceso a diferentes productos comerciales menores de carácter gratuito. En caso de que desee acceder a un mayor contenido, el usuario dispone de la opción de contratar un producto y convertirse en un cliente de la empresa, bien a través de una compra puntual, de la compra de un bono para un conjunto de productos o del pago recurrente de una suscripción.

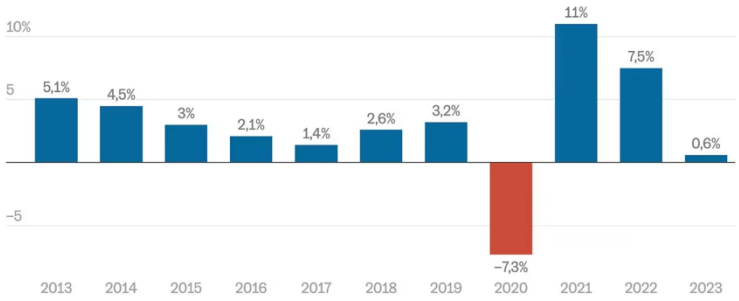


Figura 1. Crecimiento económico en Colombia (Reynoso, 2024)

Pese a que tras la crisis del COVID-19 la **economía de Colombia** experimentó un importante repunte con un crecimiento del 7,3% PIB durante el año 2022, éste vino acompañado de una inflación del 13,1% para el mismo año (Banco Mundial, 2023). Pero a lo largo del año 2023 el país apenas evitó la recesión con un crecimiento de

producto interior bruto de tan sólo el 0,6% (Europa Press, 2024), el menor en una década (Reynoso, 2024). Este dato de crecimiento se encuentra muy por debajo del experimentado tanto a nivel mundial, como en el contexto de América Latina (Banco Mundial, 2023).

La economía colombiana se enfrenta a importantes desafíos estructurales caracterizados por una tasa de crecimiento insuficiente, altos niveles de desigualdad y un elevado índice de deuda (Vargas Riaño, 2024). En el momento de redacción de este TFM el pronóstico de una mejora progresiva es todavía incierto y estará ligado a la inversión en maquinaria y equipo y obra civil, así como al consumo privado (BBVA Research, 2024). Existen, sin embargo, perspectivas de recuperación gradual para toda la región de América Latina y el Caribe, si bien el crecimiento mundial proyectado sugiere una disminución continuada (Banco Mundial, 2023).

	PIB real (%)				
	2021	2022	2023	2024f	2025f
Mundo	**6,2%**	**3,0%**	**2,6%**	**2,4%**	**2,7%**
Economías avanzadas	**5,5%**	**2,5%**	**1,5%**	**1,2%**	**1,6%**
Economías emergentes y en desarrollo	**7,0%**	**3,7%**	**4,0%**	**3,9%**	**4,0%**
Asia oriental y el Pacífico	7,5%	3,4%	5,1%	4,5%	4,4%
Europa y Asia central	7,1%	1,2%	2,7%	2,4%	2,7%
América Latina y el Caribe	7,2%	3,9%	2,2%	2,3%	2,5%
Oriente Medio y Norte de África	3,8%	5,8%	1,9%	3,5%	3,5%
Asia meridional	8,3%	5,9%	5,7%	5,6%	5,9%
África al sur del Sahara	4,4%	3,7%	2,9%	3,8%	4,1%

Figura 2. Crecimiento económico mundial

A pesar de estos datos macroeconómicos, el comercio electrónico en Colombia ha experimentado en los últimos años un crecimiento notable y es actualmente el cuarto más grande en toda América Latina, representando el 3,6% del PIB del país durante el año 2022. En ese año, el número de usuarios de internet aumento del 57% respecto al anterior, alcanzando al 59,5% de los hogares con una distribución desigual entre las áreas urbanes y las rurales. Estas cifras permiten un crecimiento sostenido, con una previsión del 35% hasta el 2027. En el primer trimestre del 2023 las ventas en línea alcanzaron la cifra de 15,1 billones de pesos colombianos, experimentando un crecimiento del 83,2% respecto al mismo período en el 2021. Respecto a su composición, en la actualidad el sector B2B está dominado en un 70% por pymes, mientras que el comercio electrónico C2C y B2G se encuentra en crecimiento (ICEX, 2023).

1.2. Justificación del Trabajo

Una de las principales aspiraciones de cualquier negocio es la de entender el comportamiento del público como medio para crear estrategias personalizadas más efectivas, capaces de satisfacer las necesidades y deseos de su mercado objetivo. Sin embargo, los criterios demográficos utilizados en el pasado para detectar patrones de compra sencillos resultan insuficientes para comprender el complejo comportamiento del público actual, expuesto a una abrumadora cantidad de estímulos. Estas carencias conducen tanto al desaprovechamiento de oportunidades de aumento del tique de venta como, en el peor de los casos, a la pérdida de clientela.

Este TFM aspira a superar las mencionadas limitaciones incorporando técnicas de aprendizaje computacional que permitan obtener una comprensión más profunda del colectivo de consumidoras y consumidores. Así, la implementación de algoritmos

de agrupamiento permite afinar el proceso de segmentación basándose en una amplia gama de variables, ayudando a los negocios a adaptarse a las necesidades y deseos cambiantes de las personas, y a focalizar sus esfuerzos en los grupos más rentables. A su vez, el uso de modelos predictivos en el análisis de la venta posibilita la identificación de patrones y tendencias ocultos, así como la toma de decisiones basadas en datos objetivos, aumentando de esta forma las posibilidades de éxito.

Por todo ello, y en un contexto de intensa competencial empresarial como el actual, la justificación de este TFM es doble:

a) Por un lado, pretende **adaptar** las técnicas de **mercadotecnia** a un nuevo contexto a través del **análisis y modelado de grandes conjuntos de datos** que transformen la información en un conocimiento aplicable, capaz de proporcionar una ventaja competitiva tangible.

b) Por otra parte, aspira a **proporcionar una herramienta práctica** para cualquier empresa que pretenda alinear sus operaciones con la demanda del mercado, mejorando así sus resultados de venta y aumentando la satisfacción de su comunidad usuaria.

1.3. Objetivos del Trabajo

El **objetivo principal** de este proyecto es la segmentación de una cartera de clientes a través de técnicas de aprendizaje computacional.

Aparejadamente, se incluyen los siguientes **objetivos parciales**, tanto preparatorios para la consecución de la segmentación, como derivados de su consecución:

1. Analizar mediante técnicas estadísticas descriptivas la cartera de clientes, poniendo el foco en sus ventas y consumos.
2. Examinar la relación entre las variables del juego de datos, incluyendo su colinealidad e independencia, así como el análisis de la varianza para identificar las variables más relevantes.
3. Valorar la influencia de las variables concurrentes en la segmentación obtenida.
4. Determinar qué parte de la clientela podría beneficiarse de estrategias comerciales encaminadas a:
    a) El *up-selling* o adquisición de un producto de mayor precio.
    b) El *cross-selling* o adquisición de otros productos relacionados.
5. Analizar la recurrencia de la clientela con objeto de establecer estrategias encaminadas a fidelizarla.
6. Implementar una metodología iterativa, tanto al proyecto como al trabajo con los datos, que posibilite la evolución de la solución a partir del aprendizaje.
7. Aplicar las competencias obtenidas durante el Máster Universitario en Ciencia de Datos a un proyecto basado en datos reales.

8. Aprender de la mentoría y experiencia profesional del tutor del proyecto durante su desarrollo.

1.4. Impactos ético-sociales, de sostenibilidad y de diversidad

Este TFM, como parte de unos estudios de máster de la Universitat Oberta de Catalunya (en adelante, UOC), busca su alineación con los Objetivos de Desarrollo Sostenible (en adelante, «**ODS**») de la **Agenda 2030** (UOC, 2023), con la cual está comprometida la institución (Gamez, 2015). Por otra parte, pretende incorporar consideraciones de sostenibilidad, equidad y responsabilidad en todas las fases del proyecto como parte de la competencia de compromiso ético y global (en adelante, **CCEG**) incorporada a estos estudios de máster por la UOC (Doñate, 2020) y vinculada a los ya mencionados ODS.

Las técnicas de segmentación y fidelización de clientes, además de ayudar a las organizaciones a cumplir con sus objetivos comerciales, son potentes herramientas de conocimiento profundo del usuario que permiten ofrecer un producto adaptado a sus necesidades en favor de un consumo responsable. Se encuadran así dentro del **objetivo 12**, centrado en garantizar modalidades de consumo y producción sostenibles (Moran, 2015a, p. 12), al permitir identificar grupos de consumidores con intereses específicos. Entre éstas necesidades y preferencias específicas, se incluyen las basadas en género conforme al **objetivo 5** (Moran, 2015b), lo que la convierte en una herramienta de reducción de la brecha en la satisfacción del cliente.

Todo ello contribuye a fortalecer la relación con los proveedores y la solidificación de cadenas de suministro más sostenibles desde un punto de vista ambiental o de la mejora de las condiciones laborales, encuadrándose dentro del **objetivo 8**, dirigido a promover el crecimiento económico inclusivo y sostenible, el empleo y el trabajo decente para todos (Moran, 2015c, p. 8).

Por otro lado, y con el objetivo de evitar el del **impacto negativo** que este trabajo podría provocar en la sostenibilidad, es esencial una implementación ética y responsable de la solución, asegurando la privacidad y la seguridad de los datos, así como la equidad en el acceso al producto resultante. Igualmente, debemos tener presente la huella de carbono que el consumo energético de los centros de datos provoca durante el entrenamiento de modelos de aprendizaje computacional, incluso estando ésta minimizada en un proyecto de pequeña envergadura como éste.

1.5. Enfoque y método seguido

En lo que respecta a la metodología seguida a lo largo de este TFM, y a pesar de que se trate de enfoques complementarios, cabe distinguir entre la empleada con el objeto de gestionar el proyecto en su totalidad y la centrada de forma específica para el

manejo, procesamiento y análisis de los datos para su comprensión y la generación de modelos.

Para la **gestión del proyecto**, la metodología escogida ha sido *Lean Startup*, la cual basa el desarrollo del producto en la experimentación rápida, la iteración y su adaptación. Así, la idea es construir un MVP (*Minimum Viable Product* o producto mínimo viable) o versión más básica del producto con el objetivo de probar las hipótesis del modelo sobre el mismo para su validación. El resultado de dichos experimentos nos permite aprender e introducir ajustes rápidos durante la siguiente iteración, asegurando la sostenibilidad y mejora continua del producto con el menor esfuerzo posible, o bien pivotar introduciendo modificaciones fundamentales en la hipótesis y el producto.

En cuanto a los **datos**, la metodología aplicada ha sido CRISP-DM (acrónimo de *Cross-Industry Standard Process for Data Mining* o proceso estándar interprofesional para la extracción de datos), al ser el proceso estándar de facto para los proyectos de Minería de Datos, Análisis y Ciencia de Datos. Esta metodología se estructura en un proceso cíclico e iterativo compuesto por 6 **fases**:

1. **Comprensión del negocio**: En la que se establecen los objetivos y requisitos del proyecto, permitiendo definir el problema a resolver y el plan que ha de conducir a su consecución.
2. **Comprensión de los datos**: Durante la cual se recopilan los datos iniciales, a través de los cuales se descubren las primeras ideas y se concreta su capacidad de responder a los objetivos definidos en la fase anterior.
3. **Preparación de los datos**: Centrada en limpiar y transformar los datos brutos para la consecución del conjunto de datos final que será empleado durante la fase de modelización.
4. **Modelización**: Donde se seleccionan las técnicas de modelado adecuadas, se aplican sobre los datos y se afinan sus parámetros a través de un proceso iterativo de construcción y evaluación de los modelos hasta encontrar el óptimo.
5. **Evaluación**: Una vez que se han construido uno o varios modelos, esta fase se focaliza en su evaluación con el propósito de garantizar el cumplimiento de los objetivos empresariales que hemos fijado durante la primera fase. Para ello es necesario evaluar el rendimiento del modelo y asegurarse de que aborda adecuadamente el problema empresarial planteado.
6. **Implementación**: Finalmente, la solución construida se despliega en un entorno operativo.

Una vez completada la totalidad de fases, y con la información obtenida durante la implantación, el proceso iterativo de CRISP-DM nos permite perfeccionar el modelo para mejorar los resultados o readaptarlo a nuevos requisitos y retos empresariales.

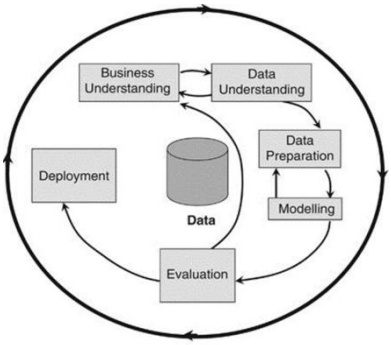


Figura 3. Ciclo de vida de CRISP-DM (Kalaidopoulou, 2017)

1.6. Planificación del trabajo

El siguiente diagrama de Gantt representa el conjunto de tareas a realizar, así como su temporización acorde a la carga de trabajo esperada. La cronología está marcada en gran parte por las dependencias existentes entre las tareas, vinculación que también queda reflejada en el gráfico. Las fases definidas, las cuales se corresponden con las metas parciales, son así las siguientes:

1. Definición y planificación del trabajo final: 22 feb. – 12 mar.
   1.1. Reunión preliminar: 22 feb.
   1.2. Planificación: 28 feb. – 12 mar.
   1.3. Fijación de los objetivos: 28 feb. – 1 mar.
   1.4. Resumen de la propuesta: 2 mar.
   1.5. Definición de la metodología: 3 mar. – 4 mar.
   1.6. Recopilación de la bibliografía: 3 mar. – 12 mar.
   1.7. Redacción de la justificación: 5 mar. – 6 mar.
   1.8. Exposición de la motivación: 7 mar.
   1.9. Incorporación de la CCEG y los ODS: 8 mar. – 10 mar.
   1.10. Elección del título: 11 mar.
   1.11. Elección de palabras clave: 12 mar.
2. Estado del arte: 13 mar. – 26 mar.
   2.1. Búsqueda de fuentes bibliográficas: 13 mar. – 17 mar.
   2.2. Actualización de la bibliografía: 18 mar. – 26 mar.
   2.3. Filtrado de las fuentes de información: 18 mar. – 19 mar.
   2.4. Organización de la información: 20 mar.
   2.5. Redacción del documento: 21 mar. – 26 mar.
3. Diseño e implementación del trabajo: 27 mar. – 21 may.
   3.1. 1ª iteración: 27 mar. – 9 abr.
   3.2. 2ª iteración: 10 abr. – 23 abr.
   3.3. 3ª iteración: 24 abr. – 7 may.
   3.4. 4ª iteración: 8 may. – 21 may.
4. Redacción de la documentación: 22 may. – 18 jun.
   4.1. Memoria del trabajo: primera entrega: 22 may. – 4 jun.
   4.2. Redacción de la memoria: entrega final: 5 jun. – 11 jun.
   4.3. Presentación audiovisual del trabajo: 12 jun. – 18 jun.
5. Defensa del proyecto: 19 jun. – 7 jul.
   5.1. Entrega de la documentación al tribunal: 19 jun. – 20 jun.
   5.2. Preparación de la defensa: 21 jun. – 23 jun.
   5.3. Defensa pública del trabajo: 24 jun. – 7 jul.

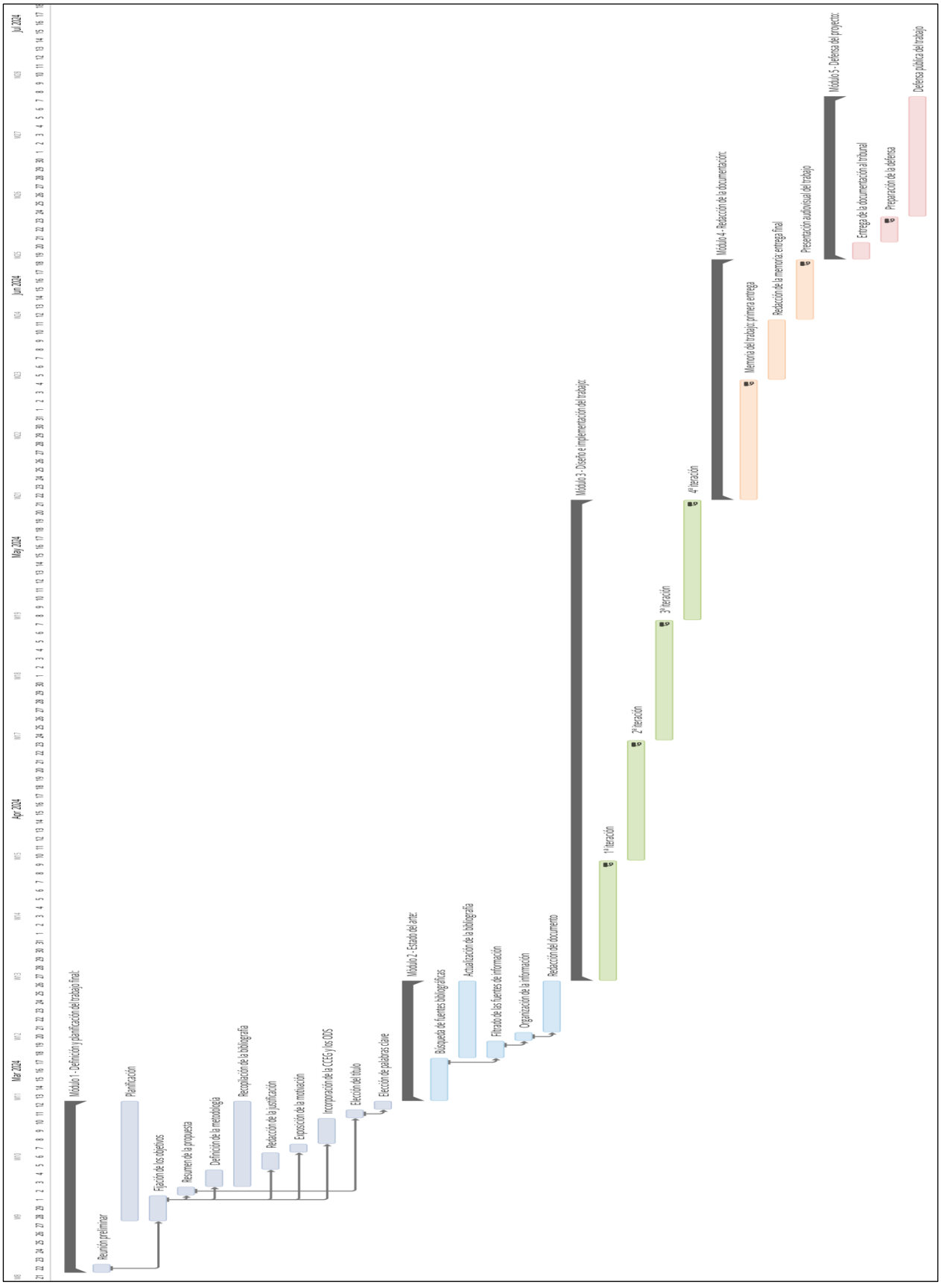


Figura 4. Cronograma del TFM

1.7. Breve sumario de productos obtenidos

A la finalización de este TFM se han obtenido los siguientes productos:

1. Segmentación en 2 fases de la cartera de clientes.
2. Fichero de clientes etiquetado de acuerdo con la segmentación.
3. Modelo de predicción de transferencia de clientes entre segmentos.
4. Fichero de clientes etiquetado de acuerdo con la predicción de transferencia.
5. Modelo de predicción de recurrencia en la compra.
6. Segmentación de la clientela de acuerdo con su recurrencia.
7. Modelo de predicción de transferencia entre segmentos de recurrencia.
8. Fichero de clientes etiquetado con la categoría de recurrencia, la segmentación de recurrencia y la predicción de transferencia de segmento de recurrencia.
9. Cuaderno Jupyter donde se recoge el proceso de creación de todos los anteriores productos, y que permite su adaptación a otros contextos de segmentación y optimización de la fidelización de la clientela en el ámbito de la inteligencia de negocios.

2. Estado del arte

2.1. Introducción

A finales de la década de 1950 se plantea la necesidad de abandonar el enfoque homogéneo de la oferta y demanda del mercado que habían mantenido tradicionalmente las técnicas de la mercadotecnia hasta el momento. Se hace así patente la necesidad de **elegir estrategias de marketing adecuadas basadas en las condiciones y dinámicas del mercado**, tanto a través de la diferenciación de productos que consigan distinguirnos de otros competidores, como precisando cuáles de dichos productos son capaces de satisfacer los requisitos de grupos definidos y distinguibles del público (Smith, 1956). Pero no será hasta finales de los años 60 cuando la **segmentación** se presente **como un concepto independiente** de la diferenciación de producto, definiéndose como el proceso de abordar un mercado heterogéneo por naturaleza como una serie de mercados más pequeños y homogéneos en respuesta a sus diferentes preferencias y necesidades (Claycamp, 2024).

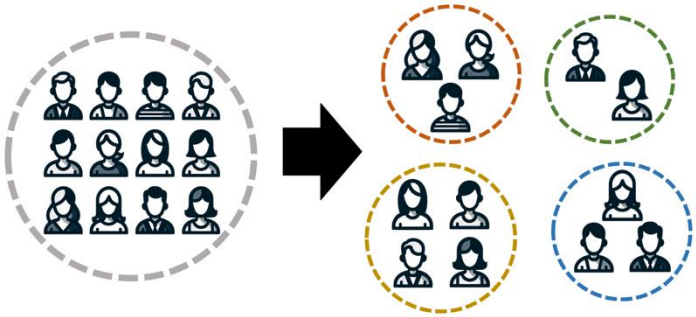


Figura 5. Proceso de segmentación del mercado

Sin embargo, este proceso se mostró **ineficiente y demasiado costoso** hasta que la tecnología permitió trabajar con grandes volúmenes de información. De esta forma, la irrupción de las **técnicas de KDD y de la minería de datos** permitieron la llegada del marketing basado en el conocimiento (Shaw et al., 2001). La literatura científica atestigua la sucesión de múltiples ejemplos de aplicación de análisis de agrupación a la mercadotecnia, donde fundamentalmente son utilizados para la segmentación del mercado, convirtiéndose así ésta en una importante herramienta tanto para la investigación académica como para el marketing aplicado (Punj and Stewart, 2024).

Destacan a partir de finales de la década de 1960 y a largo de toda la de los 70 ejemplos de aplicación de diferentes **metodologías de agrupación** a la segmentación de clientes, incluyendo métodos jerárquicos basados en distancias como el de la varianza mínima de Ward (Kernan, 1968), el del enlace promedio (Bass et al., 1969) y completo (Montgomery and Silk, 1971), o de partición iterativa como *k-means* (Schaninger et al., 1980). Posteriormente aparecerán estudios comparativos con otros algoritmos equivalentes como *k-medoids* (Aryuni et al., 2018) o basados en densidad como DBSCAN (Sembiring Brahmana et al., 2020) que analizaremos con más profundidad en los siguientes apartados.

Es en la primera mitad de la década de los 90 cuando las **redes neuronales** comienzan a aplicarse en la segmentación (Vellido, 1999), en un primer momento aplicando técnicas de aprendizaje supervisado como BPGD (Mazanec, 1992), pero

progresivamente también a través de modelos no supervisados como FSCL (Balakrishnan et al., 1996). Sin embargo, y a pesar de que las técnicas disponibles para el análisis no dejan de crecer, los estudios comparativos inciden en que la elección de **la técnica de agrupamiento adecuada resulta crucial** para conseguir una segmentación significativa (Kansal et al., 2018), y los resultados de *k-means* superan en múltiples comparativas los obtenidos por redes neuronales debido a la dificultad de parametrización de éstas en un entorno de aprendizaje no supervisado (Balakrishnan et al., 1996).

En la actualidad, la segmentación a través de técnicas de agrupamiento se enfrenta a **nuevos desafíos, fruto de los actuales entornos ricos en datos**, incluyendo la actualización en tiempo real para flujos de datos o el agrupamiento de conjuntos de datos de muy alta dimensionalidad. Así, entre las nuevas metodologías de aprendizaje automático encontramos el uso de técnicas de cuantización vectorial para la simplificación del conjunto de datos y la identificación de perfiles de cliente, el uso de modelos de temas para agrupamiento suave de textos no estructurados disponibles en las publicaciones de redes sociales, o métodos de partición de gráficos para el agrupamiento de segmentos con muchas conexiones (Reutterer and Dan, 2020).

2.2. Publicaciones de referencia

Sin embargo, y debido a lo extenso de la producción científica en el campo del aprendizaje computacional aplicado a la segmentación, centraremos nuestra atención en aquellos trabajos que abandonan el enfoque competitivo entre técnicas en favor de una estrategia colaborativa, así como en aquellos que se centran en aspectos clave de la gestión y el análisis de las interacciones y relaciones con los clientes o CRM (*Customer Relationship Management*):

- Así, en el artículo «**Comparative performance of the FSCL neural net and K-means algorithm for market segmentation**» (Balakrishnan et al., 1996), se propone el uso de una red neuronal *Frequency-Sensitive Competitive Learning* (FSCL) como paso inicial para la obtención de una primera clasificación aproximada que es posteriormente utilizada como semillas para la aplicación del algoritmo *k-means*, el cual toma estos grupos iniciales y los refina iterativamente revisando cada muestra para decidir cuál es su grupo de pertenencia y ajustando cada grupo hasta hacerlo más preciso y coherente. Este enfoque híbrido presenta ventajas significativas sobre el uso independiente de cada técnica, al conseguirse segmentos con características claras y distintivas que pueden facilitar la elección de una estrategia de marketing y venta para cada uno de ellos. En una línea similar, el estudio «A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA» (Alkhayrat et al., 2020) utiliza *autoencoders* en conjunción con PCA para reducir la dimensionalidad de los datos, para después aplicar *k-means* en la segmentación.

- Por su parte, en «**A Two Phase Clustering Method for Intelligent Customer Segmentation**» (Namvar et al., 2010), los autores utilizan un método de segmentación mediante *k-means* estructurado en dos fases. Durante la primera, la clientela es clasificada en segmentos de acuerdo con sus valores RFM (*Recency*, *Frequency*, *Monetary*), es decir, cuan reciente fue su última compra, su frecuencia de compras y el valor monetario total de éstas. En la segunda fase, estas agrupaciones son subdivididas en nuevos clústeres de acuerdo con la información demográfica de la clientela para, finalmente asignar un perfil a cada cliente de acuerdo con su LTV (*LifeTime Value*) o valor total que la empresa esperar obtener de él a lo largo de toda su relación comercial. De esta manera, se consigue que la segmentación no se limite a un único punto de vista, enfoque que ha sido replicado en otros contextos en estudios sucesivos combinando redes neuronales, algoritmos basados en distancias, así como otros enfocados a la minería de reglas de asociación (Sarvari et al., 2016).

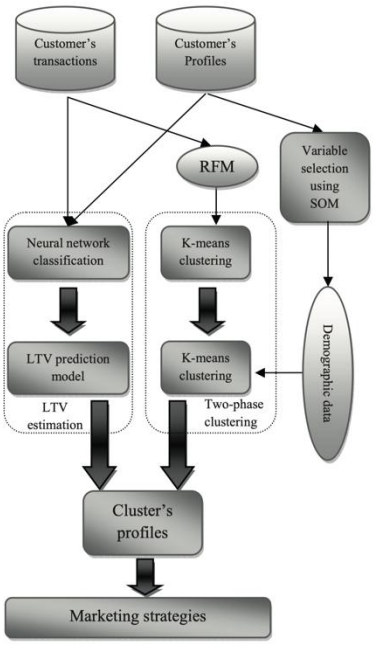


Figura 6. Marco investigador en 2 fases (Namvar et al., 2010)

- En referencia al uso de los valores RFM durante la segmentación, cabe destacar el estudio «**RFM ranking – An effective approach to customer segmentation**» (Christy et al., 2021), en el cual éstos se ordenan en secuencia ascendente en tres vectores separados (uno para cada uno de los valores: recencia, frecuencia y monto de compra), se calcula la mediana de cada vector y el resultado es utilizado como centroide inicial para la aplicación del algoritmo *k-means*. El mismo proceso se lleva a cabo de forma iterativa según el valor de $\boldsymbol{k}$, alcanzándose así la convergencia en un menor número de iteraciones y obteniendo clústeres más compactos, mejorando en conjunto la eficiencia del proceso.

- En cuanto al cálculo del LTV, es relevante hacer mención al enfoque empleado en el estudio «**Customer segmentation and strategy development based on customer lifetime value: A case study**» (Kim et al., 2006) donde, tomando como punto de partida un período de tiempo específico, el LTV es abordado desde tres puntos de vista diferentes medidos por separado:

  1. El valor actual o cantidad promedio de pagos esperados de un cliente (sin tener en cuenta los cargos pendientes).
  2. El valor potencial o beneficio esperado si el cliente contratase servicios adicionales.
  3. La lealtad del cliente o *churn rate* (tasa de abandono).

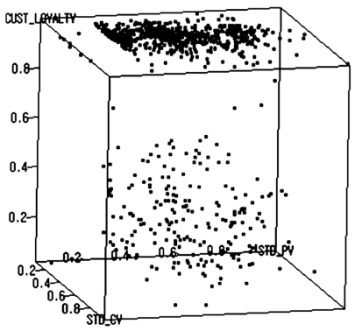


Figura 7. Segmentación tridimensional del valor (Kim et al., 200)

De esta forma cada uno de estos componentes es utilizado para la segmentación de los clientes para identificar aquellos con un alto valor actual, los que poseen un gran valor potencial para ventas futuras y aquellos que cuentan con un riesgo de deserción, y que por ende pueden requerir estrategias de retención específicas. A diferencia del estudio anterior, este enfoque no conlleva sin embargo realizar segmentaciones separadas para cada una de las dimensiones del valor, sino que las tres se integran para representar la segmentación en un espacio tridimensional donde cada eje denota el valor actual, el potencial y la lealtad del cliente respectivamente, para obtener una visión más equilibrada de su valor. Por último, se utiliza un árbol de decisión para minar las características de los clientes dentro de cada segmento y clasificarlos.

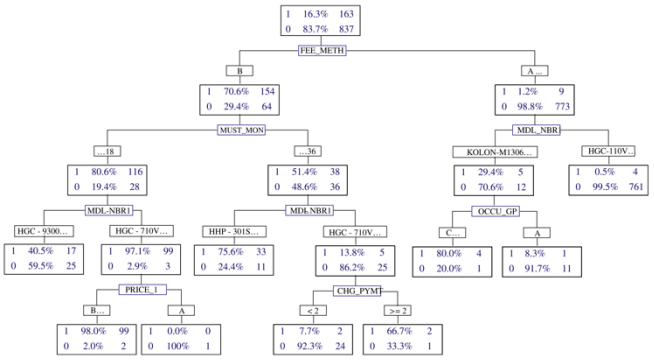


Figura 8. Árbol de decisión (Kim et al., 2006)

- Finalmente, el estudio «**Customer segmentation of multiple category data in e-commerce using a soft-clustering approach**» (Wu and Chou, 2011) analiza junto a las características demográficas del cliente su comportamiento de compra en el comercio electrónico para después aplicar un método de segmentación suave a través de una técnica derivada del modelo de asignación de Dirichlet latente para crear los segmentos, de forma que un mismo cliente puede pertenecer a más un clúster si tiene intereses que se solapan con diferentes segmentos del mercado. Sin embargo, esta visión enriquecida de la segmentación puede no ser útil en todos los contextos, funcionando especialmente bien en aquellos casos donde la mencionada superposición entre categorías sea una opción lógica y natural.

3. Metodología empleada

3.1. Algoritmia

A lo largo de los estudios analizados sobresalen algunos algoritmos utilizados durante el proceso de segmentación, tanto debido a su popularidad como a los buenos resultados obtenidos. A continuación, analizaremos a continuación los principales.

3.1.1. Algoritmos basados en distancia

Se trata de algoritmos que buscan minimizar la distancia dentro de los clústeres, así como maximizar la distancia entre éstos. Pese a que *k-means* es el más popular, mencionaremos otros algoritmos relacionados empleados en los estudios previamente citados.

K-means

K-means (Lloyd, 1982) (MacQueen, 1967) es un algoritmo de clasificación no supervisada basado en la **partición de un conjunto de n observaciones en k grupos**, de forma que cada observación pertenece a aquel grupo respecto al cual guarda una menor distancia. El total k de clústeres debe ser definido previamente a la ejecución del algoritmo (Gironés Roig et al., 2017).

Dada una serie de observaciones $(x_1, x_2, \dots, x_n)$, donde cada observación es un vector real de d dimensiones, *k-means* particionará las n observaciones en k ($k \leq n$) conjuntos $S = \{S_1, S_2, \dots, S_k\}$, minimizando la varianza dentro del clúster. Su función objetivo es:

$$J = \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2$$

donde:

- J es la función de costo total,
- S_i es el conjunto de puntos del clúster i,
- x es un punto de datos del clúster i,
- μ_i es el centroide del clúster i, calculado obteniendo el promedio de todos los puntos en S_i,
- $||x - \mu_i||^2$ es la distancia euclidiana al cuadrado entre la observación x y la media del clúster μ_i.

Inicialmente se establece un total de $\boldsymbol{k}$ centroides de forma aleatoria y, durante su ejecución, el algoritmo repite la siguiente secuencia de pasos de forma iterativa:

1. Asignación: Adscribiendo cada observación al clúster con el centroide más cercano.
2. Actualización: Calculando los nuevos centroides, es decir, la media de todas las observaciones que han sido asignada a cada clúster durante el paso anterior.

Finalmente, el algoritmo se detendrá cuando dejen de producirse cambios significativos en la asignación de centroides o una vez que las asignaciones de los puntos a los clústeres permanezcan constantes.

A pesar de la eficiencia computacional del algoritmo cuando se trabaja con un número reducido de clústeres, así como de su popularidad, *k-means* presenta ciertas limitaciones, tales como la sensibilidad a los valores extremos y sus limitaciones ante datos complejos que producen clústeres con formas irregulares y densidades variables (VanderPlas, 2016) (Ahmed et al., 2020).

K-medoids

Se trata de un algoritmo de funcionamiento parejo a *k-means*, el cual selecciona un conjunto de puntos reales del conjunto de datos para representar a los clústeres llamados «medoides» (Kaufman and Rousseeuw, 1987). Así, el propósito del algoritmo es minimizar la suma de las disimilitudes entre los puntos asignados a un *medoide* y el propio *medoide*. Su función objetivo es por lo tanto similar a la de *k-means*:

$$J = \sum_{i=1}^{k} \sum_{\forall x \in S_i} d(x, m_i)$$

con la diferencia de que $(d(x, m_i))$ es la distancia entre el punto x y el medoide i.

De esta forma, *k-medoids* resulta más **robusto frente a los valores atípicos**, dado que el *medoide* es un punto real del conjunto de datos, particularidad que lo hace menos sensible a las variaciones extremas que pueden afectar al cálculo del centroide promedio en el caso de *k-means*. A cambio, exige un mayor coste computacional, si bien recientes estudios han conseguido mejorar su eficiencia (Tiwari and Zhang, 2020).

Determinación del valor de *k*

Sin embargo, y como ha sido especificado anteriormente, en todos estos casos es preciso determinar previamente el número $\boldsymbol{k}$ de clústeres. Con este propósito se utilizan diferentes métodos, siendo algunos de los más frecuentes los siguientes:

a) La **regla del codo** (Syakur et al., 2018), basado en la de la varianza explicada a medida que aumenta el número de clústeres, buscando aquel punto donde el incremento marginal de ésta disminuye significativamente.

La técnica debe así su nombre al «codo» que forma la curva del gráfico en el punto que se produce dicha disminución.

b) El cálculo del **coeficiente de la silueta** (Rousseeuw, 1987) para diferentes valores de $\boldsymbol{k}$, midiendo de esta manera el grado de similitud un objeto a su propio clúster comparado con otros clústeres, siendo el valor óptimo de $\boldsymbol{k}$ aquel que maximice el coeficiente promedio de la silueta.

c) El **índice Davies-Bouldin** (Davies and Bouldin, 1979), el cual busca clústeres compactos alejados entre sí y con baja dispersión interna. Cuanto más bajo sea el índice, más definidas serán las agrupaciones.

d) El **índice Calinski-Harabasz** (Calinski and Harabasz, 1974) compara la dispersión entre los clústeres con la interna dentro de los propios clústeres. Cuanto más alto sea el índice, mejor definidos y separados serán los clústeres resultantes.

Selección de métricas

Respecto a la medida de las distancias, entre las métricas más habituales se encuentran las siguientes:

a) **Distancia euclidiana** o L2: La cual mide la distancia en línea recta entre dos puntos en el espacio euclídeo.

$$\mathrm{d}(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \mathrm{y_i})^2}$$

b) **Distancia Manhattan** o L1: Mide la distancia entre dos puntos a lo largo de los ejes en ángulo recto.

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

c) **Similitud del coseno**: Que mide el coseno del ángulo entre dos vectores distintos de cero, lo cual indica orientación en lugar de magnitud.

$$\text{cosine similarity} = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

d) **Distancia de Minkowski**: La cual generaliza las distancias euclídea y de Manhattan con un parámetro $\boldsymbol{p}$.

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

e) **Índice de Jaccard**: Que mide la similitud entre conjuntos de muestras finitas y se utiliza comúnmente para datos binarios o categóricos.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

La idoneidad de una u otra métrica depende en gran parte de la dimensionalidad y distribución de los datos, y su selección ha demostrado tener un importante impacto en la calidad de las agrupaciones resultantes (Shirkhorshidi et al., 2015).

3.1.2. Algoritmos basados en densidad

Se trata de algoritmos que buscan **identificar regiones de alta densidad separadas entre sí por regiones de baja densidad**.

El caso más popular es DBSCAN, siglas correspondientes a la expresión inglesa *Density-Based Spatial Clustering of Applications with Noise* (o agrupación espacial basada en la densidad de aplicaciones con ruido). Como apuntábamos, su funcionamiento se basa en identificar clústeres basados en la densidad de los puntos en un espacio, de forma que un punto se considerará que forma parte de un clúster si hay un número mínimo de puntos (MinPts) dentro de un radio específico (ε) de él, hecho que se toma como indicativo de una alta densidad; mientras, aquellos puntos que no cumplen dichos criterios serán considerados ruido. De esta manera, el algoritmo agrupa puntos densamente conectados capaces de formar clústeres con formas arbitrarias en presencia de ruido. Son por ello más eficientes en presencia de valores extremos y ruido, pero su parametrización no siempre es sencilla y no siempre responde bien a las variaciones de densidad dentro de los datos (Wang et al., 2019).

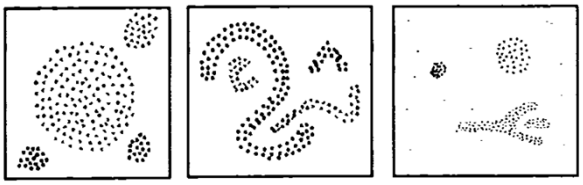


Figura 9. Clústeres identificados por DBSCAN (Ester et al., 1996)

3.1.3. Algoritmos jerárquicos

Se trata de algoritmos que construyen jerarquías de clústeres, dando como resultado un dendrograma que muestra la relación entre los clústeres identificados a diferentes niveles de agrupación. Sin embargo, estos algoritmos no son recomendables con grandes conjuntos de datos debido a su complejidad computacional en el cálculo y la necesidad de actualizar las distancias entre todos los pares de clústeres a cada iteración, si bien existen alternativas recientes que palian parcialmente esta limitación (Kobren et al., 2017).

La construcción de la jerarquía se lleva a cabo de dos maneras posibles: *top-down* o algoritmo divisivo, en el cual todos los puntos comienzan formando un único clúster que es particionado recursivamente en clústeres más pequeños; y la versión más popular, llamada *bottom-up* o algoritmo aglomerativo, que pasamos a analizar más en detalle a continuación.

Bottom-up

Al contrario de lo que sucede con el procedimiento *top-down*, *bottom-up* sigue un procedimiento **aglomerativo** partiendo de una organización inicial en que cada punto forma su propio clúster, para seguidamente ir fusionando los clústeres próximos de forma iterativa. Cabe destacar que este enfoque es computacionalmente más exigente que la estrategia divisiva del enfoque *top-down* (Gironés Roig et al., 2017).

Existen diferentes estrategias a la hora de aplicar el procedimiento aglomerativo, pero destacaremos por su popularidad las siguientes:

1. **Enlace simple**: Según la cual la fusión se basa en la distancia mínima entre cualquier par de puntos de los clústeres a unir. Gracias a esta metodología es posible detectar clústeres de forma no esférica, y de hecho tiende a crear cadenas de diferentes grupos, aunque es sensible al ruido.
2. **Enlace completo**: Donde la medida utilizada es la máxima distancia entre cualquier par de puntos de los clústeres. Así se consiguen clústeres compactos y bien separados entre sí, pero el procedimiento es sensible a la presencia de valores extremos.
3. **Método de Ward**: El cual busca minimizar la varianza total de los clústeres con objeto de fusionar los que produzcan el menor incremento de varianza total dentro de la agrupación. De esta forma los clústeres resultantes son generalmente esféricos y de medidas similares, pero en la práctica estas estructuras no siempre se corresponden con la estructura inherente de los datos.

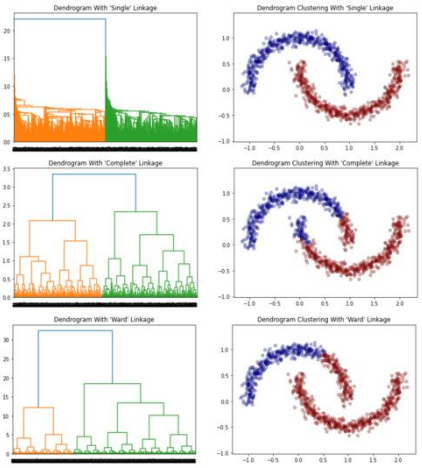


Figura 10. Enlace simple, completo y de Ward aplicado al *dataset* «Moons»

3.1.4. Redes neuronales

Las redes neuronales son un conjunto de algoritmos, dentro de la categoría de los jerárquicos, que se inspiran en el funcionamiento de las neuronas biológicas. Sus modelos se componen de unidades básicas de procesamiento (o neuronas) organizadas en **capas, interconectadas y con capacidad de aprender de los datos a través de un proceso de ajuste** de sus conexiones (o pesos). Este sistema les permite trabajar con datos imprecisos y complejos, y adaptarse a cambios y patrones emergentes, manejando con eficiencia relaciones complejas y no lineales entre características de los datos (Bosch Rué et al., 2019). Dicho esto, su interpretación resulta compleja, tienen un importante riesgo de sobreajuste a los datos de entrenamiento y, computacionalmente hablando, resultan extremadamente costosas en recursos y tiempo (Sze et al., 2017).

Entre las redes neuronales más frecuentemente aplicadas para la agrupación, destacan las siguientes:

1. **Redes de aprendizaje profundo**: Caracterizadas por la multiplicidad de capas ocultas situadas entre la capa de entrada y la de salida que permiten al modelo aprender jerarquías de características de los datos (Bosch Rué et al., 2019). En esta categoría destacan los *autoencoders*, capaces de aprender una representación comprimida de los datos de entrada, lo cual no sólo resulta efectivo para la reducción de dimensionalidad de los datos (Alkhayrat et al., 2020), sino también para gestionar la clase de datos complejos y desordenados habituales de los estudios de carácter empírico (Mangiameli et al., 1996). También se deben mencionar las redes neuronales convolucionales o CNN (*Convolutional Neural Network*), utilizadas principalmente para agrupar imágenes de las que extraen sus características, pero que también pueden ser usadas con otros datos estructurados de forma similar (Lieder et al., 2019).
2. **Redes competitivas**: En las cuales las neuronas de una capa compiten entre sí por activarse y donde la neurona con la respuesta más fuerte inhibirá al resto. En esta categoría destacan los llamados mapas autoorganizativos o SOM (por las siglas en inglés de la expresión *Self-Organizing Maps*), en los cuales se mantienen relaciones topológicas entre los patrones de entrada de forma que patrones similares activarán neuronas próximas sobre el mapa, lo que se traduce en un rendimiento robusto ante diversas imperfecciones de los datos tales como variables irrelevantes, *outliers* o clústeres con densidades no uniformes (Mangiameli et al., 1996).

3.1.5. Análisis comparativo

De toda la información expuesta relativa a los algoritmos usados en los procesos de agrupación, se extrae la siguiente tabla que sintetiza las conclusiones que deberemos tener presentes durante el desarrollo de nuestro producto:

	ALGORITMOS			
	POR DISTANCIA	**POR DENSIDAD**	**JERÁRQUICOS**	**REDES NEURONALES**
VENTAJAS	• Rápidos y eficientes. • Fácil implementación.	• Aptos para clústeres de formas arbitrarias, *outliers* y presencia de ruido.	• Ofrecen un dendrograma de fácil interpretación. • Permiten examinar diferentes niveles de granularidad.	• Adaptables a datos complejos. • Las SOM facilitan la visualización y comprensión de datos de alta dimensión.
INCONVENIENTES	• Requieren determinar previamente el total de clústeres. • Sensibles a *outliers*. • No aptos para clústeres no esféricos.	• Mala respuesta ante variaciones de densidad. • Parametrización compleja.	• Computacionalmente costoso con grandes conjuntos de datos. • Inefectivos con datos que se desvían del ideal de clústeres compactos y aislados.	• Difícil interpretación. • Computacionalmente costoso. • Riesgo de sobreajuste.

Tabla 1. Análisis comparativo de la tipología de algoritmos de agrupamiento

3.2. Técnicas suplementarias

3.2.1. Técnicas de equilibrado

En términos generales, encontramos un problema de desbalanceo en los datos cuando las clases en un conjunto de datos no están representadas de manera equitativa. Si bien es más habitual en el contexto de un problema de clasificación, también es posible encontrarlo en el caso del **aprendizaje no supervisado** cuando existe un número desigual de puntos entre los diferentes clústeres inherentes a los datos. Esta falta de equilibrio puede manifestarse:

a) Cuando encontramos **gran diferencia entre el número de puntos que forman los clústeres**, hecho que puede afectar la capacidad del algoritmo para identificar los grupos más reducidos, pudiendo éstos acabar subsumidos dentro de clústeres más grandes.

b) Cuando encontramos **gran diferencia en la densidad de los clústeres**, particularmente en el caso de los algoritmos que se basan en ésta, como DBSCAN.

De ser así, el problema puede ser detectado a través de técnicas de reducción de la dimensionalidad que nos ayuden a visualizar la distribución de los clústeres; mediante **métricas de evaluación interna**, como el coeficiente de silueta, que nos retornarán información sobre la cohesión y separación de los grupos; o bien, una vez ejecutado el algoritmo de agrupación, **examinando la distribución en clústeres** resultantes y sus tamaños. Por otra parte, el recurso a las métricas habituales de evaluación del modelo junto con la referida observación de los clústeres, pueden apuntar al desbalanceo: así, en el caso de las métricas de calidad por partición, un **diámetro** de clúster desproporcionadamente grande en comparación con el resto puede sugerir que éste es demasiado disperso; por otro lado, las métricas de calidad general, como es el caso de un **índice Dunn** bajo, puede apuntar a la existencia de agrupaciones mal definidas o solapadas, lo que podría ser un síntoma del desbalanceo del *dataset* (Gironés Roig et al., 2017).

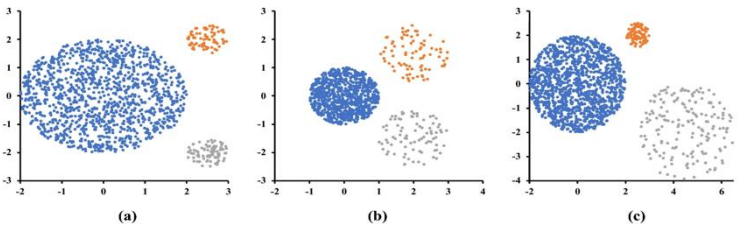


Figura 11. Agrupaciones con distribuciones sesgadas (Liu Yun et al., 2021)

Algunas de las **soluciones** que pueden paliar este desbalanceo son las siguientes:

1. Si bien los algoritmos basados en densidad son menos sensibles a este problema al no asumir una distribución uniforme de los clústeres, además del mencionado DBSCAN cabe destacar la existencia del algoritmo **HDBSCAN** (Campello et al., 2013). Éste incorpora una perspectiva jerárquica al construir una jerarquía de clústeres conectados, para después seleccionar los más

estables transformando el espacio de características en un espacio de densidad en el cual la distancia entre los puntos refleja su densidad mutua. A continuación, el algoritmo explora este nuevo espacio en busca de regiones de alta densidad persistentes a través de un rango de escalas de distancia, identificando éstas como clústeres. Dicho esto, nos encontramos ante un algoritmo de alto coste computacional y la interpretación de los resultados no es siempre evidente.

2. Existe una categoría de algoritmos de agrupación basados en un modelo probabilístico en la cual destaca EM o ***Expectation-Maximization*** (Dempster et al., 1977), el cual asume que los datos provienen de una mezcla de diversas distribuciones gaussianas cada una de las cuales representa un clúster. Para ello, en un primer paso se estima la probabilidad de que cada punto de datos pertenezca a cada uno de los clústeres en el llamado paso de expectativa; para después, basándose en dichas probabilidades, maximizar la probabilidad de los datos a partir de estos parámetros ajustando medias, varianzas y coeficientes de mezcla de las distribuciones con el objetivo de que reflejen de forma más fidedigna la estructura de los datos según las probabilidades calculadas previamente. El proceso se repite iterativamente hasta su convergencia en una solución estable o hasta que alcanza un determinado umbral. Este enfoque, pese a poder modelar clústeres de formas complejas y ofrecer la probabilidad de pertenencia de cada punto a cada clúster, es computacionalmente costoso y muy sensible a su parametrización, la cual incluye la determinación del número de agrupaciones.
3. Sobremuestreo u ***oversampling***: Incrementando artificialmente los grupos más reducidos replicando sus elementos (Patel, 2019).
4. Submuestreo o ***undersampling***: Reduciendo el tamaño de los grupos más grandes suprimiendo algunos de sus elementos, aunque esta estrategia puede provocar la pérdida de información relevante (Patel, 2019).
5. **Reducción de dimensionalidad**: Debido a que estas técnicas resaltan las estructuras naturales de los datos en espacios de menor dimensionalidad, pueden mitigar un eventual desbalanceo existente en los datos, y serán analizadas en más detalle en subsiguientes apartados.
6. **Agrupación en dos etapas**: Como hemos visto en algunos de los estudios mencionados (Namvar et al., 2010), la agrupación puede ser estructurada en dos fases para, en un primer paso, identificar agrupaciones generales y, a continuación, aplicar el algoritmo de agrupación dentro de cada uno de los grupos con objeto de identificar grupos más pequeños, los cuales pueden haber sido ignorados o incorrectamente asignados durante la primera fase.

Cabe finalmente mencionar algunas propuestas recientes para la aplicación de *k-means* sobre conjuntos de datos con distribuciones sesgadas con objeto de hacer frente a la tendencia del algoritmo original de igualar los volúmenes de los clústeres sin considerar sus densidades o tamaños reales (Liu et al., 2021). Así, el algoritmo ajusta la función objetivo incorporando a las actualizaciones una medida de densidad junto con la distancia del centro del clúster.

Por otro lado, y pese a que nuestro objetivo principal de segmentar los clientes es un problema de agrupación, no debemos perder de vista que está ligado al objetivo de aumentar el tique de compra, bien a través de técnicas de *cross-selling* como a través del *up-selling*. Nuestro *dataset* posee información sobre estas **variables numéricas continuas** (cantidad de productos comprados, monto total de las compras), las cuales, si se tratan como variables objetivo de un **problema de regresión**, podrían resultar desbalanceadas en los siguientes casos para los cuales se plantean las siguientes estrategias de mitigación:

a) Cuando exista una distribución sesgada de estas variables con valores concentrados en un determinado rango, problema que puede mitigarse normalizando las variables objetivo a través de su transformación logarítmica o de la raíz cuadrada.

b) Por la presencia de valores extremos que distorsionen los datos, los cuales deberán ser detectados y tratados, o bien emplear algoritmos menos sensibles a su presencia.

c) Por la presencia de heterocedasticidad o variación de la varianza a lo largo del rango de valores, ante lo cual se puede someter los datos a transformación, utilizar algoritmos que no asuman la homogeneidad de la varianza o utilizar un método de regresión ponderada.

De igual manera, la predicción de recurrencia, así como la transferencia entre clústeres pueden ser tratados como un problema de **clasificación**, en el cual encontraremos el caso arquetípico de desbalanceo a través de una representación desigual de las clases.

3.2.2. Técnicas de valoración de variables

Uno de los aspectos fundamentales durante la segmentación consiste en valorar la **importancia de las características que son tomadas en cuenta durante el proceso de formación de las agrupaciones**. Sin embargo, tal como ha sido expuesto durante los apartados precedentes, éste no es un proceso trivial, especialmente en el caso de los modelos de difícil interpretación que operan como «cajas negras». Por todo ello, es necesario tener presentes las siguientes técnicas.

Análisis de sensibilidad

Consiste en la **variación de una o más variables** con objeto de **identificar cómo afectan** al proceso de formación de los clústeres. Estas modificaciones pueden incluir:

a) Variaciones en la parametrización del algoritmo, tales como modificar el valor de $\boldsymbol{k}$ en el caso de *k-means*, o los *eps* y *MinPts* en el caso de DBSCAN. En el caso de los algoritmos basados en distancias, esto puede incluir la modificación de la métrica de distancia, reemplazando, por ejemplo, la distancia euclidiana por la de Manhattan.

b) Incluir o excluir determinadas variables de la ejecución del algoritmo.

c) Modificar el valor de las variables para observar su impacto en los clústeres resultantes.

d) Utilizar diferentes algoritmos de agrupación de entre los expuestos con anterioridad.

Técnicas de reducción de dimensionalidad

Destacamos aquí el análisis de componentes principales (en adelante, «**PCA**», siglas de la expresión inglesa *Principal Component Analysis*), ya que si bien se trata de una técnica de reducción de la dimensionalidad, ofrece información sobre la varianza explicada por cada componente principal, revelando qué características contribuyen en mayor medida a la variación en los datos (VanderPlas, 2016).

Por otra parte, la descomposición en valores singulares o **SVD** (siglas de *Singular Value Decomposition*) permite la descomposición de los datos en componentes que, al capturar la mayor parte de la variabilidad, ofrecen igualmente pistas sobre la relevancia de las características (Patel, 2019).

Si bien otras técnicas de reducción de la dimensionalidad, tales como **t-SNE** (Van der Maaten and Hinton, 2008) o **UMAP** (McInnes et al., 2020), no tienen por objetivo la valoración directa de variables, sí pueden ayudar durante la exploración de la estructura de los datos para identificar si ciertas combinaciones de variables son informativas.

Técnicas de interpretación de modelos

Tal como ha sido expuesto, dado que nuestro objetivo está ligado al aumento del tique de compra, y nuestro *dataset* cuenta con información como el número de compras realizadas o su importe, es posible también la aplicación de técnicas de interpretación más frecuentes en el aprendizaje supervisado. Además, estas técnicas podrían ser aplicadas en un segundo paso **una vez etiquetados los datos** como resultado de una primera fase de segmentación. Así, destacamos las siguientes:

a) *Clustering Feature Importance* (en adelante, «CFI») elimina iterativamente una característica cada vez, evaluando su impacto en la calidad de la agrupación para identificar cuáles son las características más importantes (Witten and Tibshirani, 2010) (Alelyani et al., 2018).

b) *Feature Importance* en *Random Forest* permite identificar la contribución de cada característica en la predicción del modelo (Breiman, 2001).

c) LIME (siglas de *Local Interpretable Model-agnostic Explanations*) es una técnica capaz de explicar las predicciones de los modelos de aprendizaje computacional a través de pequeñas modificaciones en los datos de entrada para generar un nuevo conjunto de datos, que será entrenado y ponderando los resultados (Ribeiro et al., 2016).

d) SHAP (*SHapley Additive exPlanations*) explica el resultado de los modelos basándose en la teoría de los juegos cooperativos para medir la importancia de

cada una de las características implicadas en su predicción (Lundberg and Lee, 2017).

3.3. Trabajo con datos

El proceso de recolección, organización, análisis y gestión de los datos puede ser llevado a cabo desde diferentes metodologías con el propósito de maximizar la utilidad que se puede extraer de ellos durante los procesos de toma de decisiones y el desarrollo de proyectos. Tal y como apuntamos en apartados precedentes de la memoria, la metodología escogida ha sido CRISP-DM, pero es relevante hacer mención aquí a las más destacadas entre las disponibles.

3.3.1. CRISP-DM

Su nombre responde a las siglas de la expresión inglesa *CRoss-Industry Standard Process for Data Mining* (proceso estándar interprofesional para la extracción de datos) (CORDIS, 1998). Concebida en 1996 como un projecto financiado por la Unión Europea y presentada en 1999, está considerado un **estándar de facto** por la industria en el campo de la minería de datos por su clara estructuración de todas las fases del proyecto, las cuales ya han sido enumeradas en apartados anteriores.

3.3.2. KDD

Siglas de Knowledge Discovery in Database (o descubrimiento del conocimiento en bases de datos). Surge en 1989 como resultado del trabajo de diversos investigadores en el campo de las bases de datos y la inteligencia artificial (Frawley et al., 1992), y pone el énfasis en un proceso de transformación iterativa enfocada al **descubrimiento y transformación de los datos en conocimiento**. Sus fases incluyen: (1) selección de datos, (2) prepocesamiento eliminando ruido e inconsistencias, (3) transformación normalizando, a través de *feature engineering* o reduciendo la dimensionalidad; (4) minería de datos a través del algoritmo oportuno, (5) interpretación y evaluación del resultado, y (6) consolidación del conocimiento obtenido incorporándolo al sistema para futuras referencias y para la toma de decisiones. Sin embargo, su resultado depende en gran medida de la calidad y mantenimiento de la información, convirtiéndolo en la práctica en un proceso complejo.

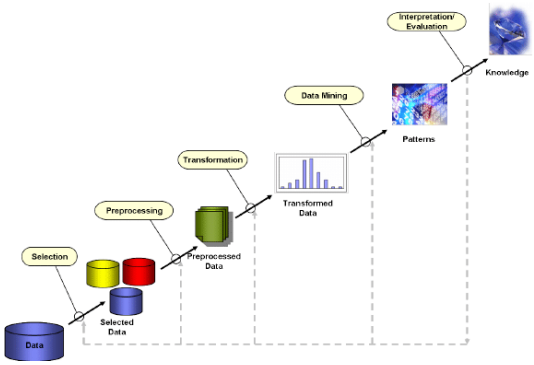


Figura 12. Diagrama de las fases de KDD (Guerra-Hernández et al., 2008)

3.3.3. SEMMA

El nombre corresponde a la expresión inglesa *Sample, Explore, Modify, Model, and Asses* (muestrea, explora, modifica, modela y evalúa), y fue desarrollada por SAS Institute Inc. (SAS, 2017) y, si bien inicialmente estaba orientado a la modelización predictiva, su enfoque estructurado es aplicable a una amplia gama de tareas de carácter analítico, incluyendo los procesos de agrupación. Sus fases consisten en: (1) seleccionar una muestra representativa de los datos que capture la variabilidad y características esenciales de la totalidad del dataset, (2) analizar los datos en busca de patrones y tendencias, (3) transformar los datos para su análisis empleando técnicas de normalización, imputación o feature engineering; (4) construir el modelo y (5) evaluar la calidad del modelo. A pesar de ello su enfoque está demasiado **ligado al uso de las herramientas que desarrolla la propia SAS**.

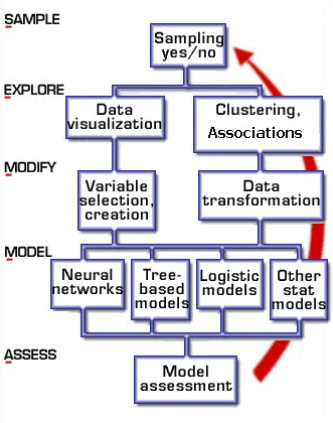


Figura 13. Diagrama de las fases de SEMMA (SAS, 2017)

3.4. Impacto ético social, de sostenibilidad y de diversidad

Los notables avances de técnicas de aprendizaje computacional vividos durante los últimos años han llevado a centrar los esfuerzos en conseguir resultados que mejoren el estado del arte, ignorando el coste ambiental y social aparejado. Así, durante la década de 2010 el coste computacional de los nuevos modelos de aprendizaje computacional se ha duplicado cada pocos meses, estimándose un aumento de hasta un total de 300.000 veces (Schwartz et al., 2020). Dicho coste ha llevado a incrementar tanto la huella de carbono como las barreras de entrada para su implementación, circunstancias que impactan sobre los siguientes ODS de Naciones Unidas (Gamez, 2015):

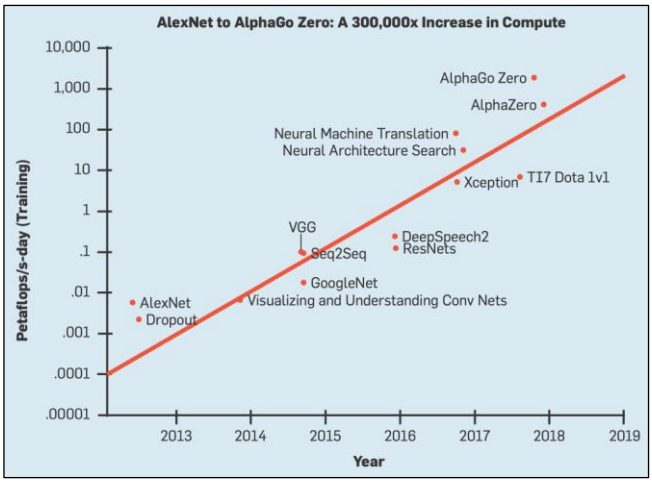


Figura 14. Incremento del coste computacional de la IA durante la década de 2010 (Schwartz et al., 2020)

- **Objetivo 7**, el cual pretende garantizar una **energía limpia y asequible**, debido a que el alto consumo energético asociado a los algoritmos de aprendizaje computacional afecta a su sostenibilidad y accesibilidad.

- **Objetivo 9**, centrado en la construcción de **infraestructuras resilientes y la promoción de la industria sostenible**, puesto que un elevado coste energético limita el acceso a las tecnologías avanzadas para las pymes y sectores desfavorecidos.
- **Objetivo 10**, sobre la **reducción de las desigualdades** en y entre los países, nuevamente debido a la barrera tecnológica y económica de acceso al aprendizaje computacional avanzado. Además, cabe destacar cómo las emisiones provocadas por los modelos varían significativamente entre los países debido a las fuentes de energía utilizadas por sus redes eléctricas (Lacoste et al., 2019) (Cowls et al., 2023), todo lo cual es susceptible de ampliar la brecha digital entre comunidades.

	Carbon emissions (CO_2eq)	Train Compute (FLOPS)	GPU	Training hours	Cloud Provider
South Africa (West)	942,330kg	3.14E+23	V100	3.11E+06	Microsoft Azure
India (South)	858,360kg	3.14E+23	V100	3.11E+06	Microsoft Azure
Australia (Central)	839,700kg	3.14E+23	V100	3.11E+06	Microsoft Azure
Europe (North)	578,460kg	3.14E+23	V100	3.11E+06	Microsoft Azure
South Korea (Central)	485,160kg	3.14E+23	V100	3.11E+06	Microsoft Azure
Brazil (South)	186,600kg	3.14E+23	V100	3.11E+06	Microsoft Azure
France (Central)	93,300kg	3.14E+23	V100	3.11E+06	Microsoft Azure

Figura 15. Coste ambiental del entrenamiento de GPT-3 en diferentes países
(Cowls et al., 2023)

- **Objetivo 12**, para la garantía de **modalidades de consumo y producción sostenibles**, ya que una alta demanda de recursos computacionales puede contribuir a prácticas de producción y consumo que ponen en peligro la subsistencia de las generaciones actuales y futuras.
- **Objetivo 13**, de **acción por el clima**, debido al incremento de las emisiones de efecto invernadero provocado por el entrenamiento de los modelos complejos de inteligencia artificial.

Sin embargo, el impacto de los algoritmos sobre los mencionados objetivos no es homogéneo, ya que, como ha sido expuesto durante el análisis de los algoritmos utilizados, el coste computacional aparejado a su ejecución difiere ampliamente entre ellos. Todo y que la literatura científica disponible se centra principalmente en la comparativa de algoritmos de aprendizaje supervisado (Verdecchia et al., 2022), tal como ha sido expuesto en apartados anteriores, la complejidad computacional de los algoritmos basados en distancia y densidad es más reducida que la de los jerárquicos y los modelos de redes neuronales (Patterson et al., 2021); sin embargo, el coste final dependerá en gran parte de la adecuada parametrización del algoritmo y del tamaño del *dataset*.

Finalmente, cabe destacar que uno de los principales problemas para la evaluación del impacto de los diferentes modelos es la **ausencia de métricas universales de sostenibilidad** adoptadas de forma general por la industria, si bien el aumento de la sensibilidad y del compromiso social han hecho emerger recientes propuestas en este sentido (Lieder et al., 2019) (Anthony et al., 2020) (Eilam et al., 2023) (Heguerte et al., 2023).

4. Implementación de la solución

4.1. Punto de partida

La solución ha sido implementada utilizando la versión 3.10 del **lenguaje** de programación Python. Para el desarrollo, se ha seleccionado un cuaderno Jupyter, ejecutado localmente sobre un **entorno** virtual creado con Conda, la herramienta de gestión de entornos y paquetes de la distribución de Python para ciencia de datos, Anaconda, en el cual se han instalado todas las bibliotecas necesarias. La información principal de partida se compone de tres **ficheros** de texto, proporcionados por el negocio, que abarcan el período desde enero de 2017 hasta julio de 2023; a saber:

- «CLIENTES.txt»: Contiene las características del perfil del cliente, así como información agregada procedente de otros archivos.
- «CONSULTAS.txt»: Con el detalle de empresas consultadas por los clientes como parte de su utilización del servicio.
- «VENTAS.txt»: Con el detalle de las compras llevadas a cabo por los clientes.

A estos archivos se suman otros dos de los cuales se hace un uso puntal:

- «CLIENTES_PF_EMAIL_RELACIONES.txt»: Igualmente proporcionado por el negocio, el cual contiene información sobre el número de empresas con que los clientes personas físicas mantienen alguna relación ejecutiva, así como el dominio del correo electrónico con que se registraron en el servicio. Nuevamente
- «DEPARTAMENTOS_DISTANCIA_PIB.txt»: De elaboración propia a partir de información pública, con el producto interior bruto (en adelante, «PIB») de los diferentes departamentos de Colombia donde están domiciliados los clientes, así como la distancia en km desde la capital de dicho departamento hasta la del país Bogotá.

4.2. Inspección de los datos

El fichero principal, «**CLIENTES**.txt», se compone de 9.512 filas y las siguientes 17 variables, de carácter categórico salvo cuando se indica lo contrario:

ID	Identificador numérico único del cliente.
FECHA_REGISTRO	Fecha que en que el cliente se registró en la web del servicio.
CANAL_REGISTRO	Canal por el que se captó al usuario.
FECHA_CLIENTE	Fecha de la primera compra del cliente.
CLIENTEPORCAMPAÑAEMAIL	Valor lógico que indica si la primera compra se produjo a raíz de una campaña de correo electrónico.

FORMAJURIDICA	Forma societaria del cliente en el caso de las empresas, o bien su condición de empresario individual o de persona física. Destacamos este atributo porque, como veremos, jugará un papel fundamental durante los pasos subsiguientes.
SECTOR	Codificación del sector de actividad del cliente.
DESC_SECTOR	Descripción aparejada al sector de actividad.
ESTADO	Situación actual de la actividad del cliente.
DEPARTAMENTO	Departamento de residencia en Colombia o su condición de residente extranjero.
TAMAÑO	Tamaño discretizado de la sociedad.
ANTIGUEDAD	Antigüedad discretizada desde la constitución de la sociedad.
DIASCLIENTE	Número de días transcurridas desde el registro del cliente hasta su primera compra.
CONSUMOSTOTAL	Número de consumos o consultas realizadas por el cliente a través del servicio.
EMPRESASUNICAS_CONSUL	Número de empresas únicas que han sido consultadas por el cliente a través del servicio.
NUM_COMPRAS	Número total de compras realizadas por el cliente.
IMPORTE_COMPRAS	Número que indica el gasto total realizado por el cliente mediante sus compras.

Tabla 2. Variables del fichero «CLIENTES.txt»

A través de su descripción, es posible clasificar estas variables en cuatro grandes **grupos**:

Relativas al **perfil empresarial** del cliente	Incluyen la forma jurídica, el sector, estado, departamento y tamaño.
Relativas al proceso de ***onboarding*** o proceso de integración en el servicio	Donde se inscribe la fecha de registro y de la primera compra, así como los días transcurridos entre una y otra, el canal de registro y si la primera compra se produjo a través de una campaña de *mailing*.
Relativas a las **consultas** llevadas a cabo por el usuario	Incluyendo el total de consumos y el número de empresas únicas consultadas, siendo éstas informaciones agregadas procedentes del fichero de consultas.
Relativas a las **compras** del usuario	Donde se incluye el número de compras hechas por éste, así como el importe total gastado en el servicio, informaciones agregadas del fichero de ventas.

Tabla 3. Clasificación de las variables del fichero «CLIENTES.txt»

Una primera inspección de los datos reveló dos particularidades llamativas:

1. Por una parte, el fichero contiene un elevado número de **valores nulos** concentrados en las variables «DEPARTAMENTO», «TAMAÑO» y «ANTIGUEDAD», todos los cuales se corresponden con las **personas físicas** de acuerdo con la variable «FORMAJURIDICA», y para los cuales esta clase de atributo no es aplicable. Además, las columnas relativas al sector y estado del cliente muestran siempre un valor único para este perfil de cliente.

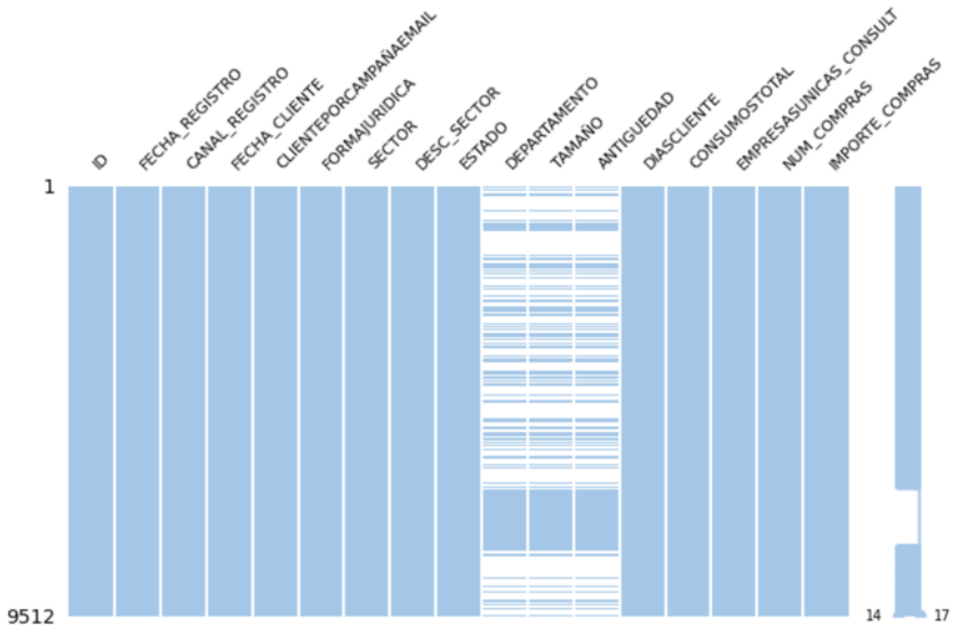


Figura 16. Análisis de valores ausentes

2. Por otro lado, las variables numéricas contienen un elevado número de **valores extremos**, la presencia de los cuales puede distorsionar gravemente el proceso de segmentación de los clientes.

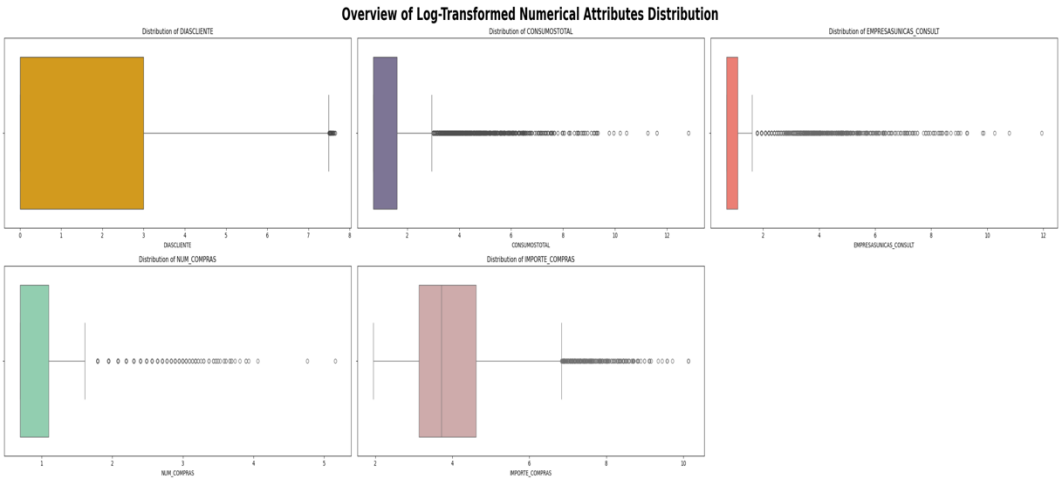


Figura 17. Distribución de los atributos numéricos transformados logarítmicamente de los clientes

Por su parte, el fichero de **consultas** se compone 910.738 filas y de los siguientes 9 atributos, categóricos salvo cuando se indica lo contrario:

IDCONSUMO	Identificador numérico único del consumo o consulta. Esta información se puede encontrar de forma agregada para cada cliente a través del campo CONSUMOSTOTAL del fichero de clientes.
ID	Identificador numérico único del cliente que llevó a cabo la consulta y que hace las veces de clave foránea.
FECHACONSUMO	Fecha en que se produjo la consulta.
PRODUCTO	Producto consultado.
EMPCONSUL_ID	Identificador único de la empresa sobre la cual se hizo la consulta. Esta información se puede encontrar de forma agregada para cada cliente a través del campo «EMPRESASUNICAS_CONSUL» del fichero de clientes.
EMPCONSUL_SECTOR	Sector de pertenencia de la empresa sobre la cual se hizo la consulta.
EMPCONSUL_TAMAÑO	Tamaño discretizado de la empresa sobre la cual se hizo la consulta.
EMPCONSUL_DEPARTAMENTO	Departamento de residencia de la empresa sobre la cual se hizo la consulta.
EMPCONSUL_ESTADO	Estado de la empresa sobre la cual se hizo la consulta.

Tabla 4. Variables del fichero «CONSULTAS.txt»

Mientras que el de **ventas** presenta 20.536 registros y los siguientes 5 campos, nuevamente categóricos salvo cuando se indica otra cosa:

ID	Identificador numérico único del cliente que realiza una compra y que hace las veces de clave foránea. Esta información se puede encontrar de forma agregada para cada cliente a través del campo «NUM_COMPRAS» del fichero de clientes.
FECHAVENTA	Fecha en que se produjo la compra/venta.
PRODUCTOCOMPRADO	Producto comprado/vendido.
CANALVENTA	El canal a través del cual se produjo la compra/venta.
IMPORTE	Importe numérico pagado por el producto comprado/vendido. Esta información se puede encontrar de forma agregada para cada cliente a través del campo «IMPORTE_COMPRAS» del fichero de clientes.

Tabla 5. Variables del fichero «VENTAS.txt»

Una vez se comprobó la inexistencia de valores duplicados, así como la coherencia entre los valores agregados contenidos en el fichero principal y la información desglosada en los ficheros de consultas y ventas, se procedió a la siguiente fase.

4.3. Preprocesado de los datos

En primer lugar, se consideró la clase de tratamiento que debía aplicarse a los mencionados ***outliers***. En este punto, el conocimiento del dominio jugó un papel relevante. Según el principio de Pareto o regla del 80/20, la mayor parte de los ingresos del negocio habitualmente provendrán de un número reducido de clientes, razón por la cual los valores atípicos aportan un significado relevante dentro de la tipología de la clientela del negocio. Por tal razón, se llevó a cabo una supresión moderada de los mismos, descartando del proceso de segmentación únicamente los que superaban el umbral de 7 veces la desviación estándar desde la media de los datos. Como resultado de esta contención, la **distribución** de los datos ni era simétrica ni cumplía con el criterio de normalidad, limitación que condicionó el trabajo posterior.

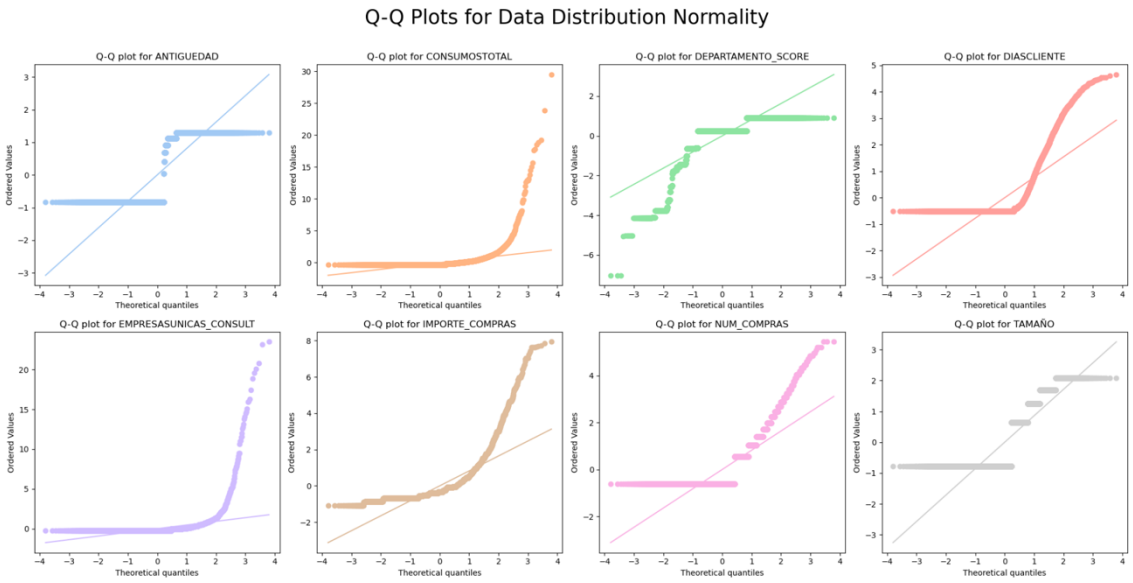


Figura 18. Gráficos Q-Q de la normalidad en la distribución de las variables numéricas

A continuación, como ha sido expuesto durante la descripción de los datos, el juego de datos se componía parcialmente de variables categóricas, las cuales requerían ser **codificadas numéricamente** para poder ser procesadas por el algoritmo de acuerdo con el siguiente criterio:

a) Aquellos cuyo contenido guarda una relación ordinal («ANTIGUEDAD» y «TAMAÑO»), fueron codificados a través de *OrdinalEncoder* asignando a cada valor único una numeración conforme a su gradación.

b) El resto fue codificado mediante *One-Hot Encoding* (en adelante, «OHE»), transformando cada categoría con n valores únicos en n columnas diferentes, representando cada columna uno de los valores posibles. El contenido de estas columnas es de carácter lógico, marcándose con un 1 cuando la observación corresponde a la categoría y con un 0 en el caso contrario.

Sin embargo, y dado el inevitable incremento de dimensionalidad que la técnica de OHE supone, para la codificación de la variable «DEPARTAMENTO» (referente a la región donde se domicilia el cliente) se procedió a su transformación mediante un proceso de ***feature engineering*** con objeto de otorgar significación a su contenido numérico. Para ello comenzó recopilando información relativa a los departamentos y se recogió en el fichero de creación propia «DEPARTAMENTOS_DISTANCIA_PIB.txt», con 31 entradas y 3 atributos:

DEPARTAMENTO	Nombre del departamento.
DISTANCIA_CAPITAL_DEPARTAMENTO	La distancia en kilómetros desde la capital del departamento hasta la capital del país, Bogotá; la cual nos ofrece información sobre su proximidad al epicentro de la economía e industria colombianas.
PIB_DEPARTAMENTO	PIB del departamento, el cuál es un indicativo de su actividad y salud económica.

Tabla 6. Variables del fichero «DEPARTAMENTOS_DISTANCIA_PIB.txt»

Una vez cargados los datos, éstos fueron seguidamente combinados en un único atributo llamado «DEPARTAMENTO_SCORE» mediante un procedimiento estructurado en dos pasos:

1. Estandarizando ambas variables para ambas medidas contribuyan de manera equitativa al nuevo atributo.
2. Considerando que, mientras que un mayor PIB es un síntoma de fuerte industrialización y acceso a un mercado relevante, en el caso de la distancia de la capital, cuanto mayor sea ésta, menor será el acceso a servicios y al mercado, por lo que la distancia debe ser valorada inversamente.

Tal como ha sido mencionado, para el caso de los clientes personas físicas se carecía tanto de la información relativa a las variables categóricas ordinales de antigüedad y tamaño, como de la relativa al departamento. Por este motivo se **imputaron** valores de acuerdo con 2 criterios diferenciados:

a) Para el caso de antigüedad y tamaño, debido a sus propias características diferenciadas de las de una sociedad o un empresario, les fue asignado el valor más bajo dentro de la gradación de la codificación ordinal.

b) En cuanto al departamento, debido a que la persona física reside sin duda en algún lugar, pero nos es desconocido, y teniendo presente la falta de normalidad de la distribución, el valor de la nueva variable «DEPARTAMENTO_SCORE» les fue asignado por imputación por la mediana.

A continuación, y con objeto de mitigar el impacto de la distribución asimétrica de los datos, se procedió a aplicar sobre éstos una **transformación** basada en la raíz cuadrada, la cual ofrece la ventaja de conservar intactos los valores binarios

correspondientes a las variables lógicas resultado del proceso de OHE. Finalmente, se **estandarizaron** los datos para facilitar su comparación y evitar distorsiones provocadas por la diferencia de escalas.

4.4. Análisis de colinealidad y valoración de variables

Tanto el cálculo del factor de inflación de la varianza o VIF, como el de la matriz de correlación, evidenciaron la **colinealidad** existente entre las personas físicas y un valor constante para las categorías referentes al estado, el sector y, en menor medida la antigüedad. De hecho, todas las variables detectadas hicieron referencia al perfil empresarial del cliente (forma jurídica, sector, estado, antigüedad), suprimiéndose las características redundantes.

Variable 1	Variable 2	Correlation
SECTOR_NOSECTOR	ESTADO_VIVA	1.00
FORMAJURIDICA_PERSONA FISICA	ESTADO_VIVA	1.00
FORMAJURIDICA_PERSONA FISICA	SECTOR_NOSECTOR	1.00
FORMAJURIDICA_EMBAJADAS Y ORGANISMOS INTERNACI...	SECTOR_ACTIVIDADES DE ORGANIZACIONES Y ENTIDAD...	1.00
ANTIGUEDAD	FORMAJURIDICA_PERSONA FISICA	-0.99
ANTIGUEDAD	ESTADO_VIVA	-0.99
ANTIGUEDAD	SECTOR_NOSECTOR	-0.99
CONSUMOSTOTAL	EMPRESASUNICAS_CONSULT	0.98
TAMAÑO	ANTIGUEDAD	0.95
TAMAÑO	ESTADO_VIVA	-0.94
TAMAÑO	SECTOR_NOSECTOR	-0.94
TAMAÑO	FORMAJURIDICA_PERSONA FISICA	-0.94
TAMAÑO	FORMAJURIDICA_SOCIEDAD	0.81
ANTIGUEDAD	FORMAJURIDICA_SOCIEDAD	0.72
FORMAJURIDICA_PERSONA FISICA	FORMAJURIDICA_SOCIEDAD	-0.72
FORMAJURIDICA_SOCIEDAD	SECTOR_NOSECTOR	-0.72
FORMAJURIDICA_SOCIEDAD	ESTADO_VIVA	-0.72

Figura 19. Análisis de variables con alta correlación

Por otra parte, se aplicó PCA sobre los datos, no con el propósito habitual de reducir la dimensionalidad del *dataset* manteniendo una importante proporción de la varianza, sino de obtener información sobre los atributos que contribuían en mayor medida a dicha varianza en el conjunto de datos. Nuevamente, con pequeñas excepciones, la mayoría de atributos en que se concentraba el peso de los componentes principales se refiereía tanto al perfil empresarial del cliente como, en menor grado, a su proceso de *onboarding* en el servicio.

PC	Main Feature	Explained Variance Ratio	Cumulative Explained Variance Ratio
PC1	ANTIGUEDAD	12.13%	12.13%
PC2	CONSUMOSTOTAL	6.77%	18.90%
PC3	FORMAJURIDICA_EMPRESARIO	4.59%	23.49%
PC4	FORMAJURIDICA_ESAL	3.71%	27.20%
PC5	DIASCLIENTE	3.51%	30.71%
PC6	CANAL_REGISTRO_SEM	3.08%	33.79%
PC7	FORMAJURIDICA_SOCIEDAD EXTRANJERA	2.98%	36.78%
PC8	SECTOR_ACTIVIDADES PROFESIONALES, CIENTÍFICAS ...	2.87%	39.64%
PC9	SECTOR_COMERCIO AL POR MAYOR Y AL POR MENOR; R...	2.82%	42.46%
PC10	ESTADO_EXTINGUIDA	2.71%	45.18%
PC11	SECTOR_INDUSTRIAS MANUFACTURERAS	2.69%	47.86%
PC12	SECTOR_INFORMACIÓN Y COMUNICACIONES	2.62%	50.48%
PC13	SECTOR_INFORMACIÓN Y COMUNICACIONES	2.60%	53.08%
PC14	SECTOR_ACTIVIDADES INMOBILIARIAS	2.58%	55.66%
PC15	SECTOR_ACTIVIDADES DE SERVICIOS ADMINISTRATIVO...	2.55%	58.20%
PC16	SECTOR_COMERCIAL / INDUSTRIAL NO DEFINIDA	2.54%	60.74%
PC17	SECTOR_AGRICULTURA, GANADERÍA, CAZA, SILVICULT...	2.53%	63.27%
PC18	SECTOR_EDUCACIÓN	2.53%	65.80%
PC19	SECTOR_DISTRIBUCIÓN DE AGUA; EVACUACIÓN Y TRAT...	2.52%	68.31%
PC20	SECTOR_SUMINISTRO DE ELECTRICIDAD, GAS, VAPOR ...	2.52%	70.83%
PC21	SECTOR_EDUCACIÓN	2.51%	73.34%
PC22	SECTOR_ADMINISTRACIÓN PÚBLICA Y DEFENSA; PLANE...	2.50%	75.85%
PC23	SECTOR_ACTIVIDADES DE LOS HOGARES INDIVIDUALES...	2.49%	78.34%
PC24	FORMAJURIDICA_EMBAJADAS Y ORGANISMOS INTERNACI...	2.44%	80.78%
PC25	ESTADO_EXTINGUIDA	2.33%	83.11%
PC26	ESTADO_INSOLVENTE	2.23%	85.34%
PC27	ESTADO_INSOLVENTE	2.20%	87.54%
PC28	SECTOR_EXPLOTACIÓN DE MINAS Y CANTERAS	2.12%	89.66%
PC29	SECTOR_ALOJAMIENTO Y SERVICIOS DE COMIDA	2.03%	91.69%
PC30	DEPARTAMENTO_SCORE	1.90%	93.59%
PC31	CLIENTEPORCAMPAÑAEMAIL_sí	1.88%	95.47%

Figura 20. Análisis de componentes principales y sus características más relevantes

4.5. Primera fase de la segmentación

Una vez preprocesados los datos se procedió a aplicar el algoritmo de agrupamiento sobre el *dataset*, optándose por la utilización del algoritmo ***k-medoids*** $\left(\text{Minimize} \quad \sum_{j=1}^{k} \sum_{x_i \in C_j} d\left(x_i, m_j\right)\right)$, por combinar la eficiencia de los algoritmos basados en distancias con un mayor grado de robustez frente a los valores extremos presentes en nuestro juego de datos, utilizando los puntos más representativos del grupo en lugar de las medias, lo que minimiza la suma de las distancias absolutas. Se utilizó como métrica la **distancia Manhattan** $(d(x,y) = \sum_{i=1}^{n} |x_i - y_i|)$, nuevamente por ser menos sensible a los valores atípicos, puesto que ésta calcula la distancia sumando las diferencias absolutas en cada dimensión, en lugar de elevarlas al cuadrado como en la métrica euclidiana, reduciendo así el impacto de los valores extremos.

Se utilizó un criterio doble como método de **determinación del valor de** k, calculándose:

1. El coeficiente de la silueta, método particularmente eficiente cuando el número posible de clústeres resultantes es inicialmente desconocido o cuando los límites entre las agrupaciones no están particularmente bien definidos.
2. El índice de Calinski-Harabasz, al tratarse de una métrica menos sensible a la densidad y tamaño de los clústeres, así como a la presencia de valores extremos dado que, cuando los clústeres presentan tamaños y densidades diferentes, los resultados del coeficiente de la silueta pueden no ser del todo óptimos.

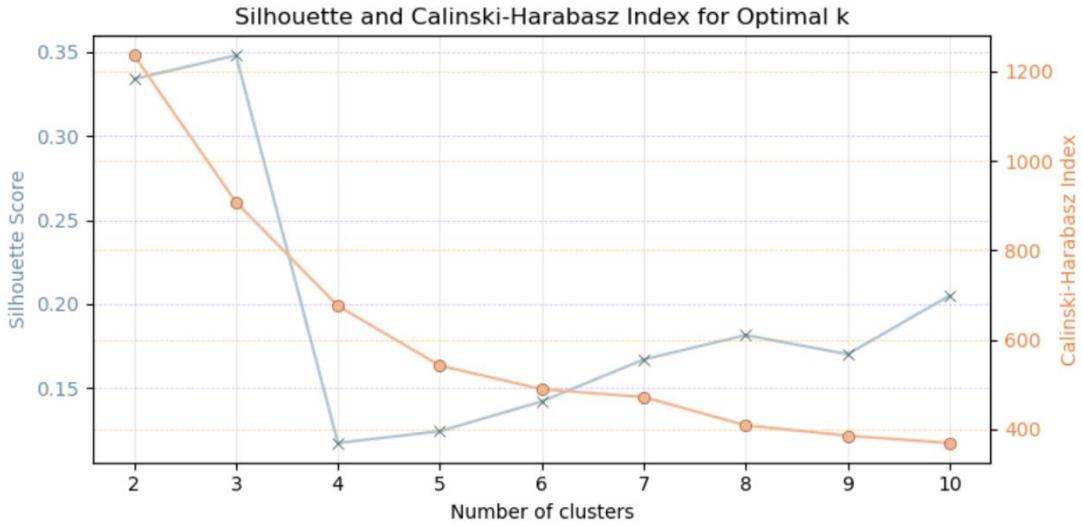


Figura 21. Determinación del valor de *k* para la primera fase de la segmentación

Como se observa en la gráfica, de acuerdo con el resultado del coeficiente de la silueta, el valor óptimo estaría entre 2 y 3 clústeres con una pequeña diferencia a favor de una segmentación en 3 grupos. Sin embargo, el índice de Calinski-Harabasz apuntaba a un marcado descenso de los resultados para $\boldsymbol{k} = 3$. Tras diferentes pruebas, la distribución destacada mediante Calinski-Harabasz produjo unos clústeres mejor definidos como se puede comprobar a continuación, motivo por el cual se optó por una agrupación en 2 clústeres durante esta primera fase de la segmentación.

Analizada la desviación estándar como medida de dispersión para identificar las características con mayor variación de un *medoide* a otro, se detectaron como variables más influyentes nuevamente las relacionadas con la forma jurídica.

Feature	Baseline Score	Score Without	Score Change
FORMAJURIDICA_PERSONA FISICA	0.333848	0.313858	0.019990
ANTIGUEDAD	0.333848	0.318582	0.015266
TAMAÑO	0.333848	0.319829	0.014019
FORMAJURIDICA_SOCIEDAD	0.333848	0.324457	0.009391
SECTOR_COMERCIO AL POR MAYOR Y AL POR MENOR; R...	0.333848	0.326870	0.006978
SECTOR_ACTIVIDADES INMOBILIARIAS	0.333848	0.328171	0.005678
SECTOR_SUMINISTRO DE ELECTRICIDAD, GAS, VAPOR ...	0.333848	0.329183	0.004666
SECTOR_ACTIVIDADES PROFESIONALES, CIENTÍFICAS ...	0.333848	0.329330	0.004518
FORMAJURIDICA_EMPRESARIO	0.333848	0.329408	0.004440
SECTOR_INDUSTRIAS MANUFACTURERAS	0.333848	0.330213	0.003635
SECTOR_ALOJAMIENTO Y SERVICIOS DE COMIDA	0.333848	0.330378	0.003470
SECTOR_INFORMACIÓN Y COMUNICACIONES	0.333848	0.331379	0.002469
ESTADO_INACTIVA	0.333848	0.331487	0.002361
SECTOR_ACTIVIDADES DE SERVICIOS ADMINISTRATIVO...	0.333848	0.331891	0.001958
SECTOR_CONSTRUCCIÓN	0.333848	0.332237	0.001611
DEPARTAMENTO_SCORE	0.333848	0.332443	0.001406
SECTOR_TRANSPORTE Y ALMACENAMIENTO	0.333848	0.332722	0.001126
SECTOR_ACTIVIDADES FINANCIERAS Y DE SEGUROS	0.333848	0.332822	0.001027
SECTOR_OTRAS ACTIVIDADES DE SERVICIOS	0.333848	0.333463	0.000385
SECTOR_ACTIVIDADES DE ATENCIÓN DE LA SALUD HUM...	0.333848	0.333508	0.000340
SECTOR_COMERCIAL / INDUSTRIAL NO DEFINIDA	0.333848	0.333698	0.000150
SECTOR_AGRICULTURA, GANADERÍA, CAZA, SILVICULT...	0.333848	0.333799	0.000049

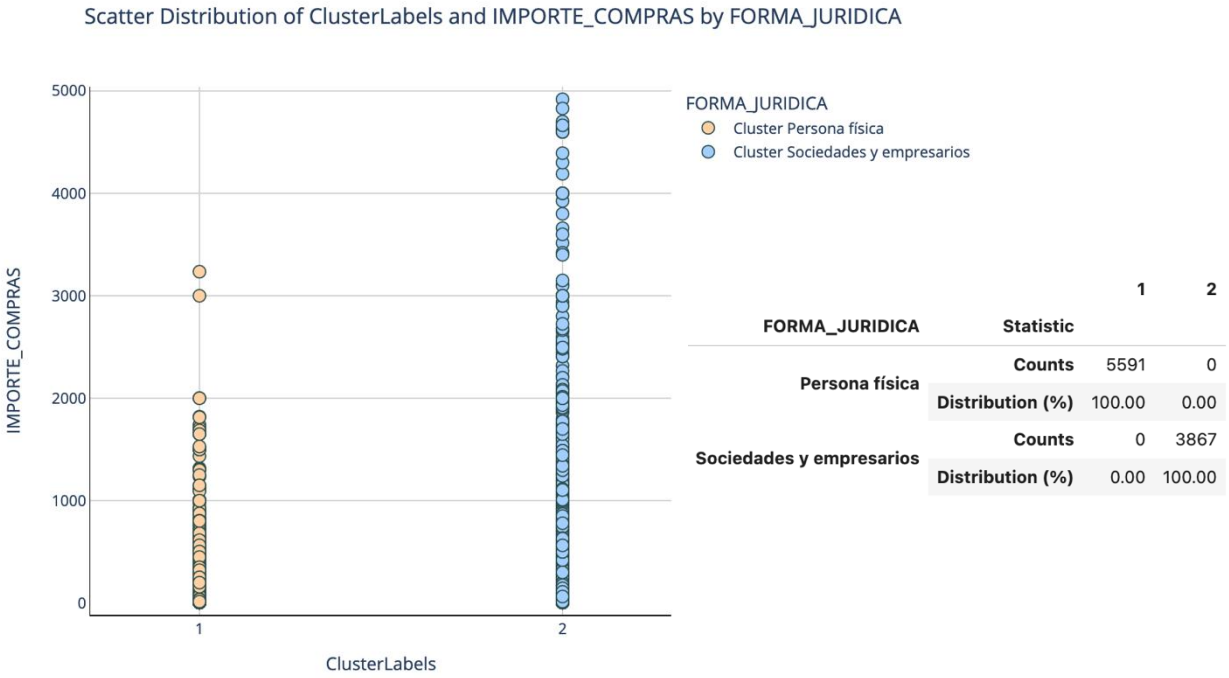
Figura 22. Importancia de las características para la primera fase de la segmentación

Y lo mismo sucedió al evaluar las variables más significativas durante el proceso de formación de los agrupamientos a través de la técnica de *Clustering Feature Importance* (en adelante, «CFI»), descubriendo cómo variaba la calidad del *clustering* según se eliminaba una determinada característica del conjunto de datos.

Variable	Importance
FORMAJURIDICA_SOCIEDAD	1.14
FORMAJURIDICA_PERSONA FISICA	1.02
ANTIGUEDAD	0.87
TAMAÑO	0.72
DEPARTAMENTO_SCORE	0.33
IMPORTE_COMPRAS	0.12

Figura 23. Análisis de las variables críticas para la primera fase de la segmentación

Así, los segmentos resultantes concentraron las personas físicas por un lado, y los empresarios individuales y toda clase de sociedades por el otro, produciendo dos agrupaciones perfectamente delimitadas a través de las cuales ambos colectivos quedaron definidos y separados.



FORMA_JURIDICA	Statistic	1	2
Persona física	Counts	5591	0
	Distribution (%)	100.00	0.00
Sociedades y empresarios	Counts	0	3867
	Distribution (%)	0.00	100.00

Figura 24. Resultado de la primera fase de la segmentación

El algoritmo se encargó así de discriminar la base usuaria según las características distintivas asociadas a su perfil empresarial, siendo la distinción más obvia la que divide a la clientela entre personas físicas por un lado y diferentes formas societarias y empresarios por el otro, y apuntando a un estudio y segmentación diferenciados de los grupos para la segunda fase. Esta primera etapa de la segmentación funcionó así como una lente de aumento, permitiéndonos en el siguiente paso poner el foco separadamente sobre cada uno de los clústeres para estudiar sus hábitos de compra de acuerdo con sus características intrínsecas, paso que facilitaría la creación de segmentos más homogéneos durante la segunda etapa y evitaría distorsiones provocadas por posibles desbalanceos en los datos.

4.6. Preprocesado durante la segunda fase

4.6.1. Creación de nuevas variables

Como resultado de la primera fase se dispuso de dos agrupaciones definidas sobre las que trabajar separadamente con objeto de completar el proceso de segmentación. Sin embargo, antes de dividir los datos, se procedió a la definición de nuevas variables. Prosiguiendo el proceso iniciado con la creación del atributo «DEPARTAMENTO_SCORE», el objetivo de esta etapa de *feature engineering* consistió en utilizar conocimientos del dominio para crear nuevos atributos, transformar los existentes en otros más significativos o ponerlos en relación con objeto de descubrir patrones ocultos en los datos.

Cabe puntualizar que las nuevas variables diseñadas no serían empleadas durante la segunda etapa de la segmentación, sino que fueron construidas con el

propósito de que su contenido pudiese revelar información valiosa de cada uno de los segmentos creados una vez completado el proceso.

Se dividió así el procedimiento en una primera etapa de *feature engineering* aplicado al juego de datos conjunto, y una segunda en la que se añadió nueva información únicamente al segmento de las personas físicas.

Antes de proceder al nuevo preprocesado sin embargo, se restablecieron los valores originales de las variables sin transformación ni escalado.

Diseño de nuevas variables para el conjunto de datos

Durante esta parte del proceso se recuperó la información desglosada contenida en los ficheros de consultas y ventas, a través los cuales se construyeron los siguientes atributos. Es necesario recordar que el ámbito temporal de los datos objeto de estudio se circunscribía al período que va desde enero de 2017 hasta julio de 2023. El análisis llevado a cabo asumió que trabajaba con información reciente, tomando como punto de partida el momento siguiente a la finalización de dicho período.

Obtenida exclusivamente del fichero «CLIENTES.txt».	
VIDACLIENTE	Contiene la antigüedad del cliente en el servicio.
Obtenidas de la combinación entre los ficheros «CLIENTES.txt» y «CONSULTAS.txt.».	
ENGAGEMENT	Variable que relaciona la recencia, la frecuencia y la diversidad presente en las consultas del usuario, considerando cuántas empresas de diferentes estados, sectores y tamaños ha consultado.
TENDENCIA_FRECUENCIA_CONSULT	Evalúa la tendencia de consultas de cada cliente a través de un modelo de regresión lineal.
Obtenidas de la combinación entre los ficheros «CLIENTES.txt» y «VENTAS.txt.».	
AOV	Contiene el *Average Order* Value o importe promediado por compra.
CLV	Contiene el *Customer Lifetime Value* o contribución financiera esperada de cada cliente a lo largo de toda su relación comercial con la empresa.
DIVERSIDAD_COMPRAS	Incluye la diversidad de productos adquiridos por el cliente.
TENDENCIA_FRECUENCIA_VENTAS	Evalúa la tendencia de compra de cada cliente a través de un modelo de regresión lineal.

Tabla 7. Diseño de nuevas variables

Incorporación de nueva información al segmento de personas físicas

Durante esta segunda parte, el proceso se centró en la creación de nuevos atributos específicos para las personas físicas. El conocimiento del dominio es clave para entender la tipología de cliente que encontramos detrás de la forma jurídica «PERSONA FISICA» en el *dataset*, ya que los usuarios que no se registran como sociedad ni empresario engloban perfiles mixtos con motivaciones muy diversas, y que

van desde personas interesadas en una oferta de trabajo o en realizar una inversión, a responsables de grandes empresas que prefieren no despertar suspicacias o, simplemente, mantener su privacidad.

Es por ello que, con objeto de aportar nuevos datos sobre cada uno de los perfiles, nos servimos de un nuevo fichero llamado «CLIENTES_PF_EMAIL_RELACIONES.txt» que contenía la siguiente información en 5.600 filas y 3 columnas:

ID	Identificador numérico único del cliente que realiza una compra y que hace las veces de clave foránea.
EMPRESASCONRELACION	Número de empresas con las que el cliente mantiene algún vínculo ejecutivo o en las que ejerce algún cargo directivo.
TIPODOMINIOEMAIL	Tipo del dominio del correo electrónico con que el cliente se registró en el servicio, el cual puede corresponder a un proveedor gratuito común, a un dominio de una institución educativa o a un dominio corporativo.

Tabla 8. Variables del fichero «CLIENTES_PF_EMAILRELACIONES.txt»

4.6.2. Selección, limpieza y transformación de los datos

Recordemos que el peso de la primera fase de la segmentación recayó sobre características definitorias del perfil empresarial del cliente, dando como resultado dos clústeres diferenciados, uno de los cuales agrupó la totalidad de sociedades y empresarios, y el otro las personas físicas. Para la siguiente etapa del agrupamiento sin embargo se descartaron las características relacionadas con el perfil, el *onboarding* y los consumos, **seleccionando** únicamente aquellas variables que definían de manera directa el valor económico del cliente, como son:

- IMPORTE_COMPRAS: El importe total gastado en el servicio.
- NUM_COMPRAS: El número total de compras realizadas.

A partir de este punto, el juego de datos se dividió en dos de acuerdo con los segmentos de la primera etapa y se llevaron a cabo nuevas tareas de preprocesado de forma similar a como se desarrollaron durante la primera parte de la segmentación:

1. Para comenzar, se hizo un segundo recorte de los **valores extremos**, dado que cada segmento tenía sus propias características y distribución. Nuevamente se trató de una limpieza muy moderada que afectó únicamente a los que superaban el umbral de 7 veces la desviación estándar des de la media de los datos.
2. Una **transformación** de los datos basada en la raíz cuadrada para mitigar el impacto de los *outliers* conservados.
3. Una **estandarización** de los valores para evitar distorsiones fruto de la diferencia de escalas.

4.7. Segunda fase de la segmentación

Para la segunda etapa de la segmentación se empleó un algoritmo aglomerativo con enlace completo ($d(A,B) = \max_{a\in A, b\in B} d\,(a,b)$), el cual mide la proximidad entre dos grupos basándose en la máxima distancia entre cualquier par de puntos de dichos grupos, con el objetivo de que los segmentos resultantes sean los más distintivos y cohesionados posibles. De esta forma, resultaría más sencillo diseñar estrategias de ventas que se dirigieran a un grupo específico de clientes.

Respecto a la métrica utilizada, se optó por la afinidad del coseno $\left(\text{cosine\_affinity}(\boldsymbol{x},\boldsymbol{y}) = 1 - \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}\right)$. Si bien esta técnica es de uso habitual en juegos de datos de alta dimensionalidad, hecho que contrasta con la selección reducida de características para esta segunda fase de la segmentación, su capacidad para medir la similitud en el comportamiento de los usuarios permitió agrupar segmentos de clientes con niveles de compra y gasto equivalentes con independencia de la magnitud.

4.7.1. Segmentación de sociedades y empresarios

De las gráficas resultantes de la aplicación del algoritmo aglomerativo se concluyó que, tanto el coeficiente de silueta como el índice de Calinski-Harabasz, valoraban muy positivamente un valor de $\boldsymbol{k}$ entre 3 y 6, mientras que el dendrograma reveló igualmente agrupaciones definidas y bien delimitadas para estos casos.

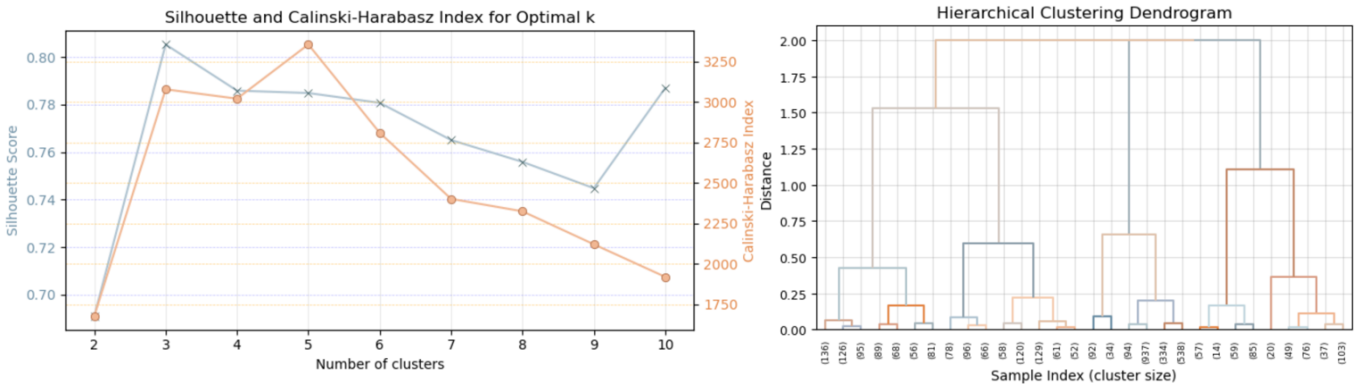
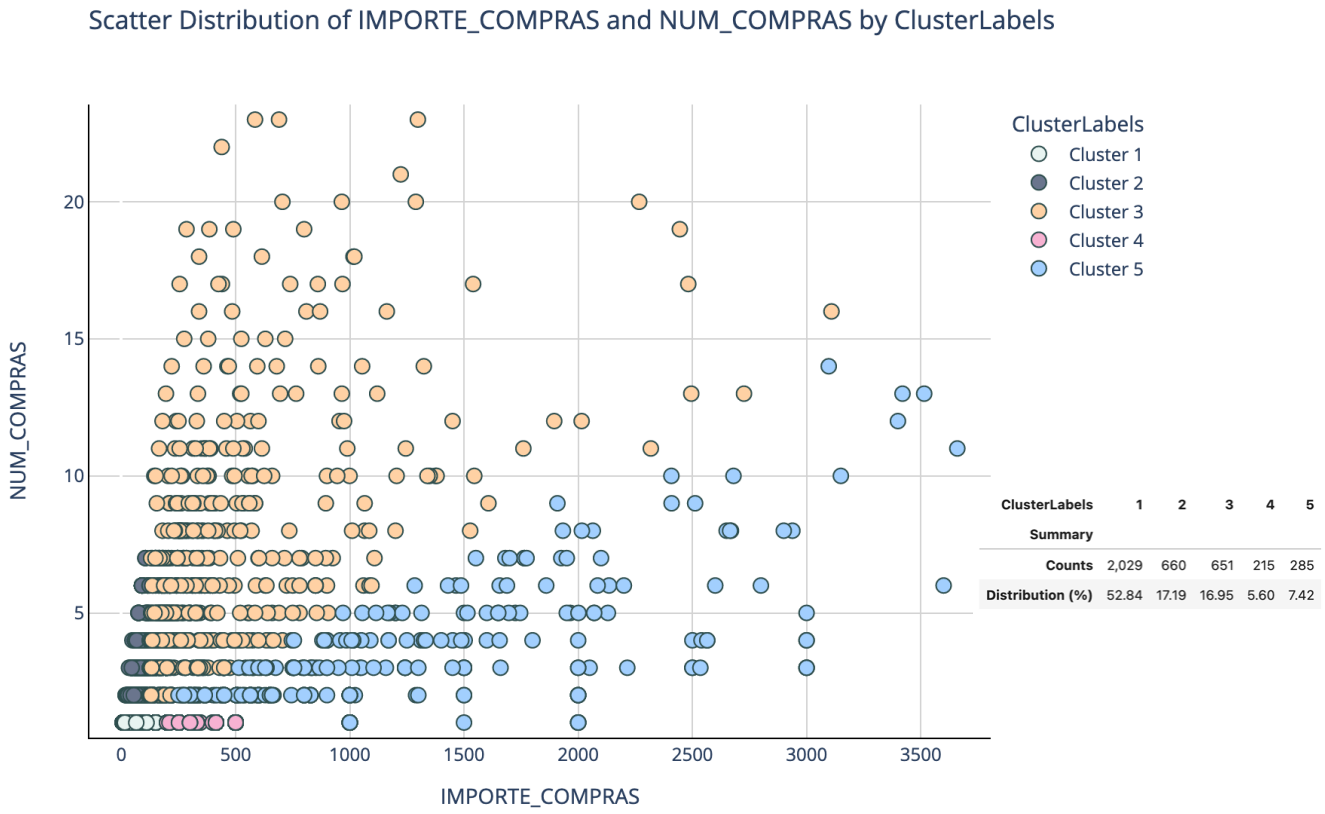


Figura 25. Determinación del valor de k para la segmentación de sociedades y empresarios

En esta ocasión se priorizó de nuevo el índice de Calinski-Harabasz, menos sensible a los valores extremos, optando por un agrupamiento en cinco segmentos, cifra que supone un nivel de granularidad manejable pero suficientemente diverso para el diseño de estrategias de ventas diferenciadas, obteniendo la segmentación que se muestra a continuación acompañada de los correspondientes perfiles prototípicos de cada clúster.



ClusterLabels Summary	1	2	3	4	5
Counts	2,029	660	651	215	285
Distribution (%)	52.84	17.19	16.95	5.60	7.42

Figura 26. Segmentación de sociedades y empresarios

	1	2	3	4	5
ANTIGUEDAD	Más de 10 Años	Más de 10 Años	Más de 10 Años	Más de 10 Años	Más de 10 Años
AOV mean	38.39	28.62	64.10	344.07	399.59
CANAL	Directorios	Directorios	WEB	WEB	WEB
CLIENTEPORCAMPAÑAEMAIL	no	no	no	no	no
CLV mean	544.35	405.86	908.89	4,878.44	5,665.71
CONSUMOSTOTAL mean	1.62	3.83	32.62	288.51	593.47
DEPARTAMENTO	BOGOTA	BOGOTA	BOGOTA	BOGOTA	BOGOTA
DEPARTAMENTO_SCORE mean	-0.35	-0.22	-0.16	-0.15	-0.09
DIASCLIENTE mean	114.97	132.18	79.86	70.75	83.44
DIVERSIDAD_COMPRAS mean	1.00	1.01	1.32	1.00	1.43
DIVERSIDAD_EMPRESAS_CONSULT mean	1.28	2.13	14.34	245.70	347.26
DIVERSIDAD_ESTADOS_CONSULT mean	1.04	1.15	1.79	2.96	3.70
DIVERSIDAD_SECTORES_CONSULT mean	1.09	1.36	3.15	5.79	7.92
DIVERSIDAD_TAMAÑOS_CONSULT mean	1.08	1.34	2.60	3.25	4.34
EMPRESASUNICAS_CONSULT mean	1.28	2.13	14.34	245.70	347.26
ENGAGEMENT mean	0.35	0.42	1.01	4.59	8.10
ESTADO	ACTIVA	ACTIVA	ACTIVA	ACTIVA	ACTIVA
FORMAJURIDICA	SOCIEDAD	SOCIEDAD	SOCIEDAD	SOCIEDAD	SOCIEDAD
FRECUENCIA_CONSULT mean	1.62	3.83	32.62	288.51	593.47
IMPORTE_COMPRAS mean	38.39	70.92	373.78	344.07	1,204.59
NUM_COMPRAS mean	1.00	2.59	6.19	1.00	3.46
RECENCIA_CONSULT mean	1,150.80	1,052.18	707.75	1,124.51	540.66
SECTOR	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...
TAMAÑO	MICRO	MICRO	PEQUEÑA	MICRO	MEDIANA
TENDENCIA_FRECUENCIA_CONSULT mean	0.00	0.01	-0.05	-0.59	-0.03
TENDENCIA_FRECUENCIA_VENTAS mean	0.00	-0.01	-0.14	0.00	-0.07
VIDACLIENTE mean	1,316.18	1,446.77	1,352.39	1,362.69	1,305.77

Figura 27. Perfiles prototípicos de los segmentos de sociedades y empresarios

De su análisis es posible concluir que:

- Aproximadamente la mitad de los clientes del *dataset* se concentraron en un primer clúster caracterizado por un número moderado de compras y un importe modesto invertido en el servicio.
- Los siguientes dos clústeres (2 y 3) aglutinaron perfiles donde, tanto las cifras de compras como la de importe, iban en progresivo crecimiento.
- En cambio, los clústeres 4 y 5 agruparon clientela con un dispendio más elevado, pero con un número de compras más reducidos.
- Tanto en el clúster 3 como el 5, donde se concentraron el mayor número de compras y de importe gastado respectivamente, fueron por ello donde coincidieron los valores extremos conservados para cada una de las variables, «IMPORTE_COMPRAS» y «NUM_COMPRAS».

4.7.2. Segmentación de personas físicas

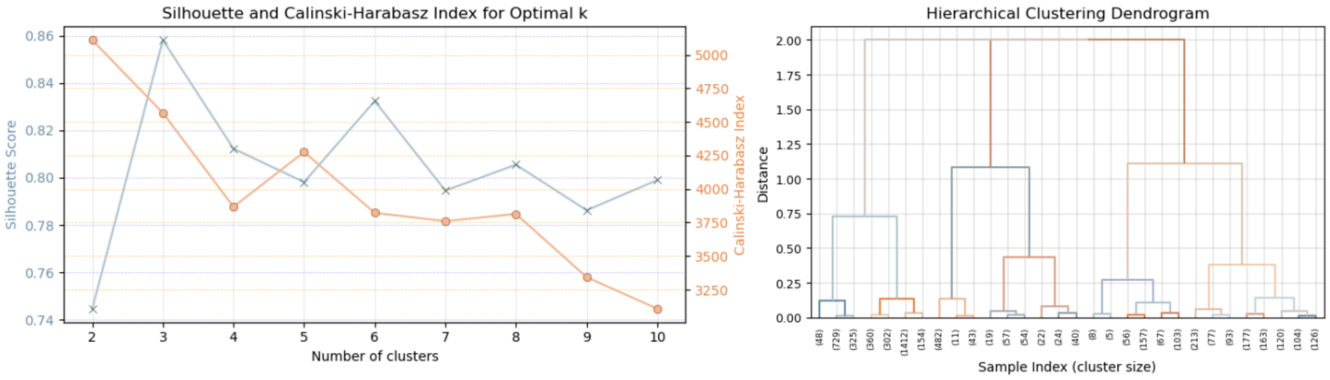


Figura 28. Determinación del valor de *k* para la segmentación de personas físicas

Como se puede observar en las gráficas, las estimaciones del coeficiente de la silueta y el índice de Calinski-Harabasz no fueron coincidentes para todos los puntos, siendo $\boldsymbol{k}$ = 3 aquel donde la valoración era considerablemente alta de acuerdo con ambas métricas, dando como resultado la siguiente segmentación con sus correspondientes perfiles prototípicos.

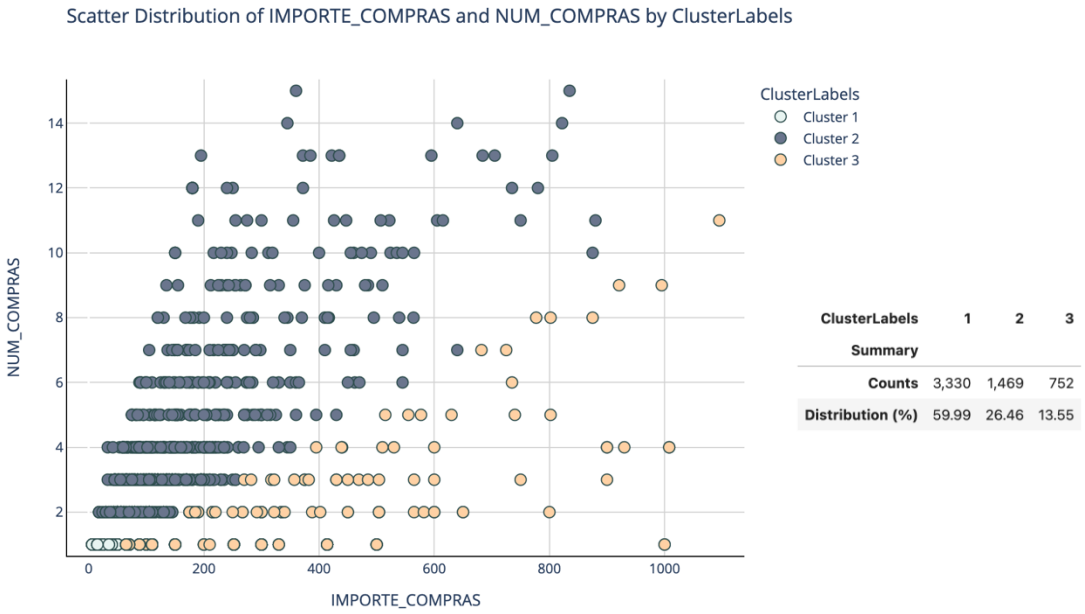


Figura 29. Segmentación de personas físicas

	1	2	3
ANTIGUEDAD	No aplicable	No aplicable	No aplicable
AOV mean	23.70	34.12	111.23
CANAL	Directorios	Directorios	WEB
CLIENTEPORCAMPAÑAEMAIL	no	no	no
CLV mean	336.03	483.78	1,577.09
CONSUMOSTOTAL mean	1.42	7.05	37.90
DEPARTAMENTO	No aplicable	No aplicable	No aplicable
DEPARTAMENTO_SCORE mean	0.18	0.18	0.18
DIASCLIENTE mean	94.57	98.76	90.50
DIVERSIDAD_COMPRAS mean	1.00	1.03	1.08
DIVERSIDAD_EMPRESAS_CONSULT mean	1.15	3.26	26.70
DIVERSIDAD_ESTADOS_CONSULT mean	1.03	1.28	1.41
DIVERSIDAD_SECTORES_CONSULT mean	1.06	1.67	2.03
DIVERSIDAD_TAMAÑOS_CONSULT mean	1.05	1.55	1.54
EMPRESASCONRELACION mean	1.59	2.57	1.93
EMPRESASUNICAS_CONSULT mean	1.15	3.26	26.70
ENGAGEMENT mean	0.34	0.51	0.91
ESTADO	VIVA	VIVA	VIVA
FORMAJURIDICA	PERSONA FISICA	PERSONA FISICA	PERSONA FISICA
FRECUENCIA_CONSULT mean	1.42	7.05	37.90
IMPORTE_COMPRAS mean	23.70	118.07	148.52
NUM_COMPRAS mean	1.00	3.37	1.26
RECENCIA_CONSULT mean	1,116.91	953.45	1,056.84
SECTOR	NOSECTOR	NOSECTOR	NOSECTOR
TAMAÑO	No aplicable	No aplicable	No aplicable
TENDENCIA_FRECUENCIA_CONSULT mean	-0.01	0.00	-0.19
TENDENCIA_FRECUENCIA_VENTAS mean	0.00	-0.05	0.01
TIPODOMINIOEMAIL	GOO-MS-YAH-APP	GOO-MS-YAH-APP	GOO-MS-YAH-APP
VIDACLIENTE mean	1,250.56	1,362.54	1,242.01

Figura 30. Perfiles prototípicos de los segmentos de personas físicas

De los segmentos formados y sus correspondientes perfiles podemos extraer las siguientes conclusiones:

- También en esta ocasión el **primer clúster** aglutinaba la mayoría de los clientes, los cuales se caracterizaron por un modesto número de compras e importe invertido en el servicio.
- Por su parte, los **clústeres 2 y 3** recogieron los clientes con un mayor número de compras y un mayor importe gastado respectivamente. Cabe destacar cómo el índice de Calinski-Harabasz apuntaba a la posibilidad de agrupar la clientela en 2 clústeres, y tras diferentes pruebas se pudo constatar que en ese caso los clientes alejados del grueso de las personas físicas, tanto en número como en importe de compras, quedaban aglutinados en un único clúster con valores crecientes en ambas direcciones. Se decidió sin embargo la división en 3 grupos al permitir al negocio adoptar una estrategia diferenciada para hábitos de compra nítidamente diferentes.
- Respecto al interrogante planteado durante el apartado de «Incorporación de nueva información al clúster de personas físicas», y en el que se planteó si una persona física con un amplio número de relaciones con otras compañías podría condicionar la pertenencia a un clúster con un mayor importe de compras, los resultados no son concluyentes, pero parecen negarlo. Tal vez optar por una mayor granularidad en el número de agrupaciones podría aportar una perspectiva diferente.

4.7.3. Resultado final del proceso de segmentación

Finalmente, el proceso de **segmentación se concluyó** con los siguientes pasos:

1. Se integraron los resultados de la segunda fase que segmentaban por separado el grupo de sociedades y empresarios por un lado, y el de las personas físicas por el otro.
2. Se recuperaron los clientes descartados como *outliers* a lo largo del proceso para conformar un segmento de clientes VIP con un monto de compra, recurrencia y consumos muy por encima del resto, el cual merece un tratamiento comercial específico.
3. Los segmentos fueron reetiquetados ordinalmente de forma ascendente de acuerdo con la media del importe de compras de cada grupo.
4. Se exportó el fichero de clientes original, «CLIENTES.txt», bajo el nuevo nombre de «CLIENTES_ETIQUETADOS.txt» y con una nueva columna llamada «SEGMENTO» que contenía las etiquetas.

	1	2	3	4	5	6	7	8	9
ANTIGUEDAD	No aplicable	Más de 10 Años	Más de 10 Años	No aplicable	No aplicable	Más de 10 Años	Más de 10 Años	Más de 10 Años	Más de 10 Años
CANAL	Directorios	Directorios	Directorios	Directorios	WEB	WEB	WEB	WEB	WEB
CLIENTEPORCAMPAÑAEMAIL	no	no	no	no	no	no	no	no	no
CONSUMOSTOTAL mean	1.42	1.62	3.83	7.05	37.90	288.51	32.62	593.47	5,493.93
DEPARTAMENTO	No aplicable	BOGOTA	BOGOTA	No aplicable	No aplicable	BOGOTA	BOGOTA	BOGOTA	BOGOTA
DIASCLIENTE mean	94.57	114.97	132.17	98.76	90.49	70.74	79.87	83.43	48.51
EMPRESASUNICAS_CONSULT mean	1.15	1.28	2.13	3.26	26.70	245.70	14.34	347.26	2,370.16
ESTADO	VIVA	ACTIVA	ACTIVA	VIVA	VIVA	ACTIVA	ACTIVA	ACTIVA	ACTIVA
FORMAJURIDICA	PERSONA FISICA	SOCIEDAD	SOCIEDAD	PERSONA FISICA	PERSONA FISICA	SOCIEDAD	SOCIEDAD	SOCIEDAD	SOCIEDAD
IMPORTE_COMPRAS mean	23.70	38.39	70.92	118.07	148.52	344.07	373.78	1,204.59	3,722.84
NUM_COMPRAS mean	1.00	1.00	2.59	3.37	1.26	1.00	6.19	3.46	19.27
SECTOR	NOSECTOR	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	NOSECTOR	NOSECTOR	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	NOSECTOR
TAMAÑO	No aplicable	MICRO	MICRO	No aplicable	No aplicable	MICRO	PEQUEÑA	MEDIANA	GRANDE

Figura 32. Perfiles prototípicos del resultado final de la segmentación

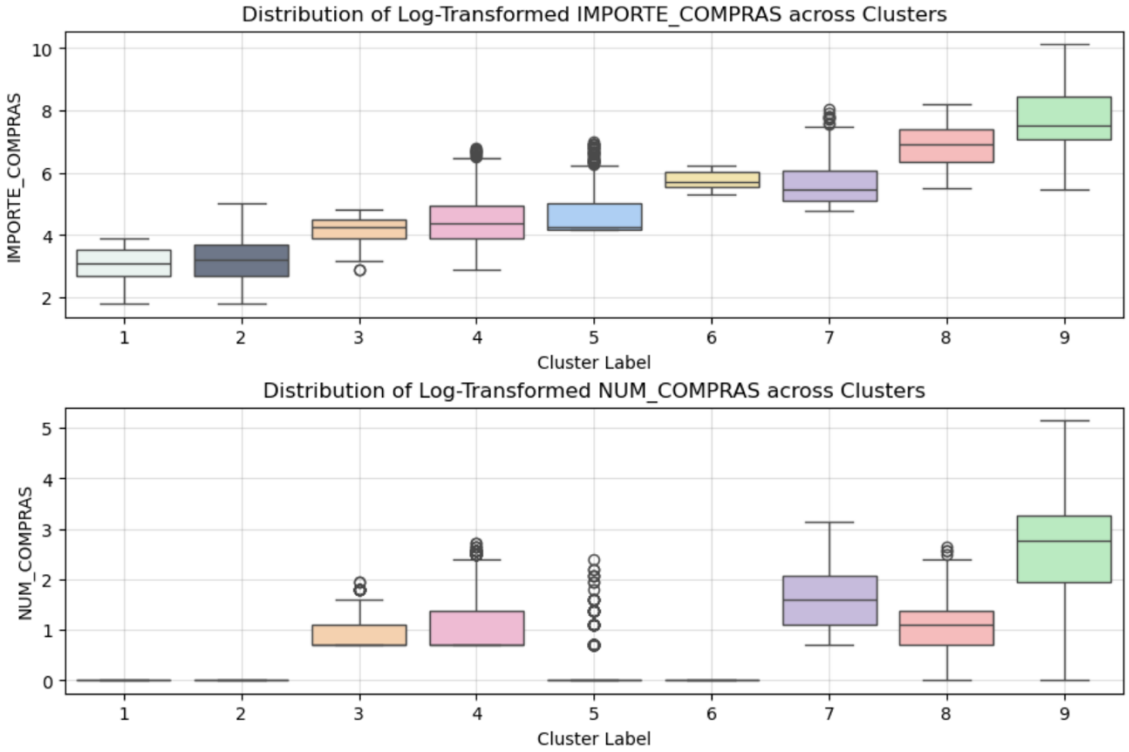


Figura 31. Distribución del importe gastado y del total de compras entre segmentos

Integrados todos los segmentos en única estructura, se pueden extraer las siguientes **conclusiones**:

- En la mayor parte de los perfiles el incremento del monto de compra va unido a un aumento en la cantidad de consultas. Este fenómeno también se puede observar en las segmentaciones diferenciadas de sociedades y empresarios por una parte, y de personas físicas por el otro, a través de

otras métricas de interacción con el servicio que hemos sintetizado mediante la variable «ENGAGEMENT». A pesar de ello, existe una tipología de clientes coincidente con el clúster 7 que se caracteriza por un importante gasto y un reducido número de consumos.

- En cambio, el aumento del monto no siempre viene de la mano de un aumento en la recurrencia de compra, y encontramos segmentos de clientes con un elevado número de compras y un importe total de compras más contenido, y a la inversa.
- Se confirma la regla 80/20 que apuntaba a que la mayoría de la clientela lleva a cabo pequeñas compras de bajo importe, y a que el mayor valor económico proviene de un grupo reducido de usuarios con un nivel de compras y gasto muy por encima de la media.
- Como una posible línea de investigación futura, el segmento que reúne los valores extremos podría ser combinado con el segundo clúster con mayor importe de compras (segmento 8) para una tercera segmentación. Si bien las cifras del segmento 8 están alejadas de las que caracterizan al segmento 9, sus métricas de importe gastado o de total de consultas realizadas se encuentran también a gran distancia de los segmentos inferiores. Esta combinación de segmentos con grandes consumos abre así la puerta a un estudio particularizado centrado en los clientes de más alto valor económico para el negocio.

Es posible acceder al detalle de todas las fases de la segmentación a través de los anexos de esta memoria.

4.8. Predicción de transferencia entre clústeres

Una vez llevada a cabo la segmentación de los clientes, debemos entender que el etiquetado responde a una foto fija del momento en que se obtuvieron los datos. Sin embargo, el **comportamiento de los clientes es dinámico** y conviene que los comerciales responsables de la fidelización del cliente permanezcan atentos a las oportunidades de transferencia entre clústeres para:

a) Dirigir sus campañas de *up-selling* a la clientela susceptible de desplazarse a un segmento de mayor importe medio de compra.

b) Dirigir campañas de cross-selling a los clientes que podrían moverse hacia segmentos de mayor recurrencia de compra.

c) Anticiparse a los *downgrades* o transferencias hacia segmentos de monto y recurrencia inferior lanzando campañas de retención y fidelización del cliente.

Con objeto de predecir estas transferencias se desplegaron 3 estrategias diferentes, aplicadas separadamente sobre el grupo de sociedades y empresarios individuales por un lado, y sobre el de personas físicas por el otro.

4.8.1. Predicción basada en distancia y similitud

Para este primer enfoque de predicción de la transferencia entre clústeres, se utilizó una combinación entre la distancia euclidiana $\left(d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}\right)$ y la similitud del coseno $\left(\text{cosine\_similarity}(x,y) = \frac{x \cdot y}{|x||y|}\right)$. Para el cálculo de la distancia, se emplean únicamente las variables escogidas durante la segunda fase de la segmentación y que son responsables de los clústeres detectados: «IMPORTE_COMPRAS» y «NUM_COMPRAS». Por su parte, para el cálculo de la similitud, se amplía el número de variables relevantes, para lo cual se recuperan los atributos creados durante la fase de *feature engineering* junto a las variables presentes en el *dataset* original referidas a ventas y consumos, y se aplica un modelo de regresión lineal HuberRegressor, caracterizado por su robustez, para seleccionar únicamente aquellas características con un mayor impacto sobre la variable dependiente «IMPORTE_COMPRAS».

Finalmente, se combinan ambas métricas para determinar la posibilidad de transferencia de clientes, **identificando aquellos clientes que cumplían dos criterios**:

1. Aquellos que estaban más próximos al *medoide* de otro clúster que al *medoide* de su propio clúster. La elección como punto de referencia del *medoide*, o punto dentro de un clúster con la menor distancia total a todos los demás puntos del clúster, vino determinada por la falta de normalidad en la distribución de los datos, siendo el *medoide* una mejor representación central al resultar menos sensible a los valores atípicos que la media.
2. Aquellos que presentaban una similitud con el perfil promedio de ese otro clúster superior a 0,8; umbral éste que indica un caso de alta similitud. A través de la similitud del coseno tratamos de entender cómo de similar es un cliente al perfil promedio de un clúster diferente del propio.

El objetivo de la combinación de ambas técnicas fue el de capturar, por una parte, la proximidad absoluta de los clientes, y por otro, la similitud en la dirección de sus características, la cual puede detectar clientes con patrones de comportamiento similares, a pesar de que las magnitudes de los atributos de uno y otro sean diferentes.

En ambos casos se aplicó un estandarizado robusto con objeto de mitigar la influencia sobre el cálculo de las distancias por parte de los valores extremos que habían sido conservados.

4.8.2. Predicción basada en clasificación multinomial

El segundo de los enfoques de predicción de transferencia entre clústeres se basó en el entrenamiento de un modelo predictivo multinomial capaz de clasificar los clientes con sus etiquetas. Un modelo bien ajustado, de acuerdo con las características del cliente, lo encuadrará dentro de uno de los segmentos y, en un porcentaje de casos, la etiqueta predicha no se corresponderá con la real. Si bien esta incongruencia entre predicción y realidad debería valorarse como un error del modelo, en este caso puede

apuntar a que los atributos y hábitos del cliente lo convierten un posible candidato a una transferencia de segmento.

Para la predicción se utilizó un **perceptrón multicapa** (en adelante, «MLP» por las siglas en inglés de la expresión MultiLayer Perceptron), aprovechando sus capacidades de poner en relación un grupo amplio de características utilizando el total de variables del *dataset*. Sin embargo, para mitigar el riesgo de añadir ruido al conjunto de datos que pudiesen desembocar en un sobreajuste del modelo, se implementaron diferentes técnicas como la validación cruzada, la detención temprana y la monitorización del desempeño del modelo. Por otra parte, y para contrarrestar el desbalanceo entre clases, se empleó tanto *over-sampling* como una versión estratificada de la validación cruzada.

4.8.3. Predicción basada en regresión lineal y reagrupamiento

Para el tercer enfoque la estrategia se basó en **dos pasos** consecutivos:

1. Primeramente, se entrenó un modelo **de regresión lineal para la predicción del importe de compras** a partir de las características del cliente. Al igual que en la estrategia de predicción multinomial, la predicción obtenida reflejará el potencial estimado por el modelo, el cual puede estar relativamente alejado del importe real gastado por el cliente en la práctica.
2. Seguidamente, **se aplicó nuevamente el algoritmo de agrupamiento** basándose en el nuevo importe de compras predicho, así como en el número de compras que utilizamos durante la segunda fase del proceso de segmentación. A la vista del potencial de gasto del cliente, el algoritmo lo reclasificó en un segmento potencial, que podía no coincidir con la etiqueta asignada durante la segmentación original.

Respecto a la selección de características empleadas, se utilizaron los atributos estadísticamente significativos identificados a través del modelo robusto que se utilizó para el cálculo de la similitud seguido durante la primera estrategia de predicción de la transferencia, estandarizados y transformados mediante PCA, así como el algoritmo XGBRegressor $\left(\mathrm{Obj}(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k)\right)$.

Además, durante la construcción y evaluación del modelo se utilizaron las siguientes técnicas:

- Para la afinación de los hiperparámetros se utilizó la librería Optuna.
- Con objeto de evitar el sobreajuste, se aplicó nuevamente la detención temprana, la regularización de parámetros mediante L1 y L2, validación cruzada y diferentes intentos; controlándose la fracción de los datos de entrenamiento y de características empleados mediante los parámetros *subsample* y *colsample_bytree* al introducir aleatoriedad en el entrenamiento.

En cuanto a la clasificación, se utilizó el mismo método (algoritmo aglomerativo basado en la afinidad del coseno) y las mismas variables (importe de compras y número de compras) utilizadas durante la segunda fase de la segmentación original.

4.8.4. Combinación de múltiples métricas

Si bien las tres estrategias predijeron las mismas posibilidades de transferencia en todos los casos, la estimación fue más o menos generosa según la métrica utilizada. El negocio dispone así de la posibilidad de escoger cualquiera de los enfoques, optando por una predicción más conservadora u optimista de acuerdo con la política comercial de la empresa, pudiendo así mismo ampliar de manera gradual la selección de acuerdo con los resultados obtenidos en la práctica.

Sin embargo, se procedió a diseñar una solución de consenso que determinase qué candidatos a la transferencia habían sido detectados mediante las tres vías o, alternativamente, por una mayoría de ellas (al menos dos de las tres totales). Con este objeto, se desarrolló una función responsable de **filtrar los resultados consensuados por unanimidad o mayoría** de acuerdo con la configuración del parámetro de entrada añadido a tal efecto, obteniéndose las siguientes **estimaciones** de trasvase entre segmentos por mayoría de métricas para sociedades y empresarios individuales:

- Segmentos 1 y 4: Se trata de un caso muy concreto de clientes con un número muy limitado de compras, pero con un importe de compras muy por encima de la media de su segmento, y que con las técnicas adecuadas de *up-selling* podrían encajar en un segmento con importe de compra promedio más generoso, como es el segmento 4.
- Segmentos 2 y 3: Lo que supondría ascender a un clúster con un importe de compra promedio más elevado y, sobre todo, con el número más elevado de compras de entre todos los clústeres, lo que abre la puerta al traslado de esta selección de clientes a los equipos comerciales tanto de *up-selling* como de *cross-selling*. Cabe también mencionar que hay una porción más moderada de usuarios encuadrados en el extremo inferior del segmento 3 que podría descender al 2 o, por el contrario, consolidar su posición a través de las políticas de fidelización adecuadas.
- Segmentos 3 y 5: Siendo así transferidos hasta el clúster de mayor importe. Se trata de clientes que podrían beneficiarse tanto de las políticas de *cross-selling* dirigidas al segmento 3 (caracterizado por su elevado número de compras), como de las estrategias de *up-selling* dirigidas al segmento 5 (caracterizado por su elevado importe de compra).
- Segmentos 5 y 4: Al contrario de lo que sucedía en la posibilidad de traspaso del segmento 1 al 4, en este caso se trata de clientes que, pese a estar alineados con un segmento de elevado importe de compra en su región más baja, su reducido número de compras lo aproxima a un clúster de compra única y con un importe de compra promedio más moderado, por lo cual el equipo comercial debería dirigir sus esfuerzos tanto a fidelizarlo, como a aumentar el importe del tique mediante técnicas de *up-selling*.

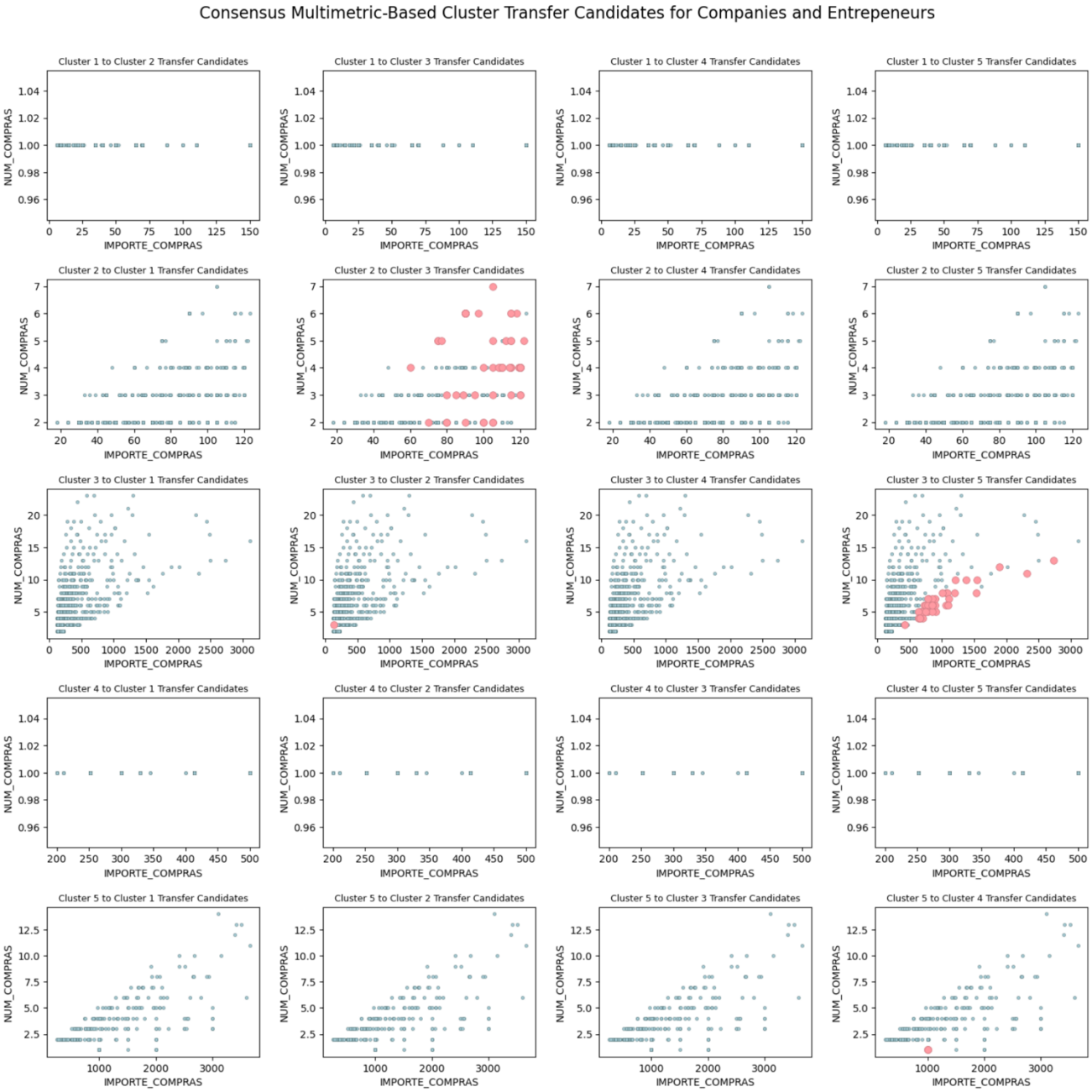


Figura 33. Predicción de transferencia entre segmentos mediante consenso de métricas para sociedades y empresarios

Por su parte, su obtuvo la siguiente estimación de traspaso entre segmentos por mayoría de métricas para las personas físicas:

- Segmentos 2 y 3: Donde un pequeño número de clientes se encuentra en la zona limítrofe entre clientes de gran importe de compra y aquellos con un elevado número de compras, y que podrían beneficiarse tanto de técnicas de *cross-selling* enfocadas al segmento 2 como de las de *up-selling* pensadas para el grupo 3.

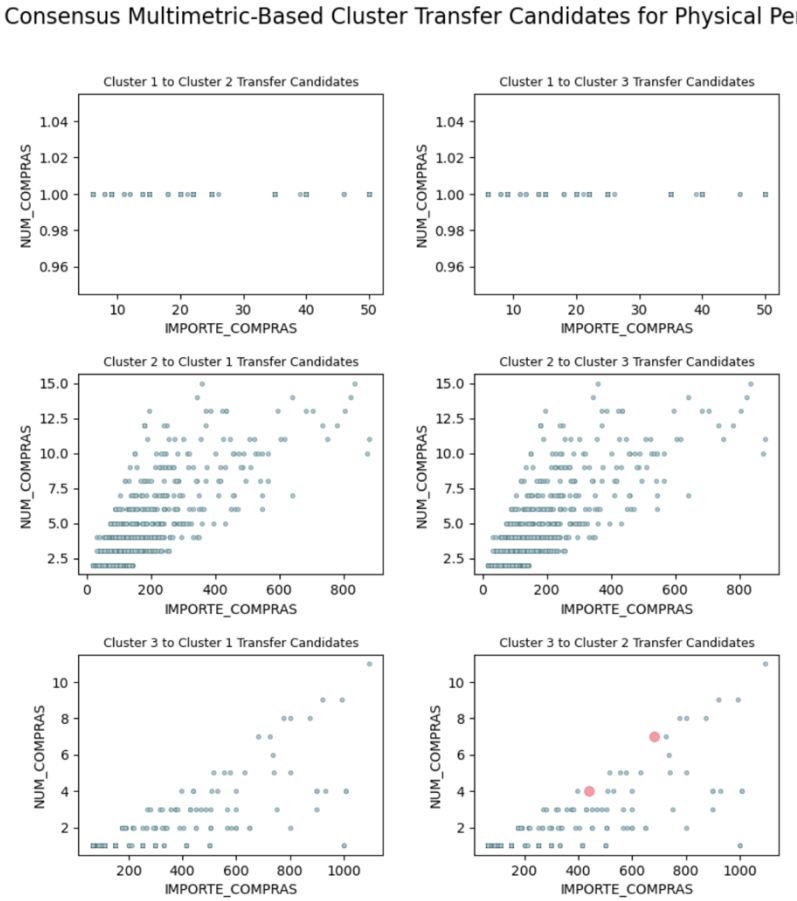


Figura 34. Predicción de transferencia entre segmentos mediante consenso de métricas para personas físicas

En el apartado de anexos de esta memoria es posible consultar el detalle completo de las predicciones anteriores.

Finalmente el resultado se exportó a un fichero de texto con el nombre «CLIENTES_TRANSFERIBLES.txt.» que, además de contar con la columna de la etiqueta, incorporó un nuevo atributo llamado «TRANSFERENCIA» con la predicción de transferencia de segmento en su caso.

4.9. Estudio y desarrollo de estrategias de fidelización

Una vez completada la segmentación de clientes y estimada la posibilidad de transferencia entre clústeres, se puso el foco en la recurrencia de compra del cliente, distinguiendo los compradores habituales de los puntuales, y diseñando estrategias basadas en datos que permitan retenerlos y ofrecerles un servicio más amplio del que obtienen.

4.9.1. Predicción de la recurrencia de compra

Partiendo de las primeras interacciones del cliente se diseñó un modelo para **predecir qué clientes llevarán a cabo compras recurrentes** y cuáles no, información a partir de la cual el equipo comercial respectivo puede anticiparse con objeto de fidelizarlo.

La primera fase del proceso fue preparatoria, **etiquetando a la clientela** como recurrente o no, entendiendo por cliente no recurrente aquel que no compra más de dos productos o que, en caso de comprar más, los adquiere en una única fecha de compra. Para ello se combinaron diferentes fuentes de datos, recuperando la información de ventas.

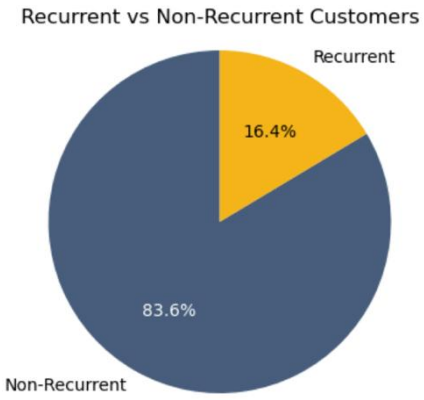


Figura 35. Proporción de clientes con compras recurrentes

Respecto a la información del cliente, se aisló el importe de los dos primeros productos comprados por cada cliente o, en caso de que hubiesen adquirido más de dos artículos en su primera fecha de compra, el importe total de éstos, así como el total de artículos adquiridos. De esta manera se dispuso de la información de importe y número de adquisiciones iniciales de cada cliente a partir de la cual predecir su condición o no de comprador recurrente.

A continuación, se construyó un modelo clasificatorio mediante XGBClassifier $\left(y = \sum_{m=1}^{M} \gamma_m \cdot h_m(x)\right)$, cuya hiperparametrización fue afinada nuevamente a través de Optuna. Dado que la proporción de compradores recurrentes y no recurrentes en los datos no estaba equilibrada, se recurrió a la validación cruzada estratificada y al *over-sampling*. Además de estandarizar los datos, éstos fueron recudidos a través de PCA. Por su parte, para evitar su sobreajuste, se incorporó un mecanismo de detención temprana.

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.96	0.96	1582
1	0.82	0.83	0.82	310
accuracy			0.94	1892
macro avg	0.89	0.89	0.89	1892
weighted avg	0.94	0.94	0.94	1892

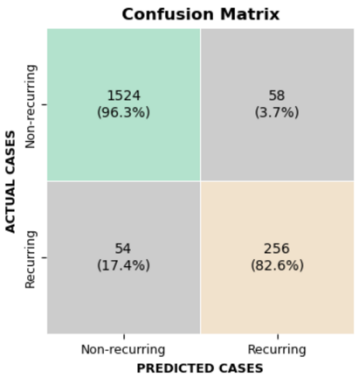


Figura 36. Evaluación del modelo de predicción de recurrencia de compra

La exactitud global del modelo indicó una capacidad de **clasificación correcta entre el 91 % y el 92 % de los casos**, si bien los resultados obtenidos mostraron que su capacidad de identificar a los clientes no recurrentes fue considerablemente superior a la de clasificar a los recurrentes. A pesar de ello, al ser la **prioridad del negocio la de identificar a los no recurrentes** para desplegar sobre ellos las políticas de fidelización correspondientes, las capacidades mostradas evidenciaron su utilidad para la empresa.

4.9.2. Segmentación por niveles de recurrencia y predicción de transferencia

El siguiente paso consistió en segmentar la clientela de acuerdo con su recurrencia de compra para distinguir diferentes niveles de fidelidad, así como qué fracción de los usuarios se podría beneficiar de las políticas de fidelización y venta cruzada de un segmento contiguo, para lo cual se analizaron nuevamente las posibilidades de transferencia entre clústeres. Si bien esta misma estrategia podría haberse llevado a cabo a partir de los datos de la primera segmentación, en los que también se pusieron de relieve las diferentes cantidades de compra correspondientes a uno y otro clúster, en este caso se optó por concentrarse únicamente en el número de compras con objeto de ofrecer una perspectiva diferente en el diseño de estrategias de *cross-selling*.

La segmentación se llevó a cabo sobre la totalidad de los clientes, descartando una moderada proporción de valore extremos y **seleccionando como única variable el número de compras**. En este caso se optó por utilizar *k-medoids* en combinación con la métrica de Manhattan debido a la distorsión en las distancias que podrían originar los valores extremos conservados.

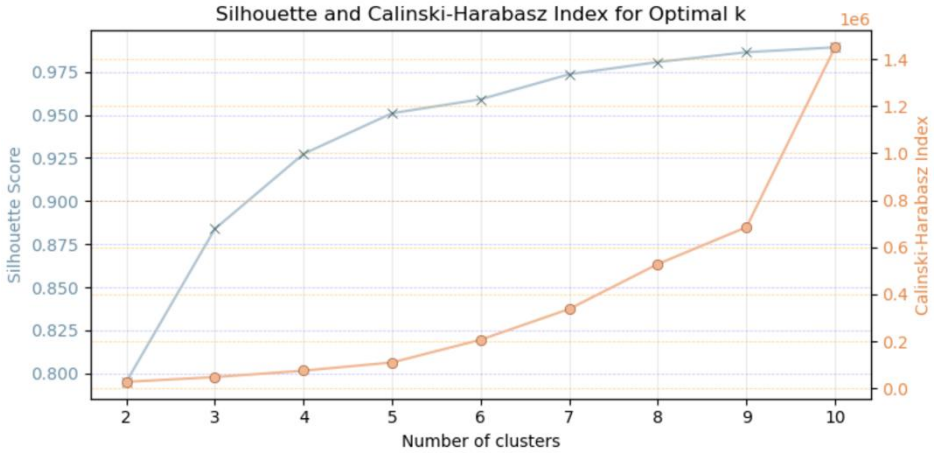
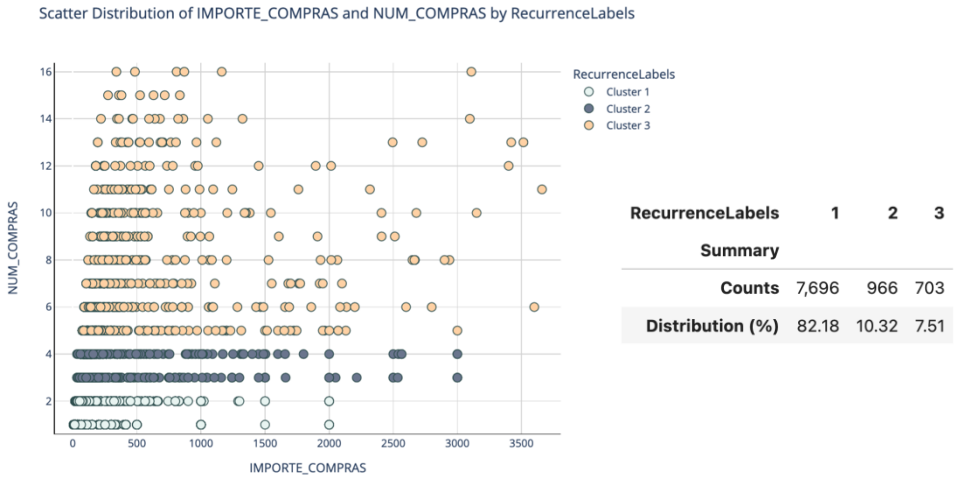


Figura 37. Determinación del valor de *k* para la segmentación por niveles de recurrencia

Si bien, tanto el coeficiente de silueta como el índice de Calinski-Harabasz, apuntaban a grupos más compactos cuanto mayor fuese el nivel de granularidad, se optó por cantidad contenida de segmentos a título ejemplificativo de la estrategia que el negocio podría llevar a la práctica y adaptarla a sus necesidades y recursos, resultando así un total de tres segmentos.



RecurrenceLabels	1	2	3
Summary			
Counts	7,696	966	703
Distribution (%)	82.18	10.32	7.51

Figura 38. Resultado de la segmentación por niveles de recurrencia

A partir de la agrupación establecida, fueron aplicadas nuevamente las **estrategias de predicción de transferencia** expuestas a partir de la segmentación original, comenzando por la predicción basada en distancia y similitud, prosiguiendo con la predicción basada en clasificación multinomial a través de un perceptrón multicapa y, finalmente aplicando la predicción basada en regresión lineal del número de compras y reagrupamiento a partir de las estimaciones de recurrencia obtenidas. Además, se proyectó así mismo una predicción por consenso mayoritario de métricas que arrojó unos resultados conservadores que descartaron la opción de un ascenso de segmento y alertaban del riesgo de que una porción de clientes pertenecientes al clúster superior, pero circunscritos a sus regiones inferiores de compra, descendiesen

hasta el segundo grupo. Una vez más, es posible consultar tanto el detalle de la segmentación como el de las predicciones en el apartado de anexos.

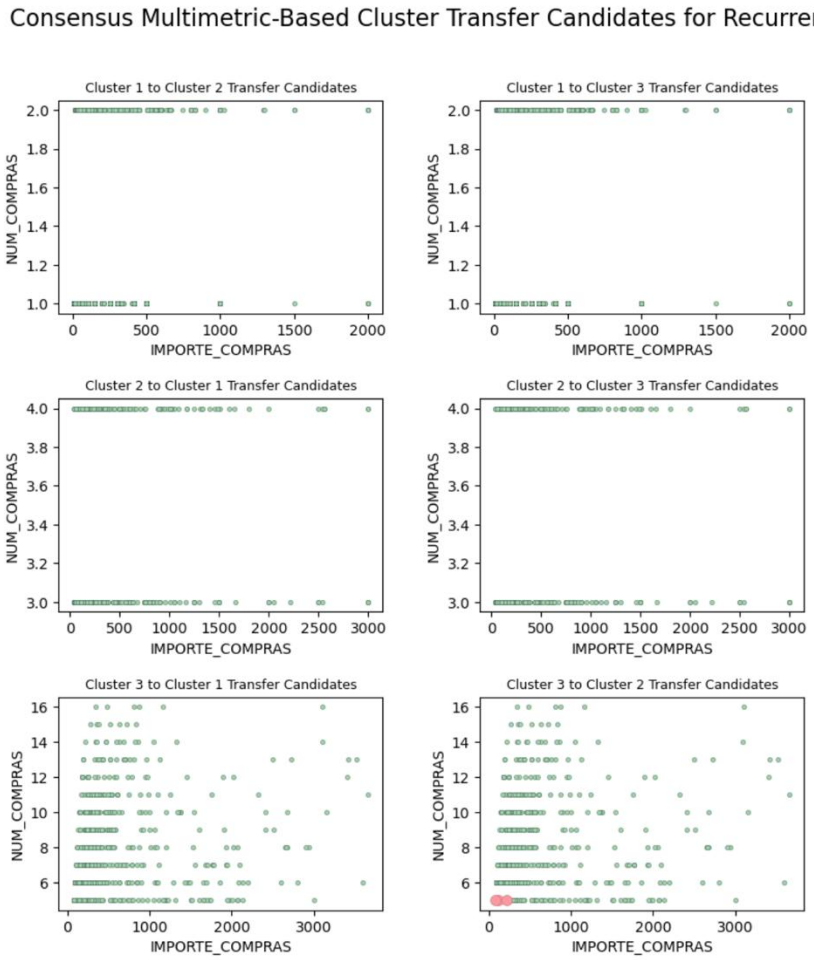


Figura 39. Predicción de transferencia entre segmentos de recurrencia mediante consenso de métricas

El resultado fue exportado en formato de fichero de texto bajo el nombre «RECURRENCIA_CLIENTES.txt», el cual contenía las columnas adicionales «RECURRENCIA», «SEGMENTO» y «TRANSFERENCIA», con la condición o no de recurrente, el segmento asignado de recurrencia y la posibilidad de transferencia de segmento en su caso para cada cliente.

5. Resultados

El Trabajo descrito en esta memoria aborda un proceso de **segmentación de clientes** estructurado en tres etapas:

1. Comenzando por la **limpieza y preprocesamiento** de los datos, etapa durante la cual el objetivo principal ha sido preparar la información para una correcta aplicación de los algoritmos durante las fases posteriores, intentando conservar las particularidades de la base de usuarios presentes en el fichero original.
2. Una **primera fase de segmentación**, durante la cual se han impuesto las características descriptivas del **perfil empresarial** de cada cliente y, en menor medida, de su proceso de *onboarding*. Debido a las limitaciones impuestas por la distribución de los datos, ha sido necesario optar por diferentes técnicas para suavizar los valores extremos y estandarizar adecuadamente las magnitudes, así como utilizar métricas diferentes a la media aritmética como medida representativa de las agrupaciones.
3. Durante la **segunda fase de segmentación**, el enfoque se ha centrado en los dos grandes segmentos que la primera etapa de la segmentación había revelado: uno conformado por diferentes tipos de sociedades y empresarios, y otro por personas físicas. Para ello, hemos alternado las variables que rigen el proceso por aquellas que definen el **aumento del tique**: el número de compras y el importe total gastado. Optando por un algoritmo jerárquico y la afinidad del coseno durante la clasificación, se ha querido incidir en aquellos paralelismos presentes en la clientela que van más allá de las magnitudes y que reflejan una afinidad en su manera de proceder como clientes.

Una vez precisados los segmentos con sus características particulares, entre las cuales el importe de compra y el número de adquisiciones tienen un papel destacado, la sección de ventas de la empresa puede dirigir sus esfuerzos de venta cruzada u ofrecer productos que pueden aportar más valor para el cliente a cambio de un precio más alto para el negocio. Sin embargo, algunos clientes podrían beneficiarse de las políticas de un segmento contiguo, razón por la cual el paso siguiente ha consistido en predecir cuáles de esos consumidores podrían **ser transferidos** hacia clústeres adyacentes, consolidando así su posición de cliente de alto valor para la empresa. Para ello, se ha experimentado con **diferentes métricas basadas en distancia y similitud, predicción multinomial mediante redes neuronales multicapa y regresión lineal**, analizando sus diferencias y puntos comunes, así como combinándolas para afinar las estimaciones obtenidas.

A continuación, como objetivo secundario a partir de los pasos seguidos durante la primera parte del cuaderno, hemos analizado la **fidelización** de la clientela, distinguiendo aquellos usuarios que compran al negocio de manera repetida de los que no pasan de una interacción puntual. A través de un modelo de predicción binomial, los clientes pueden ser clasificados tras unas pocas interacciones comerciales con el objeto de que el equipo comercial correspondiente despliegue las estrategias de

retención oportunas. Además, los clientes han sido nuevamente particionados en función de su número de interacciones comerciales con el objetivo de detectar una vez más cuáles podrían sacar mayor provecho de las ofertas de venta cruzada, recibiendo así un mayor valor.

Los datos obtenidos a la conclusión de estas operaciones han sido exportados en formato de texto, permitiendo así su integración en otros procesos de inteligencia de negocios.

Por último, el cuaderno Jupyter donde se recogen todas estas operaciones junto a las funciones encargadas de la operativa principal (ejecución de algoritmos, transformaciones, escalado, obtención de estadísticas de los clústeres y graficado) admiten ser parametrizadas con diferentes métricas y selección de variables con el objeto de convertirse en una **herramienta reutilizable**, facilitando su posterior adaptación a otros supuestos y otros juegos de datos con sus características particulares. El producto está disponible en el siguiente repositorio público: https://github.com/Marcos-A/TFM-MCD; por razones de privacidad, no se incluyen en el mismo los datos de entrada y salida.

6. Conclusiones y trabajos futuros

6.1. Conclusiones

A la finalización del trabajo, conviene retomar tanto el **objetivo principal** de segmentación de la clientela, así como los **secundarios** planteados en sus etapas iniciales entre los cuales destaca el diseño de estrategias de optimización de su fidelización, todos los cuales han sido alcanzados. Para conseguirlo ha sido sin embargo necesario hacer frente a tres **dificultades** principales:

1. La distribución sesgada de los datos, y la necesidad de respetarla y trabajar con ella ha condicionado la elección de los algoritmos y las métricas empleadas a lo largo de todo el proceso, así como la manera en que los valores atípicos descartados han sido reintegrados en la segmentación al final del proceso.
2. La comprensión del negocio y de las particularidades de los clientes ha resultado esencial para interpretar la información e idear la mejor manera de abordar el problema. Así, por ejemplo, la disparidad en el comportamiento de las personas físicas, la existencia de usuarios con un reducido nivel de recurrencia, pero un elevado nivel de gasto; o la elevada contribución económica concentrada en unos pocos clientes, han sido algunas de las circunstancias inesperadas que ha sido necesario descubrir y comprender para construir una solución adaptada a las particularidades del caso.
3. Finalmente, desde un punto de vista estrictamente organizativo, el principal reto ha sido la limitación de tiempo, la cual únicamente ha sido posible vencer gracias al seguimiento estricto de la planificación planteada desde el inicio.

Durante todo el proceso se han puesto en práctica modelos de aprendizaje computacional **supervisado y no supervisado**, así como diferentes algoritmos, jerárquicos y no jerárquicos. Se han empleado diferentes **técnicas** para evitar el sobreajuste de los modelos construidos, así como para afinar sus hiperparámetros. También se han diseñado nuevas variables combinando diversas fuentes de datos con objeto de arrojar luz sobre nuevos patrones presentes en la información o de dotarla de significado. Además, se han puesto en práctica diferentes técnicas de **representación de la información**, ya sea de manera tabular o mediante su representación gráfica. Con todo ello se ha pretendido poner en práctica la diversidad de conocimientos adquiridos a lo largo de las diferentes asignaturas del Máster.

En cuanto a los **impactos ético-sociales, de sostenibilidad y de diversidad**, si bien la solución construida permite ofrecer un producto adaptado a las necesidades del usuario en pos de un consumo responsable (objetivo 12) y de la solidificación de cadenas de suministro más sostenibles (objetivo 8), la variedad de algoritmos utilizados incluye tipologías computacionalmente costosas, como es el uso de algoritmos jerárquicos y de redes neuronales. Por otra parte, el cálculo de los *medoides* que

implica el trabajo con *k-medoids* supone calcular la distancia entre cada par de puntos dentro de un clúster, siendo ésta una operación menos directa y eficiente que la alternativa del cálculo de centroides en algoritmos como *k-means*. Si bien las dimensiones del juego de datos son relativamente modestas, la metodología escogida no es siempre la de menor coste computacional, pero se ha estimado que era la que el contexto requería para brindar la solución más apropiada al problema planteado.

6.2. Principales aportaciones del trabajo

Las principales novedades aportadas por este trabajo se concentran en tres áreas:

1. La segmentación estructurada en **2 fases** con estrategias diferenciadas:
   i. Durante la 1ª fase se ha dejado **en manos del algoritmo la selección de las características** diferenciadoras más relevantes, recayendo éstas en el **perfil empresarial** del cliente.
   ii. Durante la 2ª fase se ha llevado a cabo una **selección deliberada** y reducida de variables centradas en las ventas, optándose por una **métrica del coseno** que agrupase a los clientes por patrones de comportamiento en lugar de por magnitudes.
2. Los **valores extremos** más alejados de la distribución han sido concentrados en un segmento VIP, y se ha optado por conservar todos los demás para el proceso de agrupamiento, comprometiendo de esta manera la simetría y normalidad de la distribución en favor de una segmentación inclusiva que da cabida a usuarios de comportamiento similares a pesar de la diferencia en las magnitudes.
3. Para la predicción de transferencia entre clústeres, se ha desarrollado una estrategia mixta que **combina distancia y similitud**, y que complementa el enfoque tradicional de la clasificación multinomial y la regresión lineal del importe de compra. Además, con objeto de consensuar los resultados de las diferencias entre las predicciones se ha incluido un mecanismo de decisión por consenso.

6.3. Futuras líneas de investigación

El estudio desarrollado a lo largo de este Trabajo abre las puertas a nuevas vías que, por cuestiones de tiempo y espacio, no han podido desarrollarse en el mismo. Entre éstas, destacamos las siguientes:

- Una **segmentación basada exclusivamente en los consumos** del servicio contrastada con la tipología de producto adquirido por el cliente podría evidenciar:
  - Qué usuarios podrían optimizar su importe invertido en el servicio con la compra de bonos en lugar de adquirir productos puntuales.
  - Qué usuarios se podrían beneficiar de adquirir un plan de suscripción.

- Qué usuarios no están rentabilizando su suscripción o la compra de un bono y podrían beneficiarse de un *downgrade* en el producto adquirido.

- Como ha sido expuesto, los clientes que son **personas físicas** engloban una tipología heterogénea dentro de la cual se encuentran casos con un nivel de gasto y consumos muy por encima de lo que se espera de un particular. Parece lógico pensar que detrás de estos clientes se encuentren sociedades o empresarios que no desean dar a conocer su condición de tal.

  Con objeto de que estos clientes reciban un tratamiento comercial adecuado a sus necesidades, una posible táctica de segmentación consistiría en distinguirlos de acuerdo con su nivel de recurrencia, de manera que las personas físicas no recurrentes formasen su propio segmento, y el resto fuese subsumido dentro en los segmentos reservados para sociedades y empresarios. Para ello, este grupo de personas físicas podría ser discriminado a la finalización de la primera fase de la segmentación, y se integrado junto con las sociedades y empresarios individuales, para estimar su clusterización más apropiada en conjunto. Una segunda alternativa consistiría en adjudicarlos individualmente a los segmentos de sociedades y empresarios ya constituidos, bien en función de su nivel de gasto, bien asignándolos a aquel grupo cuyo *medoide* resulte más próximo en cada caso, bien mediante un modelo de clasificación multinomial.

- El segundo de los segmentos con unas características heterogéneas es el formado por la clientela VIP que ocupa los **valores extremos**. Teniendo en cuenta la relevante contribución económica de estos usuarios para el negocio, más allá de la posibilidad de dotarlos de un tratamiento comercial personalizado sería conveniente llevar a cabo una segmentación de éstos para estudiar posibles estrategias de aumento del tique y de fidelización. Además, como ya ha sido expuesto durante las conclusiones de la segmentación, el conjunto de partida podría enriquecerse con el segmento inmediatamente inferior, con el cual guarda algunos paralelismos. Teniendo en cuenta la amplitud de los rangos en que se mueven estos usuarios, sería necesario llevar a cabo transformaciones de los datos (por ejemplo, logarítmica o de raíz cuadrada) y trabajar con un nivel de granularidad contenido.

7. Glosario

A continuación, se recogen algunos de los términos técnicos y acrónimos más relevantes mencionados en esta memoria. Dado que este trabajo aúna técnicas de mercadotecnia y de ciencia de datos, todas las definiciones deben entenderse contextualizadas dentro de estas disciplinas.

- **Algoritmo aglomerativo**: Método de agrupamiento jerárquico que construye un árbol de clústeres o dendrograma a través de un proceso ascendente, siendo el punto de partida la creación de tantos clústeres como puntos de datos, los cuales se van fusionando progresivamente con otros clústeres más similares hasta que todos los puntos quedan en un único clúster o hasta que se alcanza un número predeterminado de grupos. Para medir la similitud pueden emplearse diferentes distancias, como por ejemplo la mínima (o enlace simple), máxima (o enlace completo) o la promediada (o enlace promedio).
- **Algoritmo del descenso del gradiente**: Método iterativo de optimización que se usa para minimizar una función objetivo, típicamente una función de error o de coste en el contexto de aprendizaje automático. El algoritmo ajusta los parámetros del modelo en la dirección opuesta al gradiente de la función objetivo con respecto a esos parámetros, siendo la magnitud de estos ajustes controlada por el hiperparámetro de la tasa de aprendizaje. El objetivo es encontrar los valores de los parámetros que minimizan la función objetivo al mejorar así el rendimiento del modelo.
- **Análisis de componentes principales (o PCA)**: Técnica de reducción de dimensionalidad que transforma las variables originales de un conjunto de datos en un conjunto alternativo de atributos no correlacionados, llamados componentes principales, reduciendo la dimensionalidad de los datos mientras se conserva la mayor parte de su varianza.
- ***Autoencoder***: Tipo de red neuronal utilizada para aprender representaciones codificadas y más eficientes de los datos, bien para reducir su dimensionalidad o para generar muestras nuevas.
- ***Average Order Value* (o AOV)**: Valor promedio de los pedidos realizados por un cliente durante un periodo de tiempo determinado. Su cálculo se obtiene de la división del total de ingresos por el número de pedidos.
- ***Batch Processing Gradient Descent* (o BPGD)**: Variante del algoritmo de descenso de gradiente que actualiza los parámetros del modelo usando todo el conjunto de datos en cada iteración, en lugar de hacerlo después de cada muestra individual o lote de muestras.
- ***Business to Business* (o B2B)**: Modelo de negocio en el que las transacciones de bienes o servicios ocurren entre dos empresas.

- ***Business to Government* (o B2G)**: Modelo de negocio en el que las empresas venden productos o servicios a gobiernos o entidades gubernamentales.
- **Centroide**: Punto central de un clúster que se obtiene calculando el promedio de todas las coordenadas de sus puntos.
- ***Churn Rate***: Tasa de deserción de clientes durante un periodo determinado y métrica clave para evaluar su retención.
- ***Clustering Feature Importance* (o CFI)**: Medida de la relevancia de las variables que participan en un proceso de agrupamiento de datos, y que permite la identificación de los atributos más influyentes en la formación de los clústeres.
- **Clúster**: Grupo de objetos más similares entre sí que respecto a los objetos pertenecientes a otros grupos. El objetivo de la agrupación en clústeres es el de encontrar estructuras y patrones en amplios y complejos conjuntos de datos que permitan su segmentación y faciliten su análisis.
- ***Consumer to Consumer* (o C2C)**: Modelo de negocio en el que las transacciones de bienes o servicios ocurren entre dos consumidores.
- **Conversión**: Proceso de transformación de un visitante en un cliente que lleva a cabo una acción valiosa por el negocio, generalmente una compra o la contratación de un servicio.
- ***Customer Lifetime Value* (o CLV, *LifeTimeValue* o LTV)**: Estimación del valor económico total que un cliente aportará a una empresa a lo largo de toda su relación comercial con la misma.
- ***Cross-selling***: Técnica de ventas consistente en ofrecer productos o servicios adicionales que complementen la compra original del cliente para aumentar el valor de la transacción.
- ***Customer Relationship Management* (o CRM)**: Estrategia y tecnología para gestionar las relaciones y las interacciones de una empresa con sus clientes o potenciales clientes.
- **DBSCAN**: Siglas de la expresión inglesa *Density-Based Spatial Clustering of Applications with Noise*, se trata de un algoritmo de agrupamiento basado en densidad que identifica clústeres como regiones densamente pobladas de puntos separadas por regiones de baja densidad, lo que le permite encontrar clústeres con formas arbitrarias a pesar del ruido presente en los datos.
- **Dendrograma**: Diagrama de árbol que muestra las relaciones de agrupamiento jerárquico entre un conjunto de objetos y que se utiliza para visualizar el proceso de agrupamiento aglomerativo o divisivo.
- **Distancia de Manhattan**: Métrica que calcula la distancia entre dos puntos en un espacio mediante la suma las diferencias absolutas de sus coordenadas. SE la conoce también como distancia de taxista, *city block* o L1.

- **Distribución gaussiana (o normal)**: Se trata de una distribución de los datos de probabilidad continua, con una curva en forma de campana simétrica definida por su media y desviación estándar.
- ***Early Stopping* (o detención temprana)**: Técnica utilizada durante el entrenamiento de modelos de aprendizaje automático con objeto de detener el proceso una vez que el rendimiento del modelo sobre el conjunto de validación comienza a deteriorarse, evitando de esta forma su sobreajuste.
- ***Engagement***: Grado de interacción de los usuarios con un producto o servicio, y que sirve como indicativo de su fidelidad, aprovechamiento y satisfacción.
- ***Eps***: En el algoritmo DBSCAN, es la distancia máxima entre dos puntos para que uno sea considerado vecino del otro, la cual determina el radio de búsqueda para los puntos vecinos.
- ***Expectation-Maximization* (o EM)**: Algoritmo iterativo que se usa para encontrar estimaciones de máxima verosimilitud de parámetros en modelos estadísticos, especialmente aquellos que dependen de variables latentes, alternando entre los pasos de expectativa (E) y maximización (M) hasta converger.
- **Fidelización**: Conjunto de estrategias y acciones desarrolladas por una empresa para formar y mantener una relación duradera con sus clientes, reforzando su lealtad y satisfacción en el largo plazo.
- **Frecuencia**: Cantidad de veces que un cliente realiza una acción específica durante un determinado período de compra, y referido habitualmente a las compras o a la utilización del servicio.
- ***Frequency-Sensitive Competitive Learning* (o FSCL)**: Variante de red neuronal competitiva que ajusta la frecuencia de actualización de los nodos para evitar la saturación y asegurar que todos participen en el proceso de aprendizaje.
- **Hiperparámetro**: Parámetro cuyo valor controla el comportamiento del algoritmo de entrenamiento.
- ***K-means*:** Algoritmo de agrupamiento que particiona un conjunto de datos en k grupos, donde cada clúster está representado por su centroide, y que minimiza la suma de las distancias cuadráticas entre los puntos y sus respectivos centroides.
- ***K-medoids***: Algoritmo de agrupamiento que divide los datos en k grupos empleando puntos reales del conjunto de datos llamados «medoides» para representar cada uno de los grupos, lo que lo hace más robusto frente la presencia de valores atípicos. El algoritmo minimiza así la suma de las distancias entre cada uno de los puntos del grupo y su *medoide* correspondiente.
- ***Mailing***: Estrategia de mercadotecnia que utiliza el envío de correos electrónicos para la comunicación con los clientes y su retención, así como para la promoción de productos o servicios.

- ***Medoide***: Objeto representativo de un clúster cuyo promedio de distancia a todos los otros puntos dentro del mismo clúster es mínimo, y que se corresponde con un punto real del juego de datos.
- ***Minimum Viable Product* (o MVP)**: Versión más simple y funcional de un producto que permite a un equipo recoger la mayor cantidad de aprendizaje validado sobre los clientes con el menor grado de esfuerzo.
- ***MinPts***: En el algoritmo DBSCAN, se refiere al número mínimo de puntos que deben estar dentro del radio *eps* para que un punto sea considerado un núcleo o *core point*, definiendo así la densidad mínima necesaria para formar un clúster.
- ***MultiLayer Perceptron* (o MLP)**: Tipo de red neuronal artificial formada por múltiples capas de nodos o neuronas, y donde cada capa está totalmente conectada a la siguiente. Habitualmente se usa en el contexto de problemas de clasificación y de regresión.
- **Modelo de asignación de Dirichlet latente**: Modelo generativo de temas para un conjunto de documentos que descubre patrones ocultos o temas en los datos de texto, donde cada documento es una mezcla de temas y cada tema es una mezcla de palabras.
- ***Onboarding***: Proceso de integración y formación de los nuevos usuarios sobre cómo en el uso de un producto o servicio para asegurar el éxito durante su adopción.
- ***Outliers* (o valores extremos o atípicos)**: Datos significativamente alejados de la mayoría de los datos en un conjunto, y que pueden indicar variabilidad extrema, errores en los datos o fenómenos fuera de lo común.
- ***Over-sampling***: Técnica empleada con objeto de equilibrar conjuntos de datos desbalanceados aumentando el total de muestras presentes en las clases minoritarias mediante la creación de copias u otras técnicas.
- **Recencia**: Tiempo transcurrido desde la última interacción o compra de un cliente, siendo ésta una métrica habitual en los procesos de segmentación y análisis del comportamiento de los clientes.
- **Recurrencia**: Frecuencia con la que un cliente compra o interacciona con la empresa durante un cierto periodo de tiempo, y que sirve como indicativo de su grado de fidelización.
- **Retención**: Capacidad de una empresa de conservar a sus clientes a lo largo del tiempo.
- **RFM**: Siglas de la expresión inglesa *Recency, Frequency, Monetary*; se trata de un modelo de segmentación de clientes que se basa en las tres dimensiones de la recencia, la frecuencia y el monto de compra, y que permite identificar los clientes más valiosos para el negocio.

- **Ruido**: Variaciones o errores aleatorios que no aportan información significativa y que puede distorsionar los resultados y hacer más difícil la identificación de patrones relevantes en los datos.
- **Segmentación**: Proceso de dividir una base usuario en grupos más reducidos y homogéneos cuyos miembros comparten características similares, con la finalidad de personalizar las estrategias comerciales y hacerlas más eficientes.
- **Similitud del coseno**: Métrica utilizada para estimar la similitud entre dos vectores en un espacio de características. Su cálculo se obtiene a través del coseno del ángulo entre ambos vectores, dando como resultado un valor entre -1 y 1 (en el que 1 indica que ambos vectores son exactamente iguales, 0 la ausencia de similitud y -1 indica que ambos vectores son exactamente opuestos).
- **Sobreajuste (o *overfitting*)**: Problema habitual en el entrenamiento de modelos de aprendizaje automático, a través del cual el modelo se ajusta en exceso a los datos de entrenamiento adaptándose al ruido presente en los datos, lo que lo que provoca un rendimiento pobre cuando se enfrenta a nuevos datos y que el modelo no sea generalizable.
- ***StratifiedKFold***: Técnica de validación cruzada que divide los datos en k pliegues (o *folds*) que constan aproximadamente de la misma proporción de clases que el conjunto de datos original, gracias a lo cual es posible mantener la distribución de clases en cada pliegue cuando se trabaja con problemas de clasificación, consiguiéndose así una mejor evaluación del modelo.
- ***Up-selling***: Técnica de ventas consistente en persuadir al cliente para que compre una versión más cara del producto o servicio que está considerando, para así incrementar el valor del tique de compra.
- **Validación cruzada**: Técnica para evaluar la capacidad de generalización de un modelo de aprendizaje automático a través de la división del conjunto de datos en múltiples subconjuntos, entrenando el modelo en unos mientras se valida en otros, y posteriormente rotando estos subconjuntos para reforzar su evaluación.

8. Bibliografía

Ahmed, M., Seraj, R., Islam, S.M.S., 2020. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. Electronics 9, 1295. https://doi.org/10.3390/electronics9081295 (consultado el 22/02/2023).

Alelyani, S., Tang, J., Liu, H., 2018. Feature Selection for Clustering: A Review, in: Aggarwal, C.C., Reddy, C.K. (Eds.), Data Clustering. Chapman and Hall/CRC, pp. 29–60. https://doi.org/10.1201/9781315373515-2 (consultado el 01/06/2024).

Alkhayrat, M., Aljnidi, M., Aljoumaa, K., 2020. A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. J Big Data 7, 9. https://doi.org/10.1186/s40537-020-0286-0 (consultado el 23/03/2024).

Anthony, L.F.W., Kanding, B., Selvan, R., 2020. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models.

Aryuni, M., Didik Madyatmadja, E., Miranda, E., 2018. Customer Segmentation in XYZ Bank Using K-Means and K-Medoids Clustering. Presented at the 2018 International Conference on Information Management and Technology (ICIMTech), IEEE, Jakarta, pp. 412–416. https://doi.org/10.1109/ICIMTech.2018.8528086 (consultado el 18/03/2024).

Balakrishnan, P.V. (Sundar), Cooper, M.C., Jacob, V.S., Lewis, P.A., 1996. Comparative performance of the FSCL neural net and K-means algorithm for market segmentation. European Journal of Operational Research 93, 346–357. https://doi.org/10.1016/0377-2217(96)00046-X (consultado el 17/03/2024).

Banco Mundial, 2023. Perspectivas Económicas Mundiales [Documento en línea]. URL https://www.bancomundial.org/es/publication/global-economic-prospects (consultado el 16/03/2024).

Bass, F.M., Pessemier, E.A., Tigert, D.J., 1969. A Taxonomy of Magazine Readership Applied to Problems in Marketing Strategy and Media Selection. J BUS 42, 337–363. https://doi.org/10.1086/295202 (consultado el 17/03/2024).

BBVA Research, 2024. Colombia 2024: de menos a más o a más más [Documento en línea]. URL https://www.bbvaresearch.com/publicaciones/colombia-2024-de-menos-a-mas-o-a-mas-mas/ (consultado el 16/03/2024).

Bosch Rué, A., Casas-Roma, J., Lozano Bagén, T., 2019. Deep learning: principios y fundamentos, Primera edición digital. ed. Editorial UOC, Barcelona.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324 (consultado el 25/03/2024).

Calinski, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Comm. in Stats. - Theory & Methods 3, 1–27. https://doi.org/10.1080/03610927408827101 (consultado el 24/03/2024).

Campello, R.J.G.B., Moulavi, D., Sander, J., 2013. Density-Based Clustering Based on Hierarchical Density Estimates, in: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (Eds.), Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 160–172. https://doi.org/10.1007/978-3-642-37456-2_14 (consultado el 24/03/2024).

Christy, A.J., Umamakeswari, A., Priyatharsini, L., Neyaa, A., 2021. RFM ranking – An effective approach to customer segmentation. Journal of King Saud University - Computer and Information Sciences 33, 1251–1257. https://doi.org/10.1016/j.jksuci.2018.09.004 (consultado el 18/03/2024).

Claycamp, H.J., 2024. A Theory of Market Segmentation. Journal of Marketing Research 5, 388–394.

CORDIS, 1998. EUNITE: European Network on Intelligent Technologies for Smart Adaptive Systems. [Documento en línea]. CORDIS | European Commission. URL https://cordis.europa.eu/project/id/25959/es (consultado el 23/02/2024).

Cowls, J., Tsamados, A., Taddeo, M., Floridi, L., 2023. The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. AI & Soc 38, 283–307. https://doi.org/10.1007/s00146-021-01294-x (consultado el 26/03/2024).

Davies, D.L., Bouldin, D.W., 1979. A Cluster Separation Measure. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1, 224–227. https://doi.org/10.1109/TPAMI.1979.4766909 (consultado el 24/03/2024).

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39, 1–38.

Doñate, À., 2020. La UOC incorpora el compromiso ético y global en sus grados y másteres [Documento en línea]. URL https://www.uoc.edu/es/news/2020/135-competencia-etica-global (consultado el 16/03/2024).

Eilam, T., Bello-Maldonado, P., Bhattacharjee, B., Costa, C., Lee, E.K., Tantawi, A., 2023. Towards a Methodology and Framework for AI Sustainability Metrics, in: Proceedings of the 2nd Workshop on Sustainable Computer Systems. Presented at the HotCarbon '23: 2nd Workshop on Sustainable Computer Systems, ACM, Boston MA USA, pp. 1–7. https://doi.org/10.1145/3604930.3605715 (consultado el 26/03/2024).

Europa Press, 2024. El PIB de Colombia creció un 0,6% en el año 2023 [Documento en línea]. URL https://www.europapress.es/economia/macroeconomia-00338/noticia-pib-colombia-crecio-06-ano-2023-20240215175319.html (consultado el 16/03/2024).

Frawley, W.J., Piatetsky-Shapiro, G., Matheus, C.J., 1992. Knowledge Discovery in Databases: An Overview. AI Magazine 13.

Gamez, M.J., 2015. 17 objetivos para transformar nuestro mundo [Documento en línea]. Desarrollo Sostenible. URL https://www.un.org/sustainabledevelopment/es/ (consultado el 16/03/2024).

Gironés Roig, J., Casas Roma, J., Minguillón Alfonso, J., Caihuelas Quiles, R., 2017. Minería de datos: modelos y algoritmos, Primera edición digital. ed. Editorial UOC, Barcelona.

Heguerte, L.B., Bugeau, A., Lannelongue, L., 2023. How to estimate carbon footprint when training deep learning models? A guide and review. Environ. Res. Commun. 5, 115014. https://doi.org/10.1088/2515-7620/acf81b (consultado el 26/03/2024).

ICEX, 2023. Informe e-País: El comercio electrónico en Colombia Noviembre 2023 [Documento en línea]. URL https://www.icex.es/content/dam/es/icex/oficinas/020/documentos/2024/01/informe-e-pa%C3%ADs-colombia-2023/Colombia%20Resumen%20Ejecutivo%202023.pdf (consultado el 16/03/2024).

Kalaidopoulou, K., Triantafyllou, S., Griva, A. and Pramatari, K., 2017. Identifying customer satisfaction patterns via data mining: The case of greek e-shops.

Kansal, T., Bahuguna, S., Singh, V., Choudhury, T., 2018. Customer Segmentation using K-means Clustering, in: 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS). Presented at the 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), pp. 135–139. https://doi.org/10.1109/CTEMS.2018.8769171 (consultado el 18/03/2024).

Kaufman, L., Rousseeuw, P.J., 1987. Clustering by means of medoids. Proceedings of the statistical data analysis based on the L1 norm conference, neuchatel, switzerland 31.

Kernan, J.B., 1968. Choice Criteria, Decision Behavior, and Personality. Journal of Marketing Research 5, 155–164.

Kim, S.-Y., Jung, T.-S., Suh, E.-H., Hwang, H.-S., 2006. Customer segmentation and strategy development based on customer lifetime value: a case study. Expert Systems with Applications 31, 101–107. https://doi.org/10.1016/j.eswa.2005.09.004 (consultado el 18/03/2024).

Kobren, A., Monath, N., Krishnamurthy, A., McCallum, A., 2017. A hierarchical algorithm for extreme clustering., in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Halifax NS Canada, pp. 255–264. https://doi.org/10.1145/3097983.3098079 (consultado el 23/03/2024).

Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T., 2019. Quantifying the Carbon Emissions of Machine Learning.

Lieder, I., Segal, M., Avidan, E., Cohen, A., Hope, T., 2019. Learning a Faceted Customer Segmentation for Discovering new Business Opportunities at Intel. Presented at the 2019 IEEE International Conference on Big Data (Big Data), IEEE, Los Angeles, CA, USA, pp. 6136–6138. https://doi.org/10.1109/BigData47090.2019.9006589 (consultado el 23/03/2024).

Liu, Y., Hou, T., Miao, Y., Liu, M., Liu, F., 2021. IM-c-means: a new clustering algorithm for clusters with skewed distributions. Pattern Anal Applic 24, 611–623. https://doi.org/10.1007/s10044-020-00932-2 (consultado el 24/03/2024).

Lloyd, S., 1982. Least squares quantization in PCM. IEEE Trans. Inform. Theory 28, 129–137. https://doi.org/10.1109/TIT.1982.1056489 (consultado el 20/03/2024).

Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions. Advances in neural information processing systems 30.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability 1, 281–297.

Mangiameli, P., Chen, S.K., West, D., 1996. A comparison of SOM neural network and hierarchical clustering methods. European Journal of Operational Research 93, 402–417. https://doi.org/10.1016/0377-2217(96)00038-0 (consultado el 23/03/2024).

Mazanec, J.A., 1992. Classifying Tourists into Market Segments [Documento en línea]. URL https://www.tandfonline.com/doi/epdf/10.1300/J073v01n01_04?needAccess=true (consultado el 17/03/24).

McInnes, L., Healy, J., Melville, J., 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Montgomery, D.B., Silk, A.J., 1971. Clusters of Consumer Interests and Opinion Leaders' Spheres of Influence. Journal of Marketing Research 8, 317–321. https://doi.org/10.1177/002224377100800306 (consultado el 17/03/2024).

Moran, M., 2015a. Objetivo 12: Garantizar modalidades de consumo y producción sostenibles. Desarrollo Sostenible. URL https://www.un.org/sustainabledevelopment/es/sustainable-consumption-production/ (consultado el 16/03/2024).

Moran, M., 2015b. Igualdad de género y empoderamiento de la mujer. Desarrollo Sostenible. URL https://www.un.org/sustainabledevelopment/es/gender-equality/ (consultado el 28/05/2024).

Moran, M., 2015c. Objetivo 8: Promover el crecimiento económico inclusivo y sostenible, el empleo y el trabajo decente para todos. Desarrollo Sostenible. URL https://www.un.org/sustainabledevelopment/es/economic-growth/ (consultado el 16/03/2024).

Namvar, M., Gholamian, M.R., KhakAbi, S., 2010. A Two Phase Clustering Method for Intelligent Customer Segmentation, in: 2010 International Conference on Intelligent Systems, Modelling and Simulation. IEEE, Liverpool, United Kingdom, pp. 215–219. https://doi.org/10.1109/ISMS.2010.48 (consultado el 18/03/2024).

Patel, A.A., 2019. Hands-on unsupervised learning using Python: how to build applied machine learning solutions from unlabeled data. O'Reilly Media.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., Dean, J., 2021. Carbon Emissions and Large Neural Network Training. arXiv preprint arXiv:2104.10350.

Punj, G., Stewart, D.W., 2024. Cluster Analysis in Marketing Research: Review and Suggestions for Application. Journal of Marketing Research 20, 134–148.

Reutterer, T., Dan, D., 2020. Cluster Analysis in Marketing Research, in: Homburg, C., Klarmann, M., Vomberg, A. (Eds.), Handbook of Market Research. Springer International Publishing, Cham, pp. 1–29. https://doi.org/10.1007/978-3-319-05542-8_11-2 (consultado el 17/03/2024).

Reynoso, L., 2024. Colombia elude la recesión por muy poco: el PIB crece 0,3% en el último trimestre de 2023 [Documento en línea]. El País América Colombia. URL https://elpais.com/america-colombia/2024-02-15/colombia-elude-la-recesion-por-muy-poco-el-pib-crece-03-en-el-ultimo-trimestre-de-2023.html (consultado el 16/03/2024).

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, San Francisco California USA, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778 (consultado el 25/03/2024).

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7 (consultado el 21/03/2024).

Sarvari, P.A., Ustundag, A., Takci, H., 2016. Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. Kybernetes 45, 1129–1157. https://doi.org/10.1108/K-07-2015-0180 (consultado el 18/03/2024).

SAS, 2017. SAS Help Center: Introduction to SEMMA [Documento en línea]. URL https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm (consultado el 24/03/2024).

Schaninger, C.M., Lessig, V.P., Panton, D.B., 1980. The Complementary Use of Multivariate Procedures to Investigate Nonlinear and Interactive Relationships between Personality and Product Usage. Journal of Marketing Research 17, 119–124.

Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O., 2020. Green AI. Commun. ACM 63, 54–63. https://doi.org/10.1145/3381831 (consultado el 26/03/2024).

Sembiring Brahmana, R.W., Mohammed, F.A., Chairuang, K., 2020. Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods. LKJITI 11, 32. https://doi.org/10.24843/LKJITI.2020.v11.i01.p04 (consultado el 18/03/2024).

Shaw, M.J., Subramaniam, C., Tan, G.W., Welge, M.E., 2001. Knowledge management and data mining for marketing. Decision Support Systems 31, 127–137. https://doi.org/10.1016/S0167-9236(00)00123-8 (consultado el 17/03/2024).

Shirkhorshidi, A.S., Aghabozorgi, S., Wah, T.Y., 2015. A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. PLoS ONE 10, e0144059. https://doi.org/10.1371/journal.pone.0144059 (consultado el 01/06/2024).

Smith, W.R., 1956. Product Differentiation and Market Segmentation as Alternative Marketing Strategies. Journal of Marketing 21, 3–8. https://doi.org/10.1177/002224295602100102 (consultado el 17/03/2024).

Syakur, M.A., Khotimah, B.K., Rochman, E.M.S., Satoto, B.D., 2018. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. IOP Conf. Ser.: Mater. Sci. Eng. 336, 012017. https://doi.org/10.1088/1757-899X/336/1/012017 (consultado el 21/03/2024).

Sze, V., Chen, Y.-H., Yang, T.-J., Emer, J., 2017. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. Proceedings of the IEEE 105 12, 2295–2329.

Tiwari, M., Zhang, M.J., 2020. BanditPAM: Almost Linear Time k-Medoids Clustering via Multi-Armed Bandits. Advances in Neural Information Processing Systems 33, 10211–10222.

UOC, 2023. Impacto global #Agenda2030 [Documento en línea]. URL https://www.uoc.edu/portal/es/compromis-social/index.html (consultado el 16/03/2024).

Van der Maaten, L., Hinton, G., 2008. Visualizing Data using t-SNE. Journal of Machine Learning Research 9, 2579–2605.

VanderPlas, J., 2016. Python data science handbook: Essential tools for working with data. O'Reilly Media, Inc.

Vargas Riaño, D.A., 2024. Deuda externa de Colombia repuntó en 2023 y se ubicó en 53,7% del PIB [Documento en línea]. www.elcolombiano.com. URL https://www.elcolombiano.com/negocios/deuda-externa-de-colombia-en-2023-subio-de-cuanto-fue-DM23940065 (consultado el 16/03/2024).

Vellido, A., 1999. Neural networks in business: a survey of applications (1992–1998). Expert Systems with Applications 17, 51–70. https://doi.org/10.1016/S0957-4174(99)00016-0 (consultado el 17/03/2024).

Verdecchia, R., Cruz, L., Sallou, J., Lin, M., Wickenden, J., Hotellier, E., 2022. Data-Centric Green AI An Exploratory Empirical Study, in: 2022 International Conference on ICT for Sustainability (ICT4S). IEEE, Plovdiv, Bulgaria, pp. 35–45. https://doi.org/10.1109/ICT4S55073.2022.00015 (consultado el 26/03/2024).

Wang, D., Lu, X., Rinaldo, A., 2019. DBSCAN: Optimal Rates For Density-Based Cluster Estimation.

Witten, D.M., Tibshirani, R., 2010. A Framework for Feature Selection in Clustering. Journal of the American Statistical Association 105, 713–726. https://doi.org/10.1198/jasa.2010.tm09415 (consultado el 01/06/2024).

Wu, R.-S., Chou, P.-H., 2011. Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. Electronic Commerce Research

and Applications 10, 331–341. https://doi.org/10.1016/j.elerap.2010.11.002 (consultado el 18/03/2024).

9. Anexos

9.1. Anexo 1

El presente anexo recoge el detalle de los resultados de la primera fase de la segmentación.

ClusterLabels	1	2
Summary		
Counts	5,591	3,867
Distribution (%)	59.11	40.89

Figura 40. Distribución entre clústeres en la primera fase de la segmentación

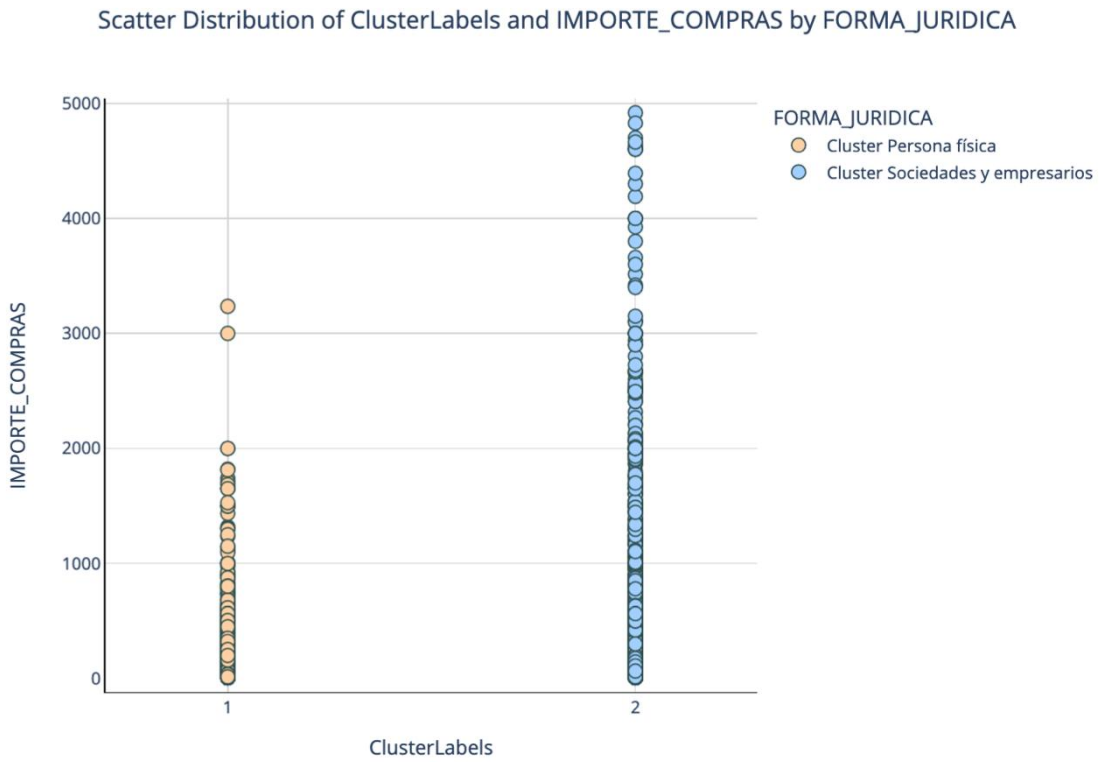


Figura 41. Representación de los clústeres resultado de la primera fase de la segmentación

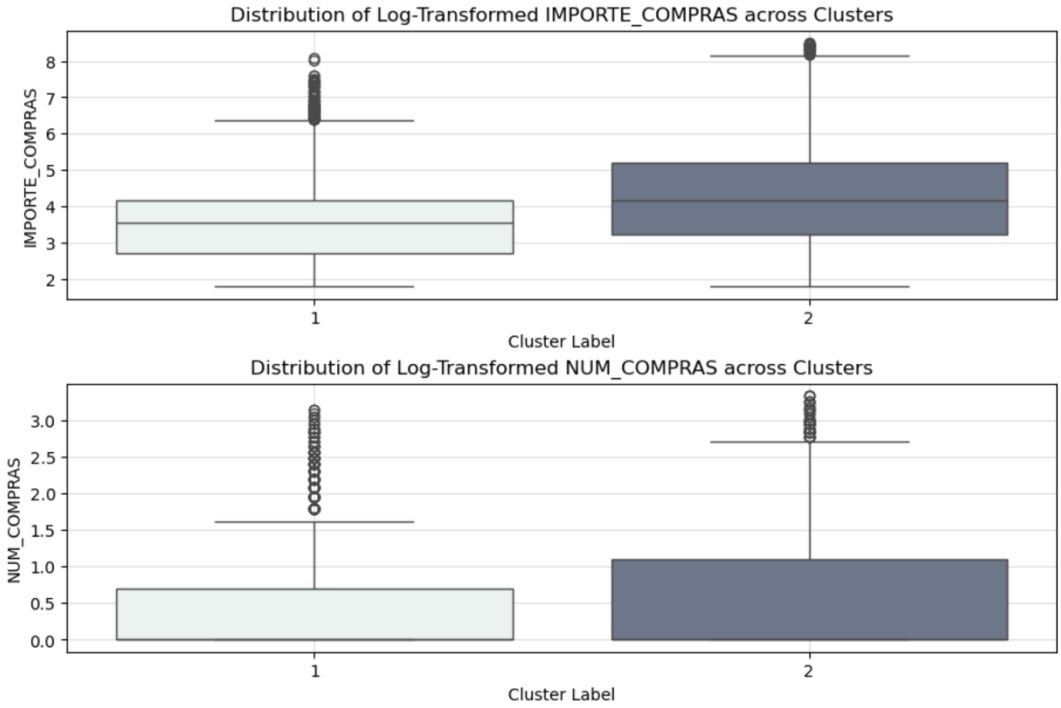


Figura 42. Distribución del importe gastado y del total de compras en la primera fase de la segmentación

	1	2
ANTIGUEDAD	No aplicable	Más de 10 Años
CANAL	Directorios	Directorios
CLIENTEPORCAMPAÑAEMAIL	no	no
CONSUMOSTOTAL mean	9.62	68.65
DEPARTAMENTO	No aplicable	BOGOTA
DEPARTAMENTO_SCORE mean	0.18	-0.27
DIASCLIENTE mean	95.04	106.60
EMPRESASUNICAS_CONSULT mean	6.15	43.38
ESTADO	VIVA	ACTIVA
FORMAJURIDICA	PERSONA FISICA	SOCIEDAD
IMPORTE_COMPRAS mean	74.12	226.67
NUM_COMPRAS mean	1.74	2.42
SECTOR	NOSECTOR	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...
TAMAÑO	No aplicable	MICRO

Figura 43. Perfiles prototípicos de la primera fase de la segmentación

		1	2
Variable	**Statistic**		
IMPORTE_COMPRAS	**min**	6.00	6.00
	max	3,235.00	4,918.00
	mean	74.12	226.67
	median	35.00	65.00
	std	151.60	502.01
NUM_COMPRAS	**min**	1.00	1.00
	max	23.00	28.00
	mean	1.74	2.42
	median	1.00	1.00
	std	1.94	2.95

Figura 44. Estadísticas detalladas del importe gastado y del total de compras en la primera fase de la segmentación

9.2. Anexo 2

El presente anexo recoge el detalle de los resultados de la segunda fase de la segmentación.

ClusterLabels	1	2	3	4	5
Summary					
Counts	2,029	660	651	215	285
Distribution (%)	52.84	17.19	16.95	5.60	7.42

Figura 45. Distribución entre clústeres de sociedades y empresarios en la segunda fase de la segmentación

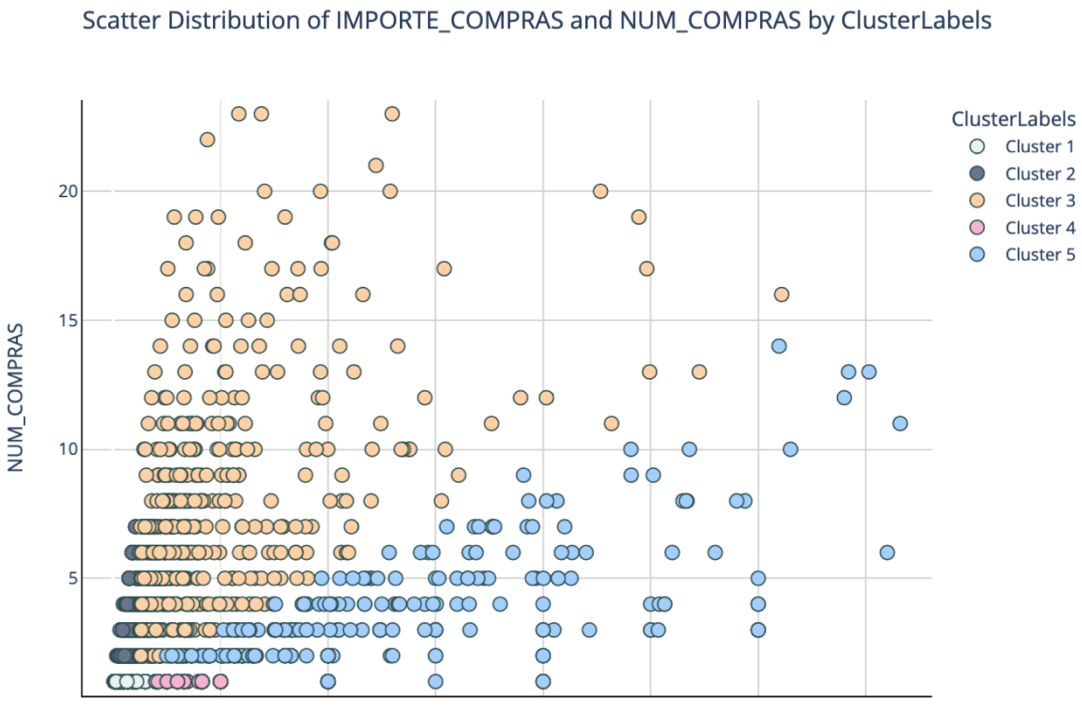


Figura 46. Representación de los clústeres de sociedades y empresarios resultado de la segunda fase de la segmentación

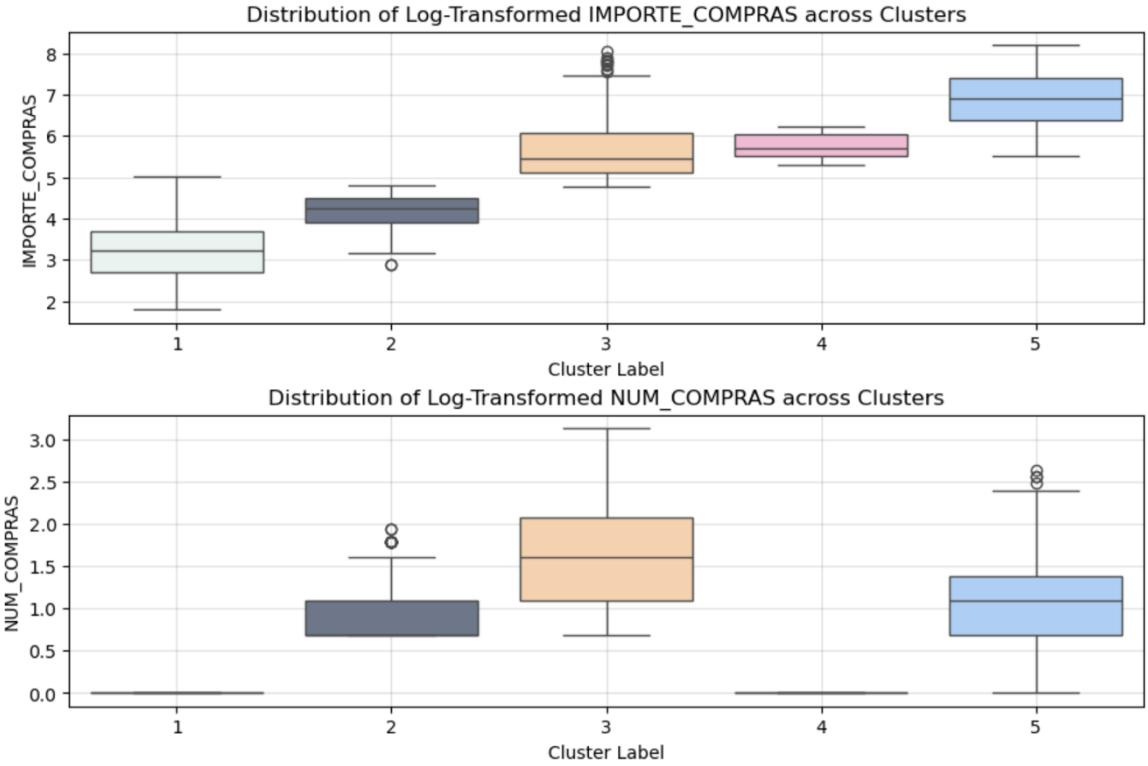


Figura 47. Distribución del importe gastado y del total de compras de sociedades y empresarios en la segunda fase de la segmentación

		1	2	3	4	5
Variable	**Statistic**					
CONSUMOSTOTAL	**min**	1.00	1.00	1.00	1.00	2.00
	max	150.00	37.00	1,644.00	9,638.00	26,657.00
	mean	1.62	3.83	32.62	288.51	593.47
	median	1.00	3.00	11.00	16.00	70.00
	std	4.35	3.60	92.35	1,020.09	2,339.74
ENGAGEMENT	**min**	0.24	0.24	0.26	0.25	0.28
	max	3.74	2.06	23.02	127.97	288.91
	mean	0.35	0.42	1.01	4.59	8.10
	median	0.32	0.39	0.71	0.85	1.82
	std	0.14	0.17	1.34	13.73	27.04
IMPORTE_COMPRAS	**min**	6.00	18.00	120.00	200.00	250.00
	max	150.00	123.00	3,109.00	500.00	3,660.00
	mean	38.39	70.92	373.78	344.07	1,204.59
	median	25.00	70.00	237.00	300.00	1,000.00
	std	31.63	25.84	371.58	96.42	784.03
NUM_COMPRAS	**min**	1.00	2.00	2.00	1.00	1.00
	max	1.00	7.00	23.00	1.00	14.00
	mean	1.00	2.59	6.19	1.00	3.46
	median	1.00	2.00	5.00	1.00	3.00
	std	0.00	0.95	4.14	0.00	2.25

Figura 48. Estadísticas detalladas de los consumos, el *engagement score*, el importe gastado y el total de compras para sociedades y empresarios en la segunda fase de la segmentación

	1	2	3	4	5
ANTIGUEDAD	Más de 10 Años	Más de 10 Años	Más de 10 Años	Más de 10 Años	Más de 10 Años
AOV mean	38.39	28.62	64.10	344.07	399.59
CANAL	Directorios	Directorios	WEB	WEB	WEB
CLIENTEPORCAMPAÑAEMAIL	no	no	no	no	no
CLV mean	544.35	405.86	908.89	4,878.44	5,665.71
CONSUMOSTOTAL mean	1.62	3.83	32.62	288.51	593.47
DEPARTAMENTO	BOGOTA	BOGOTA	BOGOTA	BOGOTA	BOGOTA
DEPARTAMENTO_SCORE mean	-0.35	-0.22	-0.16	-0.15	-0.09
DIASCLIENTE mean	114.97	132.18	79.86	70.75	83.44
DIVERSIDAD_COMPRAS mean	1.00	1.01	1.32	1.00	1.43
DIVERSIDAD_EMPRESAS_CONSULT mean	1.28	2.13	14.34	245.70	347.26
DIVERSIDAD_ESTADOS_CONSULT mean	1.04	1.15	1.79	2.96	3.70
DIVERSIDAD_SECTORES_CONSULT mean	1.09	1.36	3.15	5.79	7.92
DIVERSIDAD_TAMAÑOS_CONSULT mean	1.08	1.34	2.60	3.25	4.34
EMPRESASUNICAS_CONSULT mean	1.28	2.13	14.34	245.70	347.26
ENGAGEMENT mean	0.35	0.42	1.01	4.59	8.10
ESTADO	ACTIVA	ACTIVA	ACTIVA	ACTIVA	ACTIVA
FORMAJURIDICA	SOCIEDAD	SOCIEDAD	SOCIEDAD	SOCIEDAD	SOCIEDAD
FRECUENCIA_CONSULT mean	1.62	3.83	32.62	288.51	593.47
IMPORTE_COMPRAS mean	38.39	70.92	373.78	344.07	1,204.59
NUM_COMPRAS mean	1.00	2.59	6.19	1.00	3.46
RECENCIA_CONSULT mean	1,150.80	1,052.18	707.75	1,124.51	540.66
SECTOR	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...
TAMAÑO	MICRO	MICRO	PEQUEÑA	MICRO	MEDIANA
TENDENCIA_FRECUENCIA_CONSULT mean	0.00	0.01	-0.05	-0.59	-0.03
TENDENCIA_FRECUENCIA_VENTAS mean	0.00	-0.01	-0.14	0.00	-0.07
VIDACLIENTE mean	1,316.18	1,446.77	1,352.39	1,362.69	1,305.77

Figura 49. Perfiles prototípicos de sociedades y empresarios en la segunda fase de la segmentación

ClusterLabels	1	2	3
Summary			
Counts	3,330	1,469	752
Distribution (%)	59.99	26.46	13.55

Figura 50. Distribución entre clústeres de personas físicas en la segunda fase de la segmentación

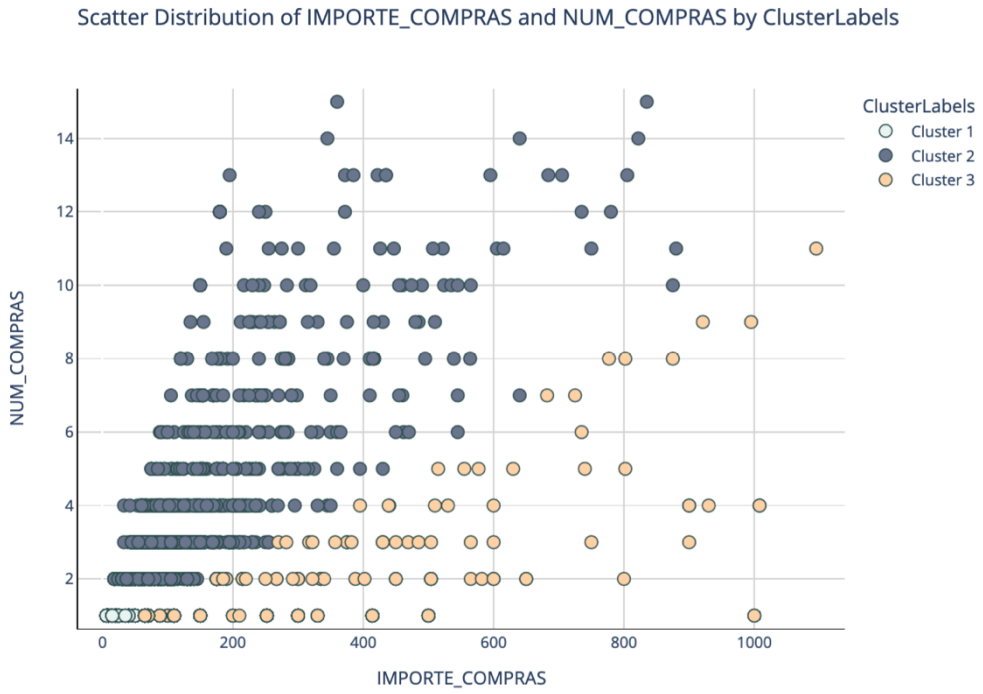


Figura 51. Representación de los clústeres de personas físicas resultado de la segunda fase de la segmentación

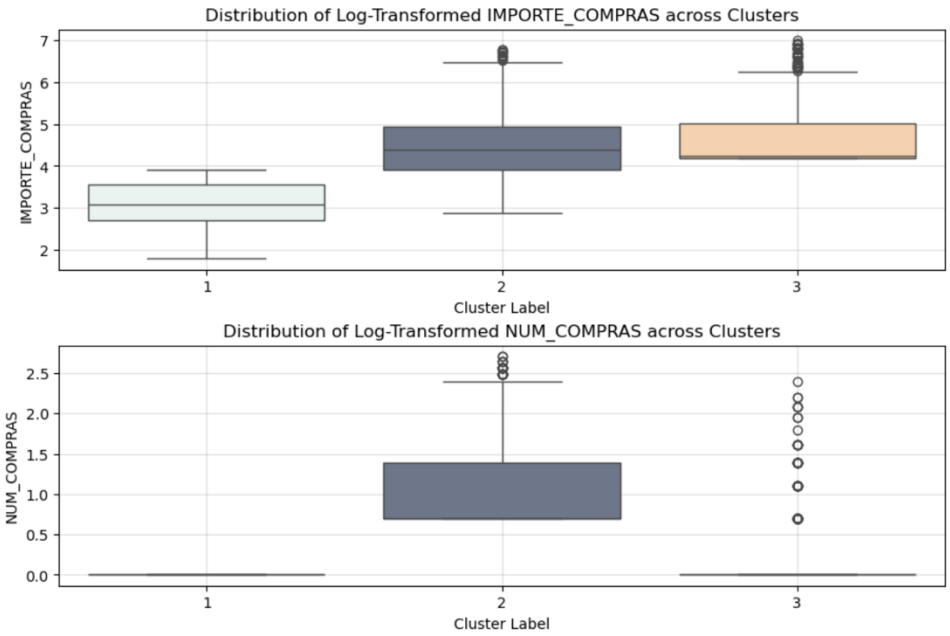


Figura 52. Distribución del importe gastado y del total de compras de personas físicas en la segunda fase de la segmentación

	1	2	3
ANTIGUEDAD	No aplicable	No aplicable	No aplicable
AOV mean	23.70	34.12	111.23
CANAL	Directorios	Directorios	WEB
CLIENTEPORCAMPAÑAEMAIL	no	no	no
CLV mean	336.03	483.78	1,577.09
CONSUMOSTOTAL mean	1.42	7.05	37.90
DEPARTAMENTO	No aplicable	No aplicable	No aplicable
DEPARTAMENTO_SCORE mean	0.18	0.18	0.18
DIASCLIENTE mean	94.57	98.76	90.50
DIVERSIDAD_COMPRAS mean	1.00	1.03	1.08
DIVERSIDAD_EMPRESAS_CONSULT mean	1.15	3.26	26.70
DIVERSIDAD_ESTADOS_CONSULT mean	1.03	1.28	1.41
DIVERSIDAD_SECTORES_CONSULT mean	1.06	1.67	2.03
DIVERSIDAD_TAMAÑOS_CONSULT mean	1.05	1.55	1.54
EMPRESASCONRELACION mean	1.59	2.57	1.93
EMPRESASUNICAS_CONSULT mean	1.15	3.26	26.70
ENGAGEMENT mean	0.34	0.51	0.91
ESTADO	VIVA	VIVA	VIVA
FORMAJURIDICA	PERSONA FISICA	PERSONA FISICA	PERSONA FISICA
FRECUENCIA_CONSULT mean	1.42	7.05	37.90
IMPORTE_COMPRAS mean	23.70	118.07	148.52
NUM_COMPRAS mean	1.00	3.37	1.26
RECENCIA_CONSULT mean	1,116.91	953.45	1,056.84
SECTOR	NOSECTOR	NOSECTOR	NOSECTOR
TAMAÑO	No aplicable	No aplicable	No aplicable
TENDENCIA_FRECUENCIA_CONSULT mean	-0.01	0.00	-0.19
TENDENCIA_FRECUENCIA_VENTAS mean	0.00	-0.05	0.01
TIPODOMINIOEMAIL	GOO-MS-YAH-APP	GOO-MS-YAH-APP	GOO-MS-YAH-APP
VIDACLIENTE mean	1,250.56	1,362.54	1,242.01

Figura 53. Perfiles prototípicos de personas físicas en la segunda fase de la segmentación

Variable	Statistic	1	2	3
CONSUMOSTOTAL	min	1.00	1.00	1.00
	max	126.00	232.00	11,429.00
	mean	1.42	7.05	37.90
	median	1.00	3.00	1.00
	std	2.71	12.09	436.48
EMPRESASCONRELACION	min	0.00	0.00	0.00
	max	81.00	90.00	71.00
	mean	1.59	2.57	1.93
	median	0.00	0.00	0.00
	std	4.50	5.77	5.45
ENGAGEMENT	min	0.24	0.25	0.24
	max	2.92	4.86	139.95
	mean	0.34	0.51	0.91
	median	0.33	0.41	0.36
	std	0.10	0.32	5.43
IMPORTE_COMPRAS	min	6.00	18.00	65.00
	max	50.00	880.00	1,095.00
	mean	23.70	118.07	148.52
	median	22.00	80.00	70.00
	std	10.37	109.78	170.97
NUM_COMPRAS	min	1.00	2.00	1.00
	max	1.00	15.00	11.00
	mean	1.00	3.37	1.26
	median	1.00	2.00	1.00
	std	0.00	2.30	0.98

Figura 54. Estadísticas detalladas de los consumos, el *engagement score*, el importe gastado y el total de compras para personas físicas en la segunda fase de la segmentación

SEGMENTO	1	2	3	4	5	6	7	8	9
Summary									
Counts	3,330	2,029	660	1,469	752	215	651	285	121
Distribution (%)	35.01	21.33	6.94	15.44	7.91	2.26	6.84	3.00	1.27

Figura 55. Distribución completa entre clústeres resultado de la segunda fase de la segmentación

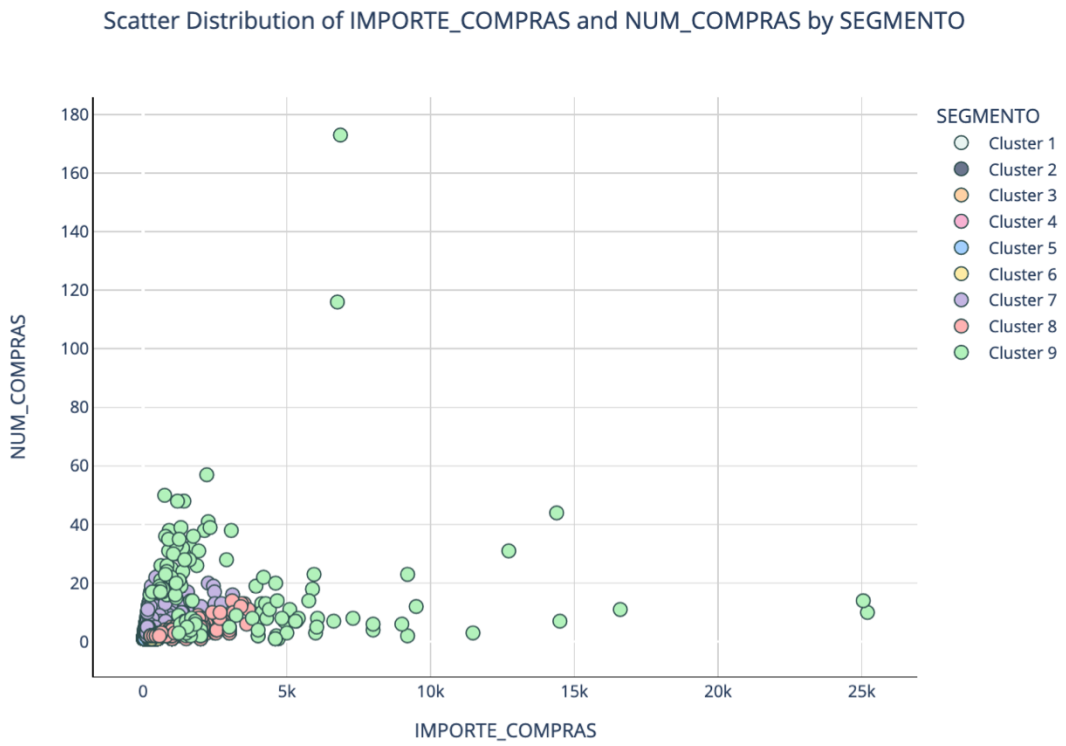


Figura 56. Representación completa de los clústeres resultado de la segunda fase de la segmentación

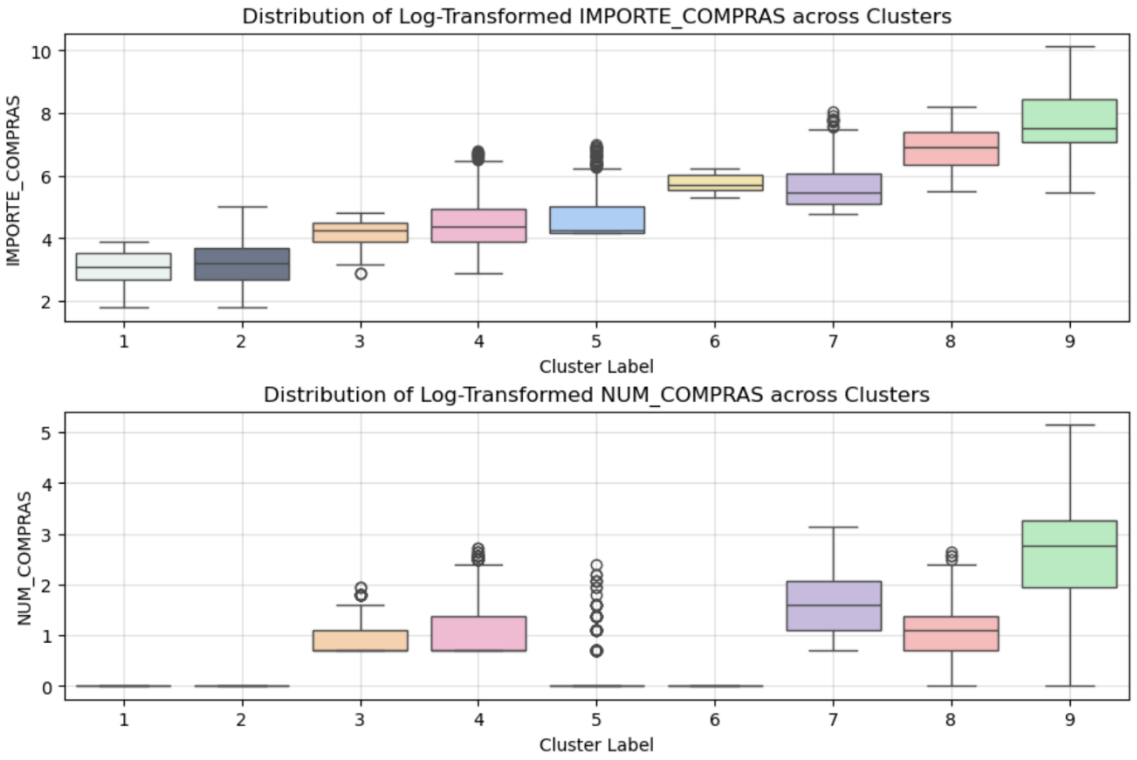


Figura 57. Distribución completa del importe gastado y del total de compras en la segunda fase de la segmentación

	1	2	3	4	5	6	7	8	9
ANTIGUEDAD	No aplicable	Más de 10 Años	Más de 10 Años	No aplicable	No aplicable	Más de 10 Años	Más de 10 Años	Más de 10 Años	Más de 10 Años
CANAL	Directorios	Directorios	Directorios	Directorios	WEB	WEB	WEB	WEB	WEB
CLIENTEPORCAMPAÑAEMAIL	no	no	no	no	no	no	no	no	no
CONSUMOSTOTAL mean	1.42	1.62	3.83	7.05	37.90	288.51	32.62	593.47	5,493.93
DEPARTAMENTO	No aplicable	BOGOTA	BOGOTA	No aplicable	No aplicable	BOGOTA	BOGOTA	BOGOTA	BOGOTA
DIASCLIENTE mean	94.57	114.97	132.17	98.76	90.49	70.74	79.87	83.43	48.51
EMPRESASUNICAS_CONSULT mean	1.15	1.28	2.13	3.26	26.70	245.70	14.34	347.26	2,370.16
ESTADO	VIVA	ACTIVA	ACTIVA	VIVA	VIVA	ACTIVA	ACTIVA	ACTIVA	ACTIVA
FORMAJURIDICA	PERSONA FISICA	SOCIEDAD	SOCIEDAD	PERSONA FISICA	PERSONA FISICA	SOCIEDAD	SOCIEDAD	SOCIEDAD	SOCIEDAD
IMPORTE_COMPRAS mean	23.70	38.39	70.92	118.07	148.52	344.07	373.78	1,204.59	3,722.84
NUM_COMPRAS mean	1.00	1.00	2.59	3.37	1.26	1.00	6.19	3.46	19.27
SECTOR	NOSECTOR	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	NOSECTOR	NOSECTOR	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACI...	NOSECTOR
TAMAÑO	No aplicable	MICRO	MICRO	No aplicable	No aplicable	MICRO	PEQUEÑA	MEDIANA	GRANDE

Figura 58. Perfiles prototípicos completos de la segunda fase de la segmentación

		1	2	3	4	5	6	7	8	9
Variable	**Statistic**									
IMPORTE_COMPRAS	**min**	6.00	6.00	18.00	18.00	65.00	200.00	120.00	250.00	240.00
	max	50.00	150.00	123.00	880.00	1,095.00	500.00	3,109.00	3,660.00	25,200.00
	mean	23.70	38.39	70.92	118.07	148.52	344.07	373.78	1,204.59	3,722.84
	median	22.00	25.00	70.00	80.00	70.00	300.00	237.00	1,000.00	1,870.00
	std	10.37	31.63	25.84	109.78	170.97	96.42	371.58	784.03	4,221.73
NUM_COMPRAS	**min**	1.00	1.00	2.00	2.00	1.00	1.00	2.00	1.00	1.00
	max	1.00	1.00	7.00	15.00	11.00	1.00	23.00	14.00	173.00
	mean	1.00	1.00	2.59	3.37	1.26	1.00	6.19	3.46	19.27
	median	1.00	1.00	2.00	2.00	1.00	1.00	5.00	3.00	16.00
	std	0.00	0.00	0.95	2.30	0.98	0.00	4.14	2.25	21.03

Figura 59. Estadísticas detalladas completas del importe gastado y el total de compras en la segunda fase de la segmentación

9.3. Anexo 3

El presente anexo recoge el detalle de los resultados de predicción de transferencia entre segmentos.

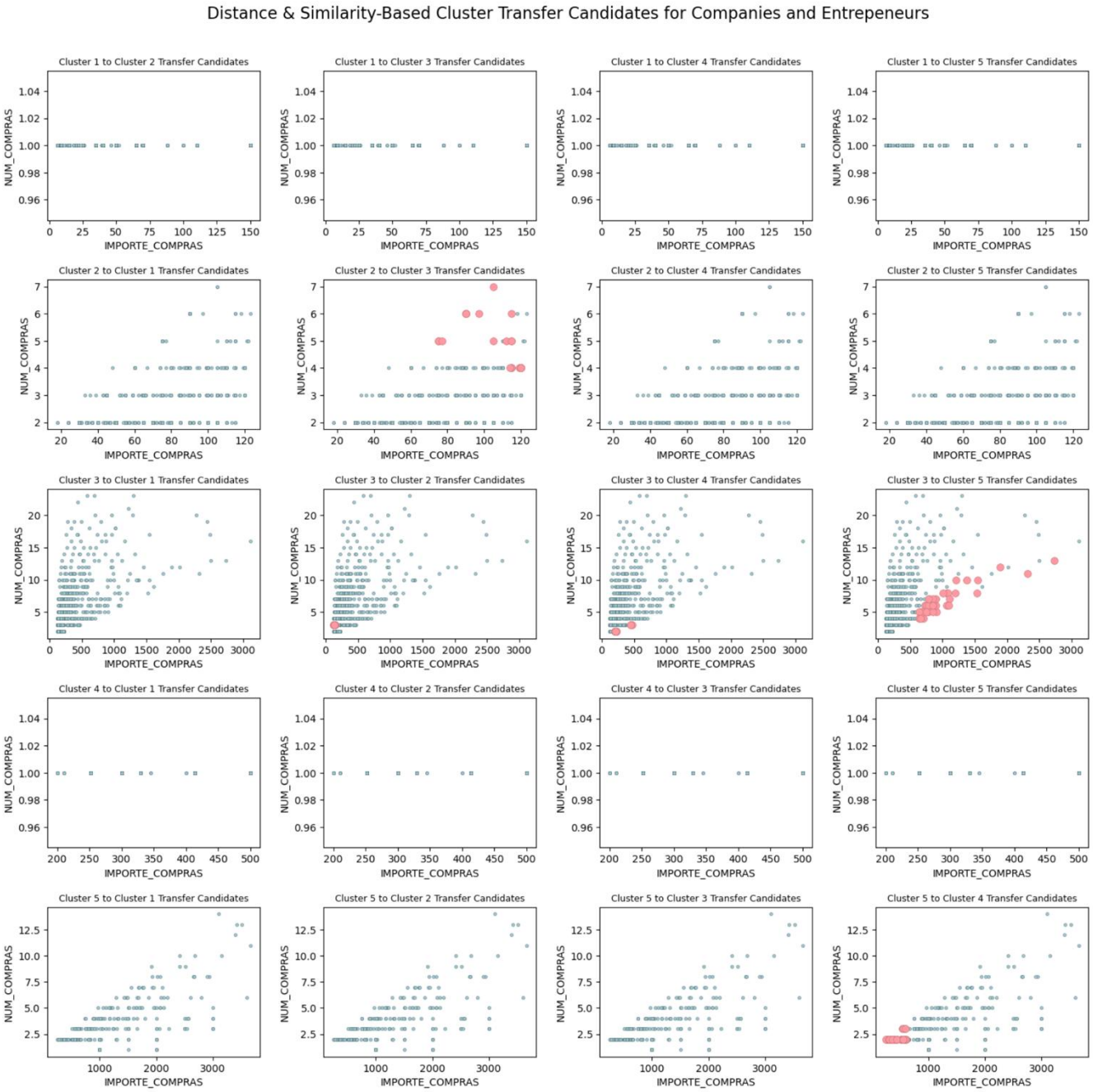


Figura 60. Predicción de transferencia entre segmentos de sociedades y empresarios basada en distancia y similitud

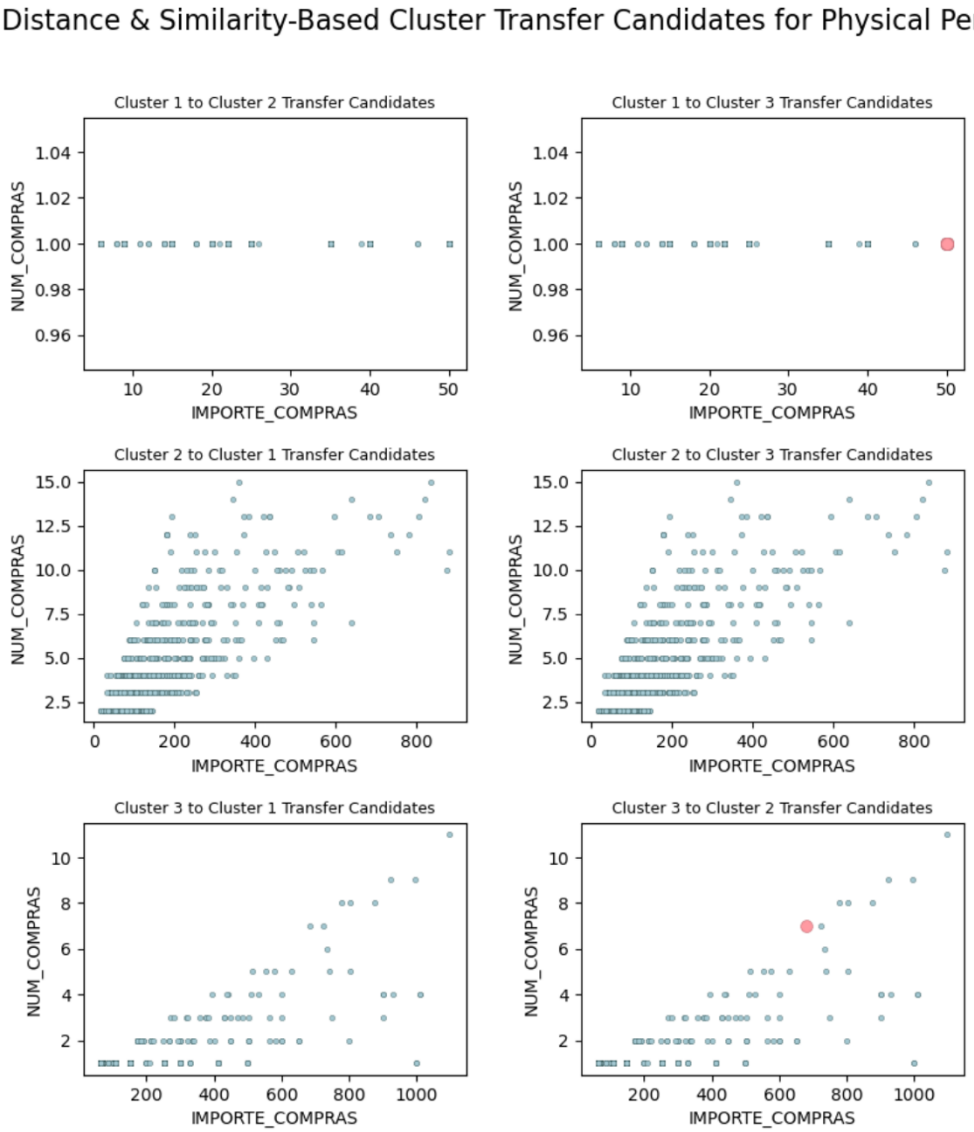


Figura 61. Predicción de transferencia entre segmentos de personas físicas basada en distancia y similitud

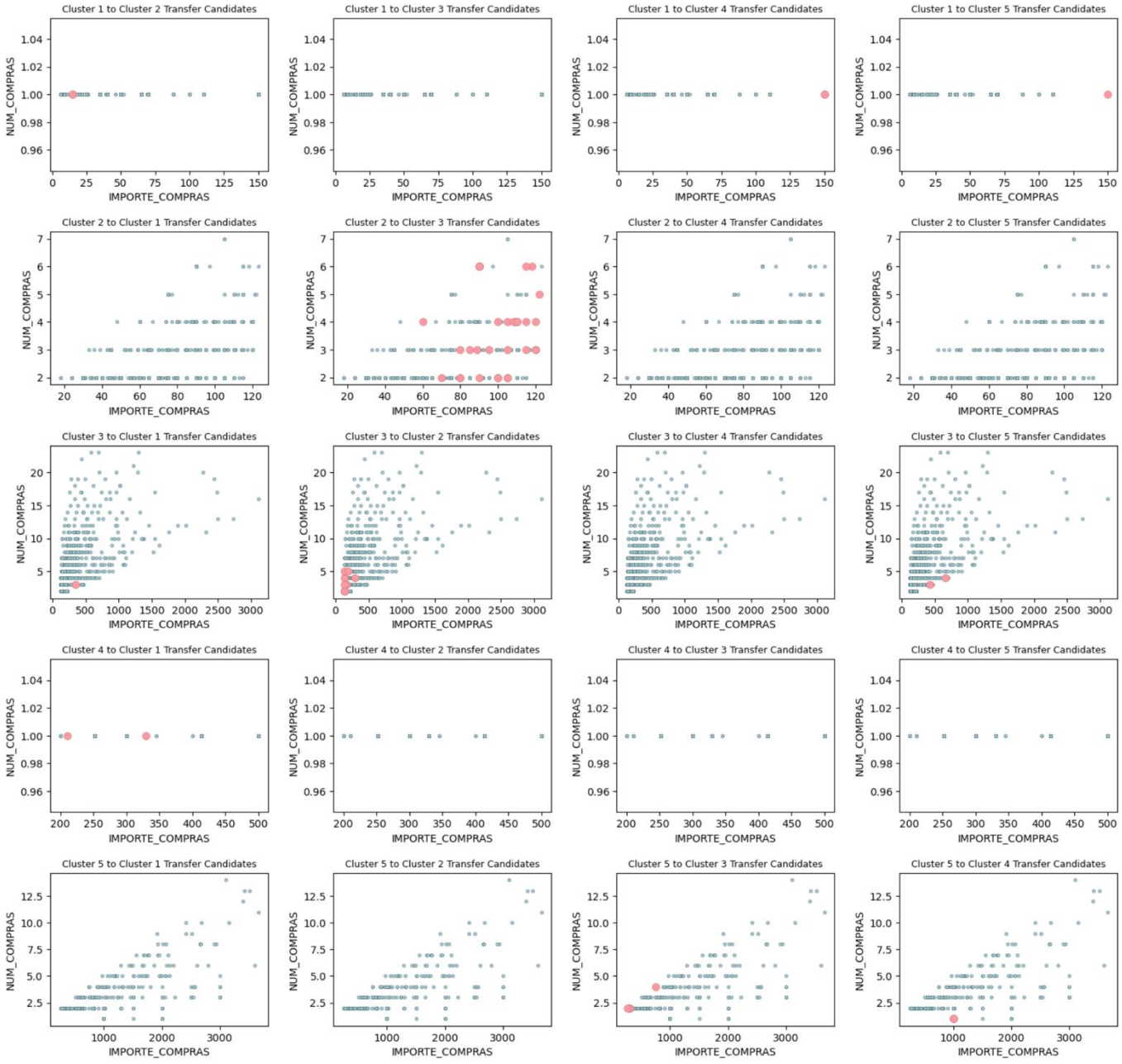


Figura 62. Predicción de transferencia entre segmentos de sociedades y empresarios basada en clasificación multinomial

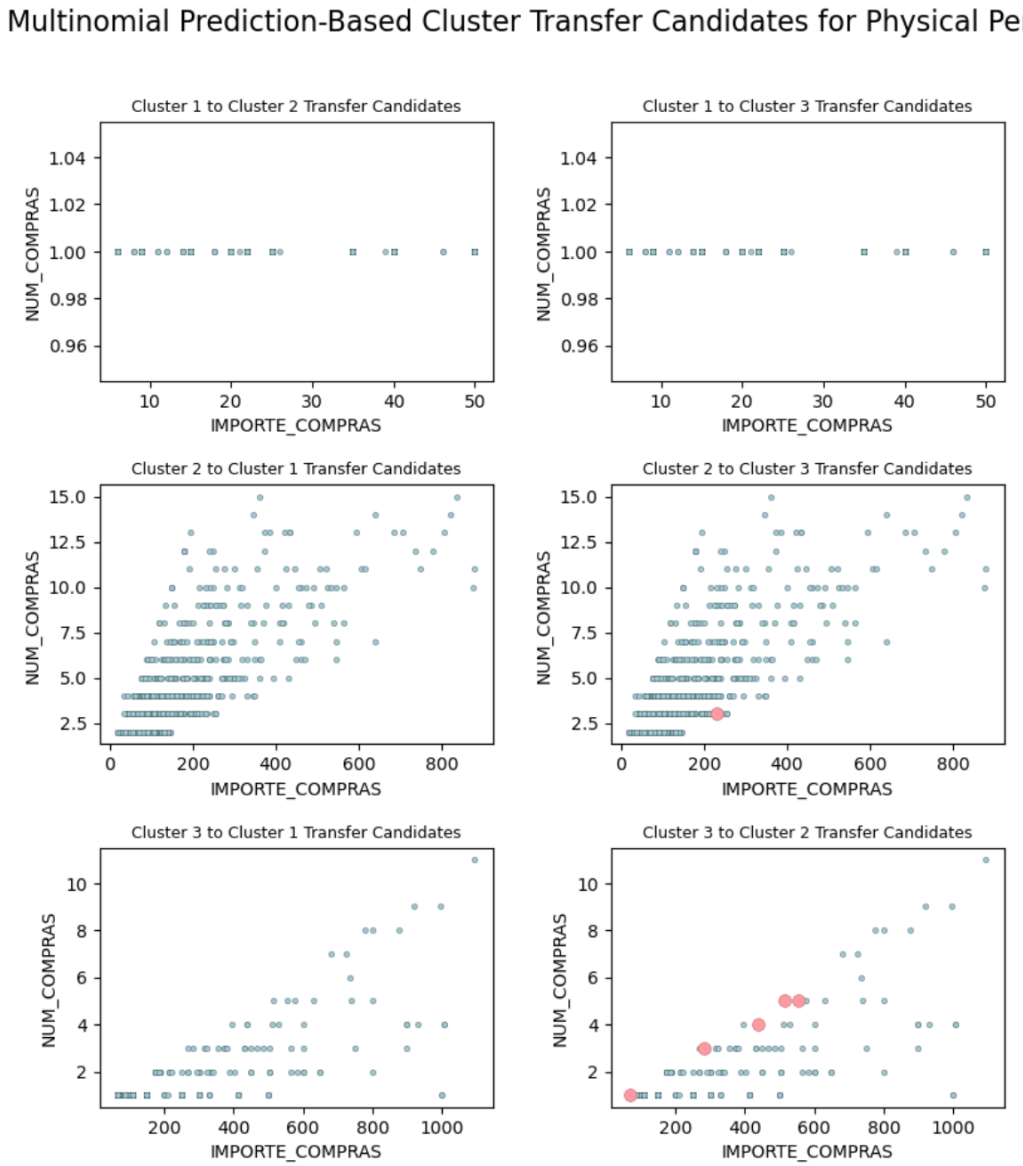


Figura 63. Predicción de transferencia entre segmentos de personas físicas basada en clasificación multinomial

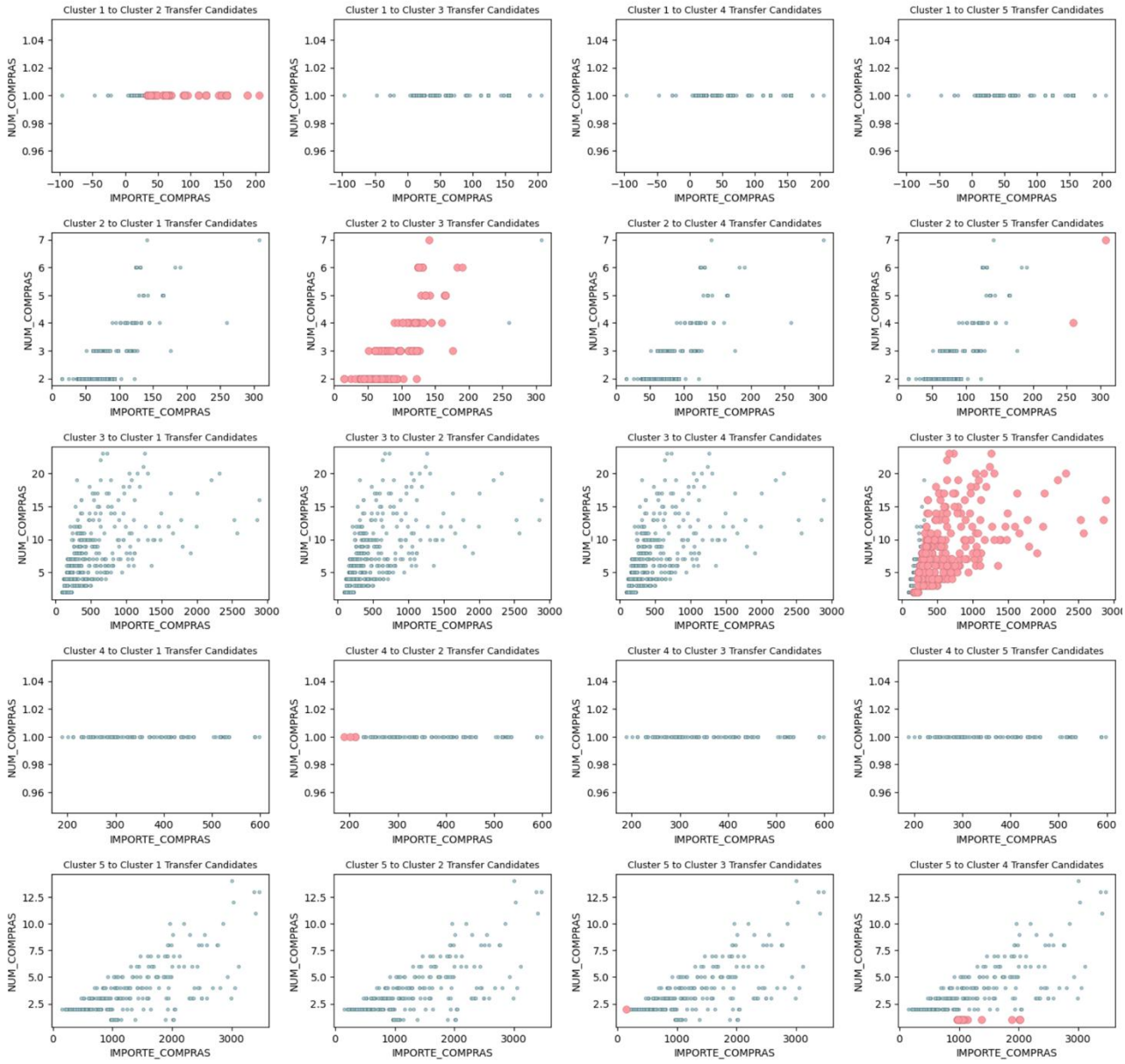


Figura 64. Predicción de transferencia entre segmentos de sociedades y empresarios basada en regresión lineal y reagrupamiento

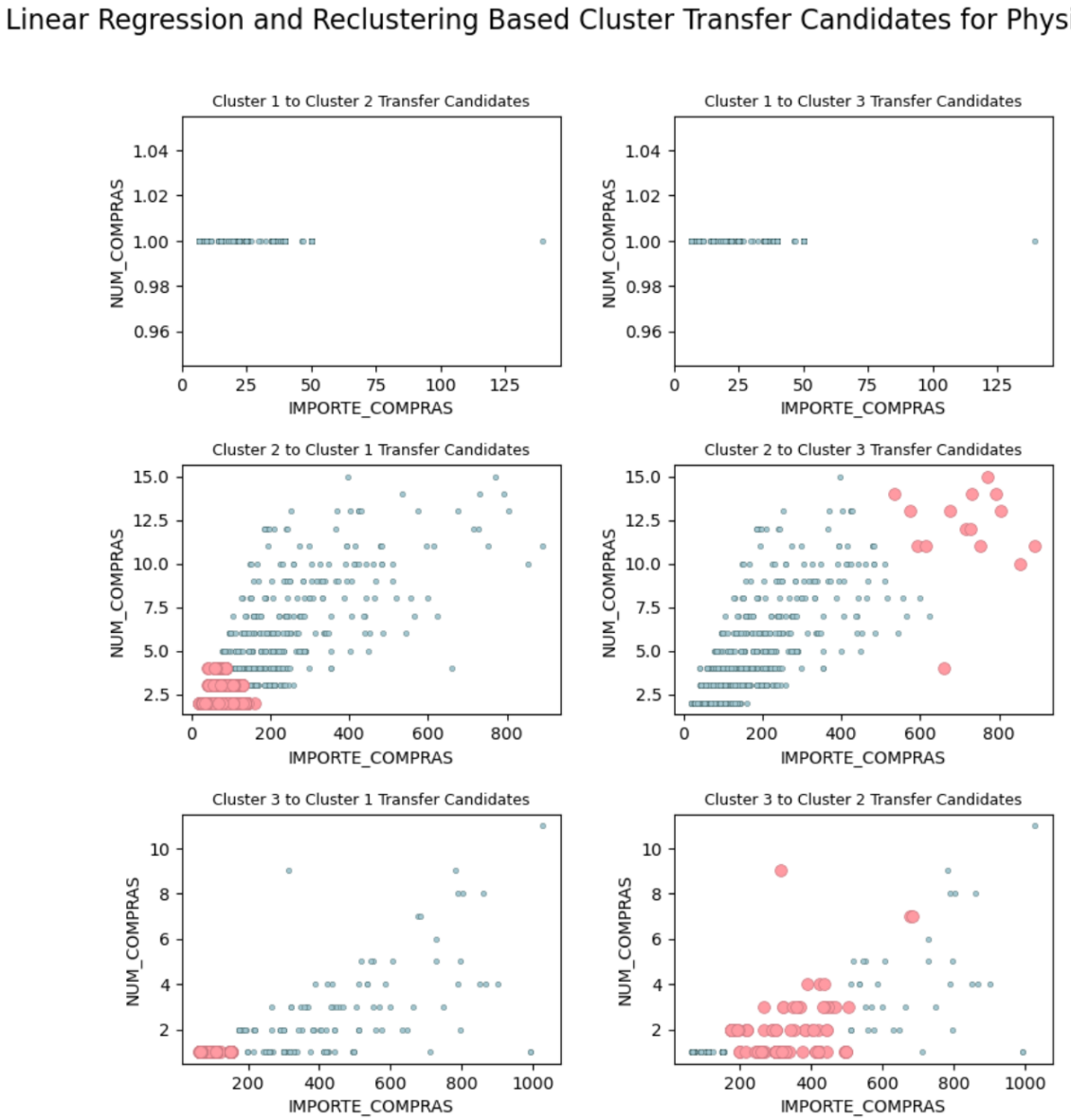


Figura 65. Predicción de transferencia entre segmentos de personas físicas basada en regresión lineal y reagrupamiento

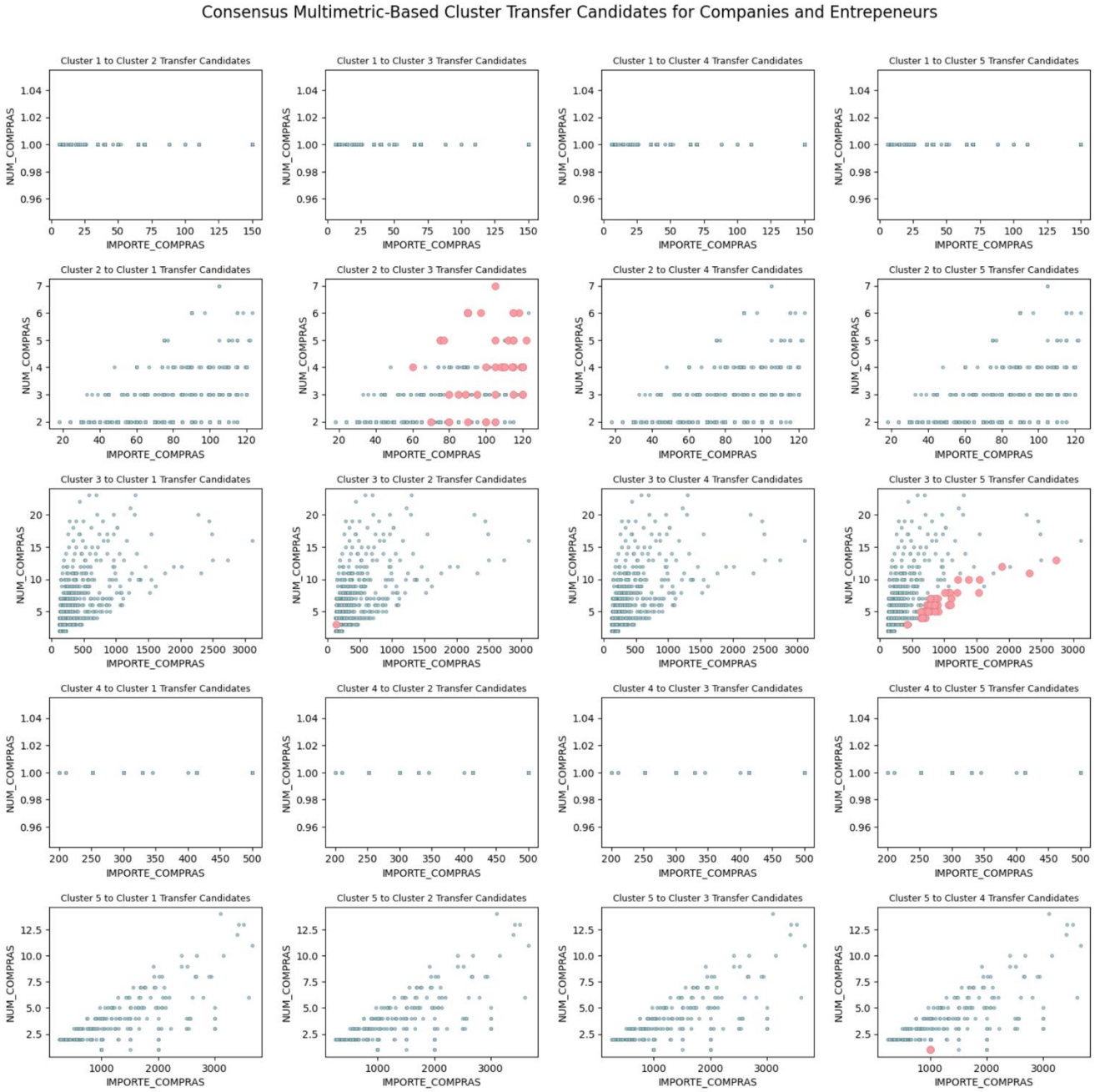


Figura 66. Predicción de transferencia entre segmentos de sociedades y empresarios por consenso de métricas

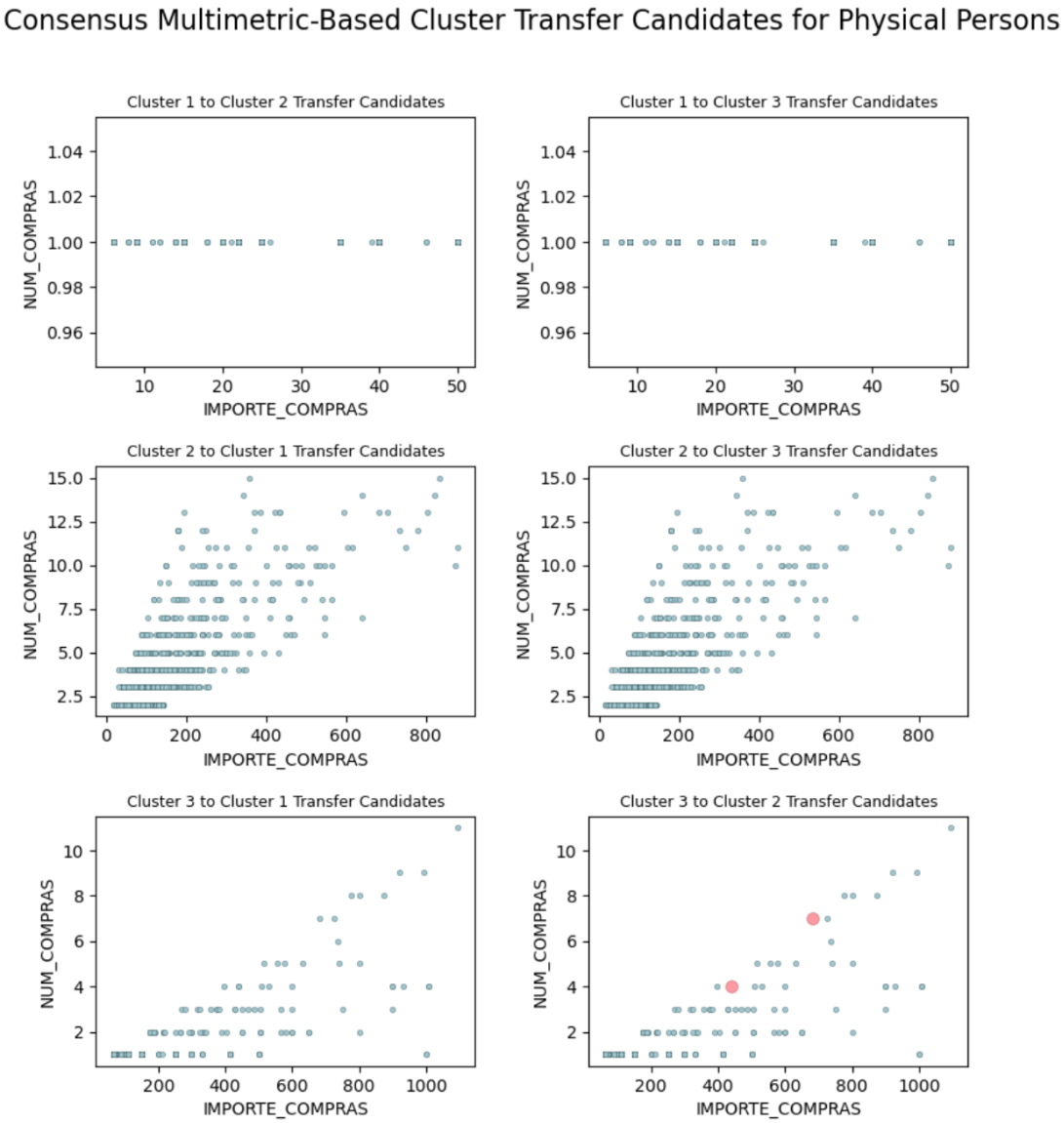


Figura 67. Predicción de transferencia entre segmentos de personas físicas por consenso de métricas

9.4. Anexo 4

El presente anexo recoge el detalle de la segmentación por recurrencia de compra.

RecurrenceLabels	1	2	3
Summary			
Counts	7,696	966	703
Distribution (%)	82.18	10.32	7.51

Figura 68. Distribución entre clústeres para la segmentación por recurrencia de compra

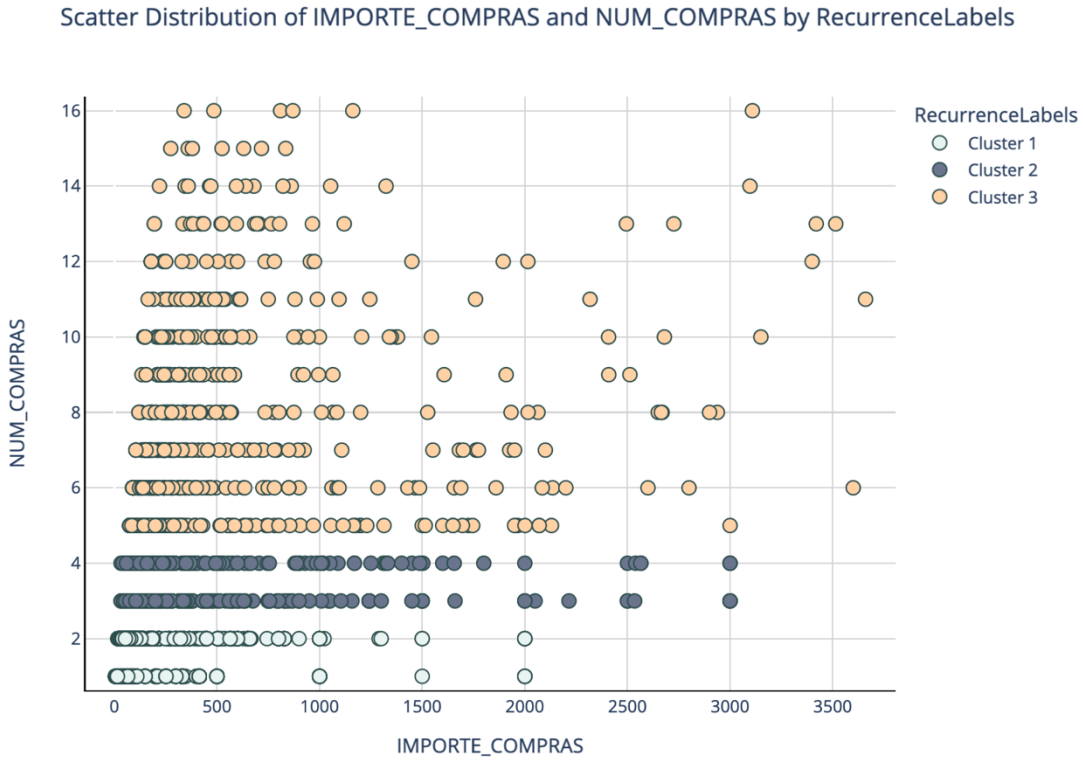


Figura 69. Representación de los clústeres resultado de la segmentación por recurrencia de compra

		1	2	3
Variable	**Statistic**			
NUM_COMPRAS	**min**	1.00	3.00	5.00
	max	2.00	4.00	16.00
	mean	1.19	3.37	7.56
	median	1.00	3.00	7.00
	std	0.39	0.48	2.65

Figura 70. Estadísticas detalladas del total de compras para la segmentación por recurrencia de compra

9.5. Anexo 5

El presente anexo recoge el detalle de los resultados de predicción de transferencia entre segmentos de recurrencia de compra.

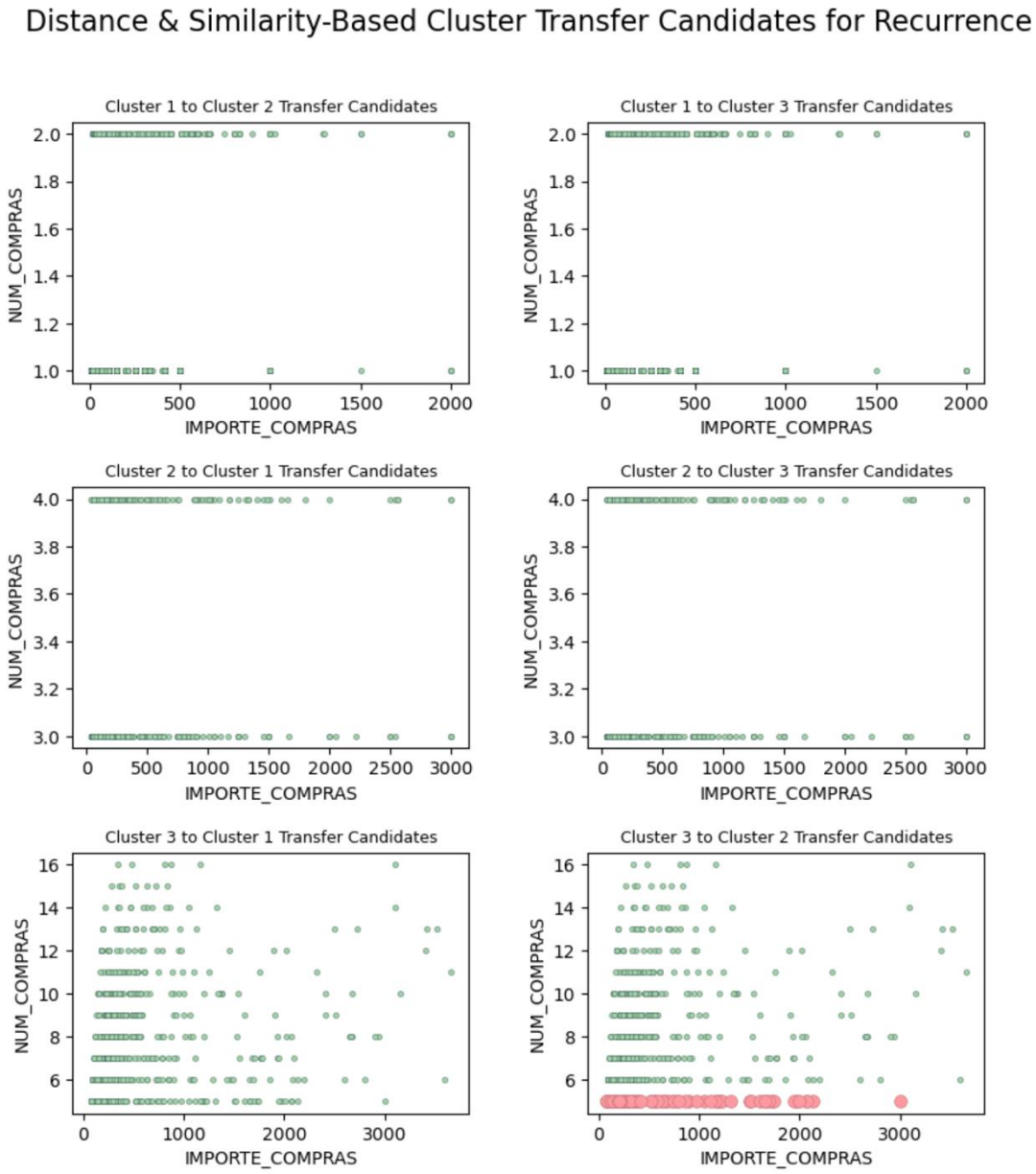


Figura 71. Predicción de transferencia entre segmentos de recurrencia de compra basada en distancia y similitud

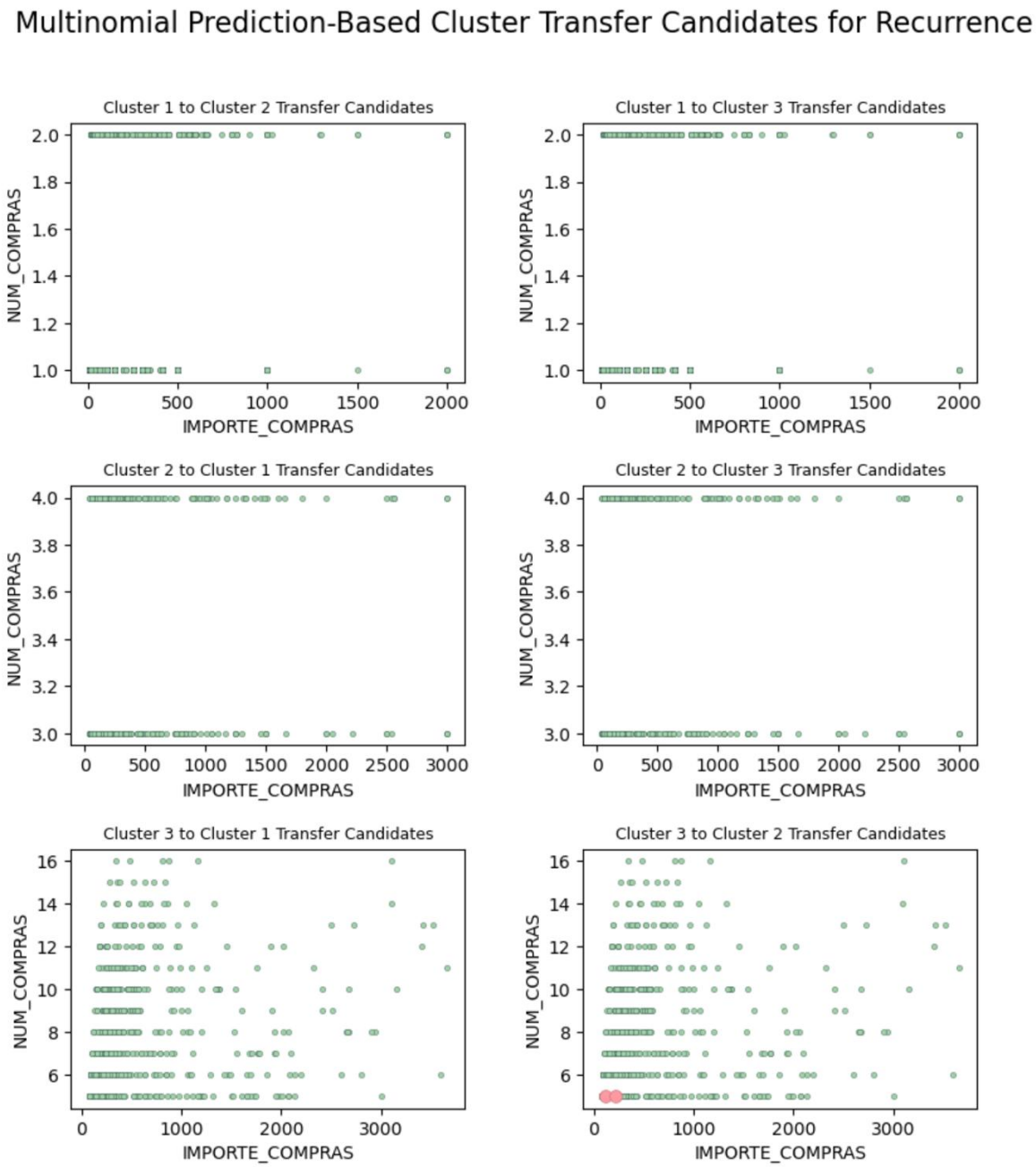


Figura 72. Predicción de transferencia entre segmentos de recurrencia de compra basada en clasificación multinomial

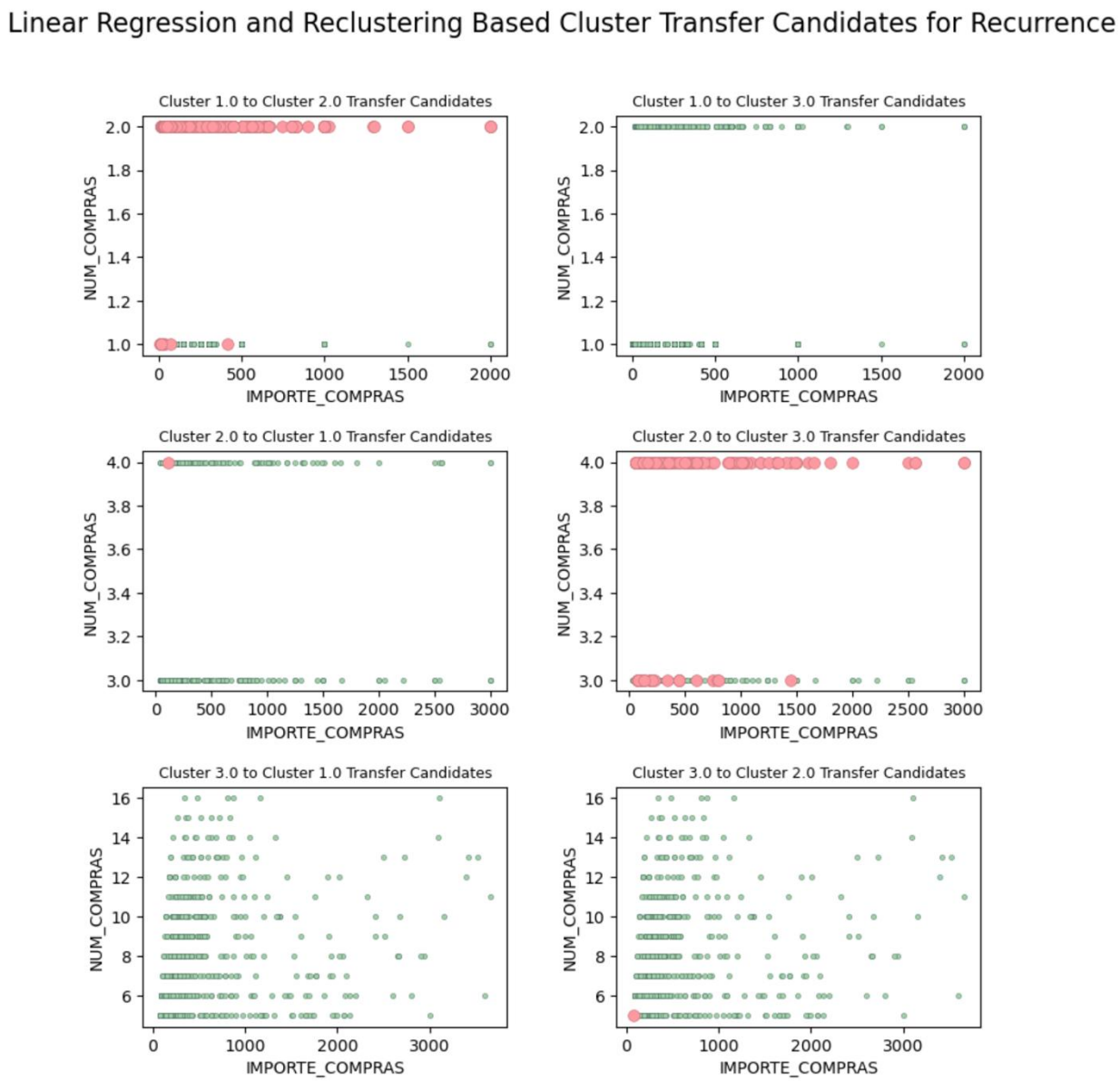


Figura 73. Predicción de transferencia entre segmentos de recurrencia de compra basada en regresión lineal y reagrupamiento

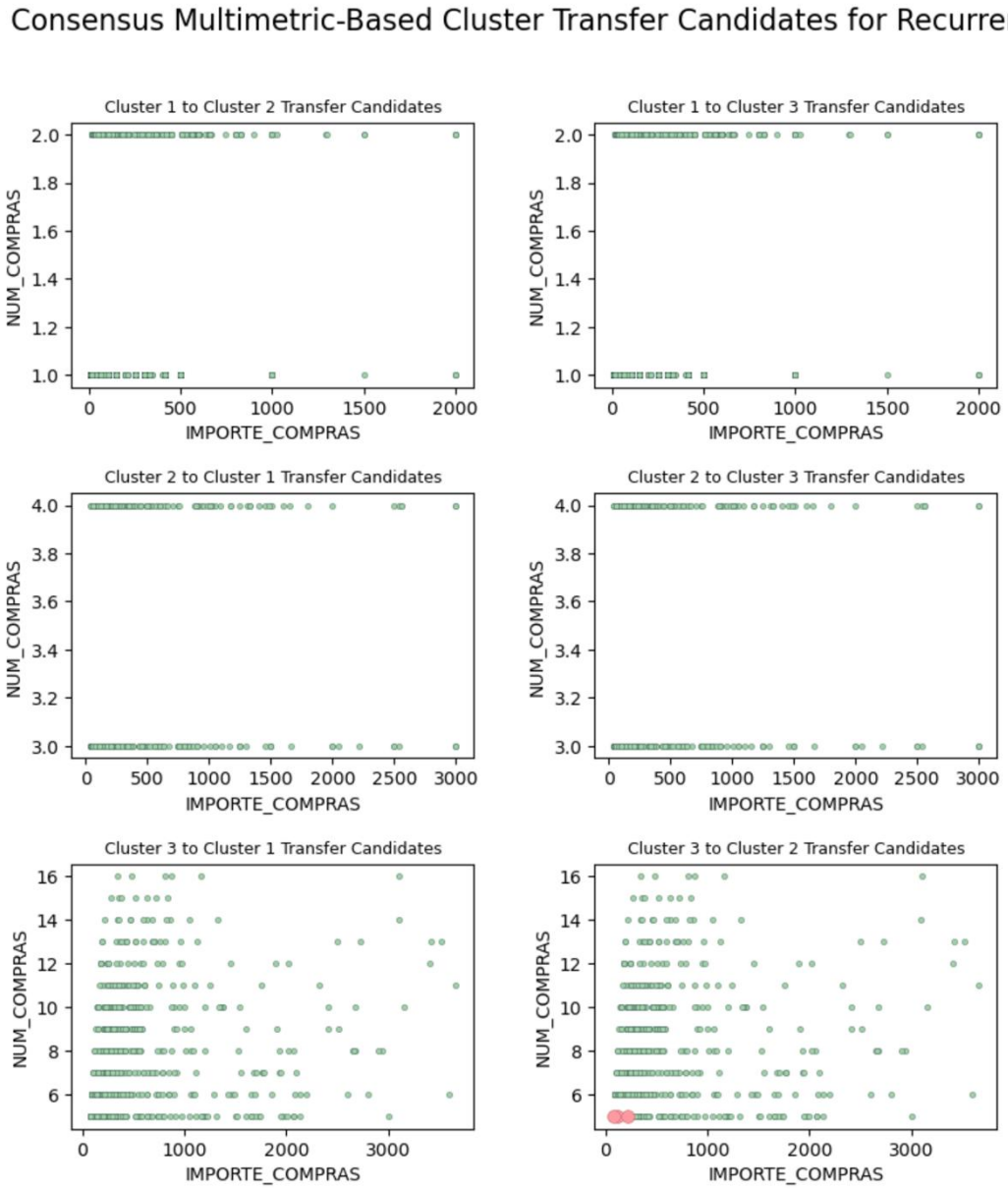


Figura 74. Predicción de transferencia entre segmentos de recurrencia de compra por consenso de métricas