PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD INGENIERÍA

MAESTRÍA EN SISTEMAS DE INFORMACIÓN MENCIÓN CIENCIA DE DATOS

DISERTACIÓN PREVIA A LA OBTENCIÓN DEL TÍTULO DE MÁSTER EN SISTEMAS DE INFORMACIÓN MENCIÓN CIENCIA DE DATOS

SEGMENTACIÓN DE CLIENTES (SOCIOS) PARA LA RECOMENDACIÓN DE PRODUCTOS DE COLOCACIÓN Y/O CAPTACIÓN PARA INSTITUCIÓN FINANCIERA EN ECUADOR

CRISTINA VICTORIA CHAPA ZUMBA

DIRECTOR: JHONNY PINCAY NIEVES

QUITO, JUNIO 2023

DECLARACIÓN DE AUTENTICIDAD Y RESPONSABILIDAD

Yo, Cristina Victoria Chapa Zumba portadora de la cédula de ciudadanía No. 0104965702 declaro que los resultados obtenidos en la investigación de trabajo de titulación, previo la obtención del título "Máster en Sistemas de Información Mención Ciencia de Datos" son absolutamente originales, auténticos y personales.

En tal virtud, declaro que el contenido, las conclusiones y los efectos legales y académicos que se desprenden del trabajo propuesto de investigación son y serán de mi sola y exclusiva responsabilidad y académica.

Cristina Victoria Chapa Zumba

0104965702

Índice General

Índice de Figuras

Índice de Tablas

Índice de Anexos

Resumen Ejecutivo

La segmentación de clientes permite a las instituciones financieras mejorar su estrategia de negocio, centrándose en su mercado objetivo mediante campañas de marketing dirigidas. Las instituciones al no disponer de técnicas para segmentar a sus clientes se les dificultan llegar a sus objetivos, asumiendo mayor riesgo en nichos de mercados desconocidos. El objetivo del presente estudio es recomendar productos financieros a clientes potenciales de una institución financiera del Ecuador mediante la técnica de machine learning, clúster k-means. La base de información son los clientes de la institución financiera del Ecuador con sus características corte diciembre del 2022; se emplea un enfoque transversal para la aplicación del modelo. Los resultados muestran que la técnica de clúster recomienda segmentar en diez grupos a los clientes, siendo potenciales tres, desde el punto de vista de la demanda que tienen en cualquiera de los productos de crédito, tarjetas de crédito, ahorro, y depósitos a plazo fijo.

Palabras clave: Segmentación clientes, mercado objetivo, estrategia, machine learning, institución financiera, minería de datos, k-means.

Abstract

Customer segmentation allows financial institutions to improve their business strategy, focusing on their target market through targeted marketing campaigns. Institutions, by not having techniques to segment their clients, find it difficult to reach their objectives, assuming greater risk in unknown market niches. The objective of this study is to recommend financial products to potential clients of a financial institution in Ecuador using the machine learning technique, k-means clustering. The base data are the clients of the financial institution of Ecuador with their characteristics as of December 2022; a transversal approach is used for the application of the model. The results show that the cluster technique recommends segmenting customers into ten groups, three of which are potential, from the point of view of the demand they have for any of the credit products, credit cards, savings, and time deposits.

Keywords: Customer segmentation, target market, strategy, machine learning, financial institution, data mining, k-means.

1. Capítulo I: Introducción

La minería de datos es el proceso automático de descubrir patrones y tendencias útiles en grandes volúmenes de datos, utilizando métodos de la machine learning capaces de transformar los datos en conocimiento.

Con la era digital y la incorporación de inteligencia artificial (máquinas capaces de replicar el razonamiento humano) en distintos sectores, áreas y procesos productivos de las empresas es importante aplicar técnicas de machine learning, a fin de tomar decisiones en base a la información que dispone la organización, y ser resilientes ante los cambios tecnológicos. Actualmente, la institución financiera objeto de análisis posee más de un millón de clientes, sin embargo, al no disponer de técnicas para segmentar a estos se le dificulta llegar a sus objetivos. De esta forma, asume mayor riesgo en nichos de mercados desconocidos al no disponer de información necesaria para segmentar a sus clientes e identificar a los potenciales.

Desde este punto de vista, se considera importante identificar segmentos de clientes, para recomendar productos de ahorro y/o crédito en base a sus características, lo cual permitirá mejorar la relación entre la empresa y el cliente, establecer estrategias de negocio personalizadas, pudiendo obtener aspectos positivos e impacto en la institución como: incremento de la cuota de mercado, incremento del rendimiento, mayor ventaja competitiva frente a otras instituciones del sector, y fidelización de sus clientes.

Es así que, apostar por técnicas de machine learning para la toma de decisiones, apuntará a que la institución se convierta en una entidad que asume riesgo de forma controlada. De esta forma, se plantean las siguientes preguntas de investigación: *¿Cuáles son las características sociodemográficas de los clientes potenciales de la institución financiera?; ¿Cómo los grupos de clientes homogéneos utilizan los productos de ahorro y crédito de la institución?*

El *objetivo general* del presente estudio es recomendar productos financieros a clientes potenciales de una institución financiera del Ecuador mediante la segmentación de clientes. Es pertinente mencionar que la minería de datos juega un papel importante en la determinación de patrones ocultos de la base de datos. En cuanto a los *objetivos específicos* se centra en utilizar técnicas de machine learning para la agrupación de clientes, y plantear estrategias de negocio personalizadas para los clientes de acuerdo al segmento al que pertenezcan.

En base al objetivo de la investigación se emplea la base de clientes de la entidad financiera corte diciembre del 2022. Se aplica la metodología CRISP-DM, Cross-Industry Standard Process for Data Mining, desarrollado por empresas europeas en 1999 (Dáderman & Rosander, 2018), el cual dispone de varias etapas para la minería de datos, desde la comprensión del negocio; tratamiento de la información; aplicación de métodos; evaluación y puesta en producción. En la fase de la aplicación de métodos se utiliza técnicas de machine learning no supervisado, método clúster k-means. Este método agrupa observaciones con características similares u homogéneas y diferentes a otros grupos de observaciones.

Los resultados muestran que la metodología clúster k-means, recomienda la segmentación de clientes en diez grupos, estos resultados son validados mediante los métodos: elbow, average silhouette y gap (Hung, Phan Duy; Thi, Nguyen; Duc, 2019), además se comprueba con diferentes muestras aleatorias, demostrando la robustez de los mismos. Respecto a las características de los grupos potenciales que demandan en mayor medida los productos de crédito, ahorro y depósitos a plazo fijo, destacan aquellos clientes con mayores niveles de ingresos, activos y patrimonio, así como personas en edades promedio de 44 años, que pertenecen al segmento independiente, entre otras características. Por su parte, las características de los clientes con mayores saldos en tarjetas de crédito son las personas con edades promedio de 36 años, con niveles de ingresos medios, así mismo registran niveles medios en sus activos y patrimonio. Es importante mencionar, que los resultados muestran las características sociodemográficas de los clientes, siendo una limitante, ya que el estudio podría enriquecerse con otras características como las psicográficas, de comportamiento, y geográfica.

El presente trabajo de titulación se organiza de la siguiente manera. En el capítulo II se presenta la revisión de la literatura. El capítulo III explica la metodología a utilizar. En el capítulo IV y V se expone los resultados y su respectivo análisis. Finalmente, en la sección VI se presentan las conclusiones y recomendaciones.

2. Capítulo II: Revisión de la literatura

En la presente sección se revisa las teorías sobre la segmentación de clientes, así como, el marco metodológico en el que se sostiene el presente estudio.

2.1 Revisión de literatura

Varias teorías sostienen que las técnicas de machine learning permiten mostrar patrones ocultos de las características de los clientes, y así mejorar la relación entre empresa-cliente mediante estrategias de negocios dirigidas. Existen varias formas de segmentar a los clientes, ya sea por sus características sociodemográficas, conductuales, de perspectiva, etc. A continuación, se presenta la revisión de la literatura relacionada al tema de análisis:

Características sociodemográficas:

Autores como (Jayant, Tikmani; Sudhanshu, Tiwari; Sujata, 2015) realizan una segmentación de clientes utilizando características sociodemográficas de sus clientes, así agrupan a clientes con patrones similares en función de atributos como la edad, intereses, hábitos, gastos, etc.

(Yefta, Christian; Oktaviani, 2022) realizan una segmentación de mercado para un startup utilizando el marco CRISP-DM y la técnica k-means, emplean datos demográficos relacionados con los intereses, comportamiento y antecedentes de clientes.

Características por utilización de productos:

Por otro lado, (Harrison, 2006) analiza la segmentación de clientes bajo la perspectiva del uso del producto, así señalan que la segmentación debe centrarse en: las percepciones; actitudes y motivaciones que tiene hacia los productos financieros. Segmenta en cuatro grupos a los clientes: i) inversionistas cautelosos; ii) acumuladores de capital; iii) financieramente confundidos, y iv) minimalistas apáticos.

En la misma línea, (Kansal et al., 2018) implementan la técnica de segmentación de clientes, a fin de identificar clientes VIP en una tienda minorista local, utilizando dos características: promedio de visitas; y cantidad de compras, donde aplican tres algoritmos de agrupamiento: k-means; agglomerative; meanshift. Como resultado obtienen siete clústeres etiquetados como: i) clientes descuidados; ii) cuidadosos; iii)

estándar; iv) objetivo; v) sensible; vi) compradores altos y visitantes frecuentes; vii) compradores altos y visitantes ocasionales, estos dos últimos considerados como VIP.

Por su parte (Yang et al., 2015), aplican el método de segmentación de clientes para un banco privado en China, a fin de identificar clientes actuales y potenciales de alto valor para el banco, utiliza variables de depósitos y productos financieros de alto valor. Como resultado obtienen tres grupos de clientes: i) core value cuya característica principal es la gran proporción de depósitos y compra de un producto financiero de alto valor; ii) clientes orientados a productos financieros, y iii) clientes orientados a depósitos.

Caracterización combinación de características:
Por su parte, (Sulekha, 2011) recomienda cuatro bases de información para la segmentación de clientes: geográfica; demográfica; psicográfica (variables de estilo de vida); conductual (comportamiento hacia los productos como lealtad, disposición a comprar).

Por el contrario, (Monil et al., 2020), destacan en su análisis que la combinación de algoritmos de clúster puede mejorar los resultados del análisis de agrupamiento mediante algoritmos individuales. Realiza una comparación de las diferentes técnicas utilizando cuatro técnicas: k-means; clúster jerárquico; Density-based spatial clustering of applications with noise (DBSCAN); y propagación de afinidad. Señala que, por velocidad de cómputo y tiempo de agrupamiento de los datos, k-means es mejor, mientras que por eficiencia en los resultados DBSCAN tiene mejores resultados. Por su parte para el manejo de datos dinámicos es menos eficiente el clúster jerárquico en relación a los otros métodos.

De las limitaciones más comunes que señalan en la literatura es la disponibilidad de información sobre variables conductuales y psicográficas de los clientes, mencionan que, para la obtención de estas, es necesario un levantamiento de información a través de entrevistas o encuestas.

2.2 Minería de datos

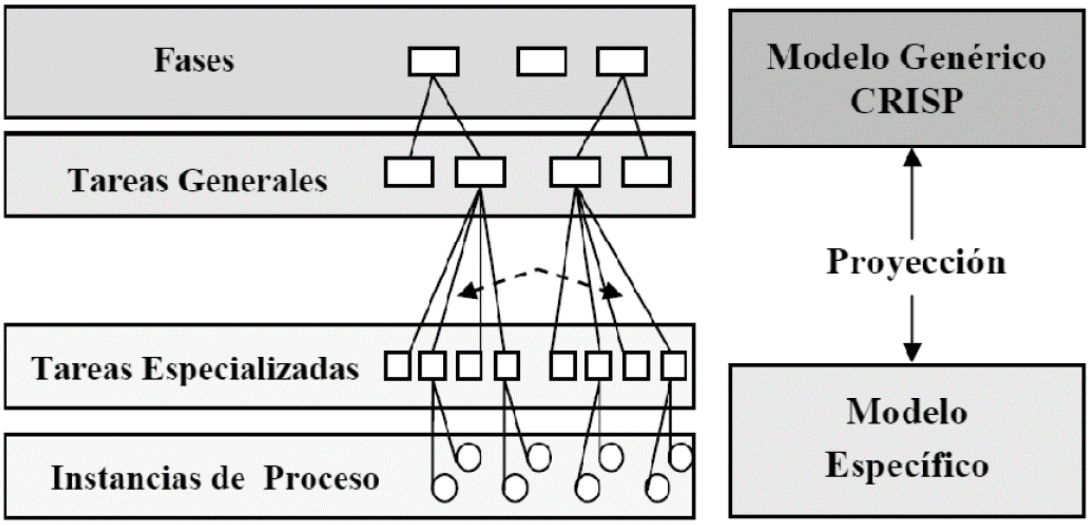
La minería de datos o data mining es el proceso automático de descubrir patrones y tendencias útiles en grandes volúmenes de datos, mediante un conjunto de técnicas y herramientas aplicadas para la exploración y análisis (Piatetsky-Shapito, 1991).

Es así que, diversas empresas han desarrollan una metodología estándar intersectorial para aplicar un proceso de minería de datos. SAS propone la metodología SEMMA (Sample, Explore, Modify, Model, Assess). En 1999 empresas europeas, NCR (Dinamarca), AG (Alemania), SPSS (Inglaterra) y OHRA (Holanda), desarrollaron la metodología CRISP-DM (Cross- Industry Standard Process for *Data Mining*). Estas dos metodologías son las principales utilizadas para un proyecto de *Data Mining* (Dáderman & Rosander, 2018).

2.2.1 CRISP-DM

CRISP-DM Cross Industry Standard Process for Data Mining es un proceso jerárquico de cuatro niveles que van desde lo general a lo específico: fases, tareas genéricas, tareas especializadas, e instancias de proceso; estableciendo el ciclo de vida de un proyecto de minería de datos, donde la salida de una fase es la entrada para la siguiente (Gallardo, 2010).

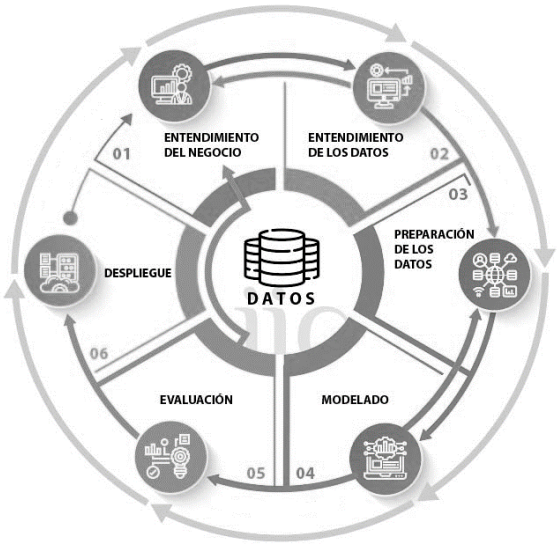
Ilustración 1 *Esquema de los cuatro niveles de CRISP-DM*



Fuente: (Gallardo, 2010).

El proceso jerárquico acoge un conjunto de seis tareas: entendimiento del negocio; entendimiento de los datos; preparación de los datos; modelo; evaluación; y despliegue. Además, la continuidad de las fases no es rígida, por lo que pueden ser interactivas, por ejemplo, se puede volver de la fase modelado a la fase de entendimiento de los datos, si dentro del proceso de minería de datos se requiere retroalimentar elementos o en caso de no estar encaminado al cumplimiento de objetivos.

Ilustración 2 *Ciclo de vida de un proyecto de minería de datos*



Fuente: (Gallardo, 2010).

Tabla 1 *Ciclo de vida de un proyecto de minería de datos*

Entendimiento del negocio	Entendimiento de los datos	Preparación de los datos	Modelado	Evaluación	Despliegue
Determine los objetivos	**Recolectar los datos iniciales**	**Dataset**	**Seleccionar técnica de modelado**	**Evaluar los resultados**	**Plan de implementación**
Background	Reporte de recolección de los datos	Descripción del Dataset	Técnica seleccionada	Valoración de los resultados	Plan de implementación
Objetivo del negocio	**Descripción de los datos**	**Seleccionar los datos**	Supuestos del modelo	Modelos aprobados	**Plan de monitoreo y mantención**
Criterios de éxito del negocio	Reporte de descripción de datos	Inclusión/exclusión de datos	**Generar el plan de prueba**	**Revisión del proceso**	Plan de monitorización y mantención
Valoración de la situación	**Exploración de los datos**	**Limpiar los datos**	Plan de pruebas	Revisión del proceso	**Informe final**
Inventario de recursos	Reporte de exploración de los datos	Reporte de calidad de los datos	**Construir el modelo**	**Determinar próximos pasos**	Informe final
Requisitos, supuestos y restricciones	**Verificar la calidad de los datos**	**Estructurar los datos**	Análisis de parámetros	Técnica modelada	Modelos aprobados
Riesgos y contingencias	Reporte de calidad de los datos	Derivación de atributos	Modelo	Listado de las posibles acciones	**Revisión del proyecto**
Terminología		Generación de registros	Descripción del modelo		Documentación
Costes y beneficios		**Integrar los datos**	**Evaluar el modelo**		Experiencias
Determinar los objetivos de DM		Unificación de datos	Evaluar el modelo		
Metas de Data Mining		Formato de los datos	Revisión de los parámetros		
Criterios de éxito de DM		Reporte de calidad de los datos			
Realizar el plan del proyecto					
Plan de proyecto					
Valoración inicial					

Fuente: (Gallardo, 2010).

Comprensión del negocio

En esta fase inicial es necesario conocer los requisitos y problemas desde una perspectiva empresarial, a fin de resolver mediante un proyecto de minería de datos. Se define los objetivos y requisitos del negocio, se evalúa la situación, a través de la definición de requisitos del problema, y se delimita los objetivos del proyecto de data mining en base a los objetivos empresariales (Ibidem).

Comprensión de los datos

En esta segunda fase de la metodología CRISP-DM trata sobre la compresión de los datos, inicialmente se debe realizar la recolección de datos; descripción de los datos; exploración, ésta a fin de definir una estructura de los datos, a través de pruebas y técnicas estadísticas; y la verificación de la calidad de los datos, a fin de identificar problemas en los datos y determinar la consistencia de estos (Ibidem).

Preparación de los datos

Esta fase incluye todas las actividades de preparación de los datos para su posterior modelamiento, dado que en función de la técnica de modelado los datos deberán ser preparados. Principalmente se realiza la selección de datos, limpieza, definición de relaciones entre las variables, establecimiento de atributos y transformación de variables.

Esta fase y la anterior son consideradas las que mayor tiempo consume, dado a la diversidad de técnicas que existen para aplicar. Respecto a la limpieza de los datos, se puede aplicar la normalización, discretización, imputación, entre otros (Ibidem).

Modelado

En la fase de modelado se seleccionan las técnicas adecuadas para el proyecto de minería de datos que permita cumplir los objetivos planteados, generalmente existen varias técnicas que se pueden realizar en un proyecto. Posteriormente se debe generar un plan de pruebas para determinar la validez del modelo, se puede aplicar el particionamiento de los datos de entrenamiento (train) y evaluación (test).

En la etapa de la construcción del modelo se ejecuta la técnica seleccionada donde se debe establecer los mejores parámetros a utilizar a fin de obtener los resultados más óptimos. Finalmente, se realiza la evaluación del modelo, esto dependerá de la técnica que se emplee para establecer los criterios de evaluación, entre los más populares se encuentra la razón de error como medida de calidad del modelo (Ibidem).

Evaluación

En la fase de evaluación del modelo se debe considerar los criterios de acuerdo a la técnica de modelamiento que se esté empleando, se pueden utilizar diversas herramientas para la interpretación de los resultados, y establecer si el modelo es válido, o si es necesario repetir algún proceso anterior a fin de mejorar los resultados.

Las etapas que se incluye en esta fase son i) la *evaluación de los resultados*, a fin de verificar si responden a los objetivos planteados del negocio y del proyecto de minería de datos; ii) **proceso de revisión** que refiere a una evaluación integral del proceso CRISP-DM y si es necesario mejorar alguna fase; y *determinación de próximos pasos,* ya sea para pasar a la siguiente fase o determinar nuevos proyectos de minería de datos (Ibidem).

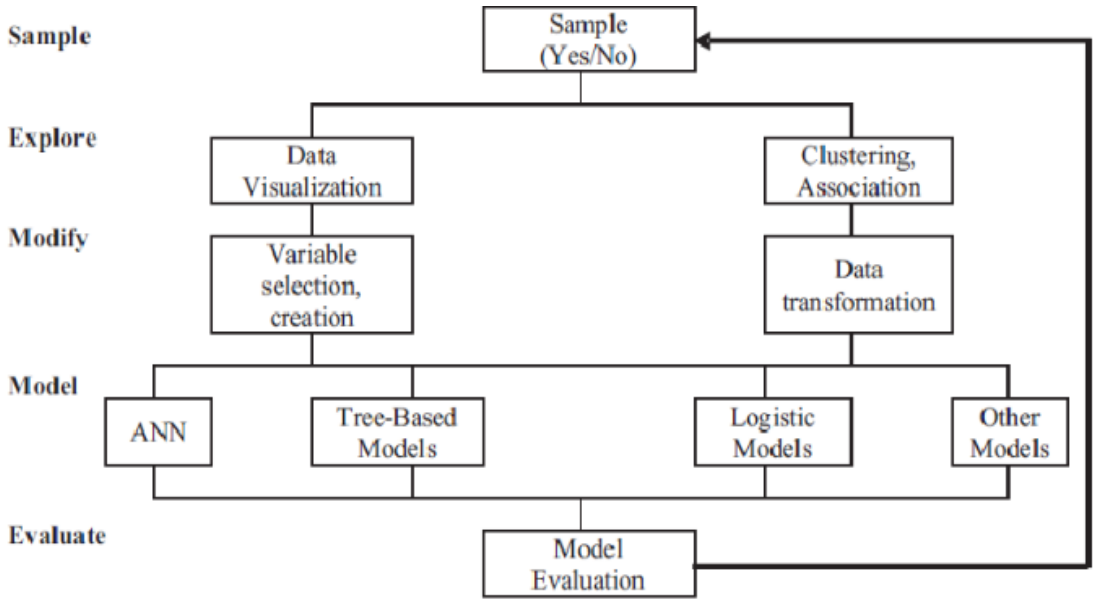
Despliegue

Una vez concluido las etapas anteriores, donde el modelo haya sido validado, se establecen acciones o estrategias del negocio en base a los resultados obtenidos, así como la puesta en marcha en producción para su monitorización y mantenimiento del modelo, el cual generará una retroalimentación para conocer si el modelo sigue siendo apropiado. Finalmente, se documenta los resultados del proyecto de minería de datos, donde puede incluir experiencia positivas y negativas del proyecto (Ibidem).

2.2.2 SEMMA

La metodología CRISP-DM ha sido fuente de inspiración de otras técnicas de minería de datos como SEMMA (Sample, Explore, Modify, Model, Assess), propuesto por SAS Institute, el cual define como un *proceso de selección, exploración y modelamiento de grandes cantidades de datos para descubrir patrones de negocios desconocidos.* La siguiente figura muestra el proceso de minería de datos desde el enfoque de la metodología SEMMA (Dáderman & Rosander, 2018).

Ilustración 3 *Metodología SEMMA*



Fuente: (Dáderman & Rosander, 2018)

La metodología SEMMA es un proceso que permite extraer información útil de forma ordenada y eficiente, se divide en cinco fases interrelacionadas: muestra, exploración, modificación, modelamiento, y evaluación, convirtiendo el proceso de minería de datos en un proceso iterativo en interactivo al igual que la metodología del CRISP-DM.

Muestra:

La metodología SEMMA inicia con la extracción de la población muestral representativa, la muestra determinada deber ser lo suficiente grande que contenga información necesaria para el modelo, y suficiente pequeño para que el procesamiento sea fácil. En esta fase se incluye la participación de los datos en entrenamiento, validación y muestras de prueba, así como la identificación de variables (Ibidem).

Explorar:

Posteriormente la metodología SEMMA establece la exploración de los datos, mediante la aplicación de herramientas estadísticas y de visualización. Se lleva a cabo el análisis univariado y multivariado, y determinar relaciones entre las variables, y establecer variables explicativas como entradas del modelo (Ibidem).

Modificar:

Este paso hace referencia a la manipulación de los datos, en función a la exploración de estos. En esta etapa se evalúa la calidad de la información, se aplican la limpieza de

los datos, se establecen nuevos atributos, transformación de variables, en definitiva, se establece un formato adecuado de los datos para que puedan ser modelados (Ibidem).

Modelo:

En la fase del modelado se aplican las técnicas de minería de datos acorde al objetivo del proyecto de minería de datos, donde se emplea los datos limpios de los pasos anteriores, a fin de obtener los resultados confiables que respondan a los objetivos planteados, y que permitan predecir resultados de un nuevo set de datos (Ibidem).

Evaluar:

Esta última fase consiste en la valoración de los resultados, a fin de determinar el nivel de confiabilidad de los resultados. Las técnicas utilizadas en esta fase dependerán del modelo que se aplique en la fase anterior. Para la evaluación y rendimiento de los modelos se utiliza las muestras de validación y prueba, a fin de determinar si el modelo permite predecir buenos y fiables resultados (Ibidem).

2.2.3 Comparación SEMMA y CRISP-DM

La metodología SEMMA y CRISP-DM definen el proceso de un proyecto de minería de datos, a través de fases interrelacionadas, donde los procesos son iterativos e interactivos. La metodología SEMMA se centra más en el proceso de comprensión y análisis de los datos, mientras que la metodología CRISP-DM se centra tanto en los objetivos empresariales como en los objetivos del proyecto de minería de datos.

Tabla 2 *Comparación metodología SEMMA y CRISP-DM*

SEMMA y CRISP-DM son metodologías orientadas a procesos de minería de datos		
	CRISP-DM	**SEMMA**
Enfoque	Encaminado a los objetivos empresariales	Encaminado al desarrollo de proceso de minería de datos
Uso	Metodología abierta y gratuita	Ligado a productos SAS
Metodología	Metodología de gestión de proyectos	Metodología aún no definida
Complejidad	Fácil de entender y aplicar, cuenta con una curva de adaptabilidad muy amplia para cualquier desarrollador	Simple y ágil, sus fases están enfocadas a desarrollo ágil
Siglas	Cross-Industry Standard Process	Sample, explore, modify, model an access

En el presente proyecto se emplea la metodología CRISP-DM donde se considera a la perspectiva empresarial como la base para el desarrollo del proyecto de minería de datos.

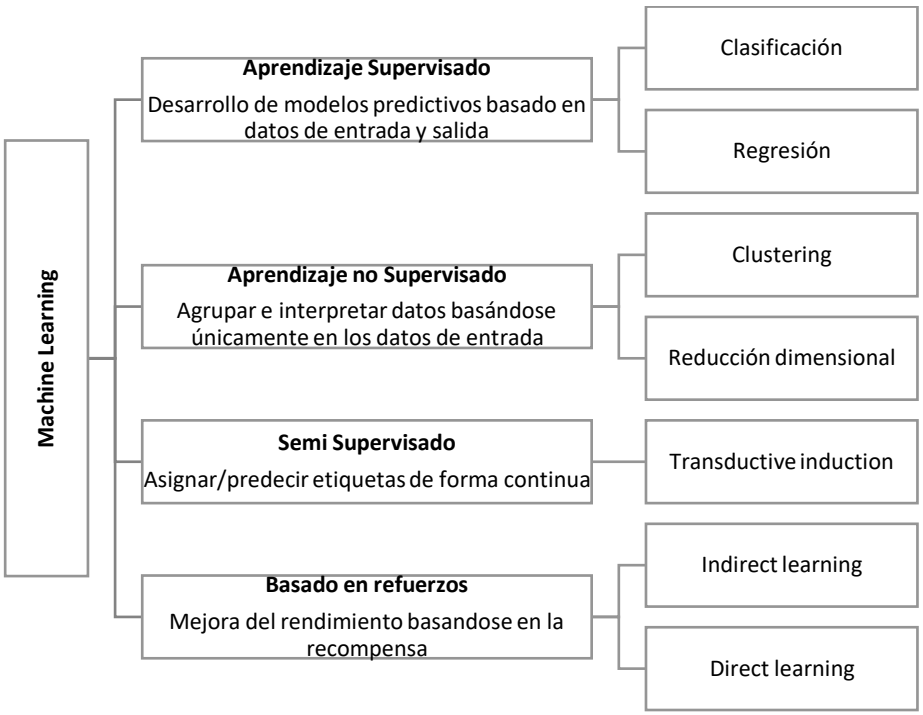
2.3 Inteligencia Artificial

La inteligencia artificial estudia el diseño y construcción de máquinas capaces de replicar el razonamiento humano y, por tanto, de resolver tareas para las que no se las ha programado previamente. Los sistemas de inteligencia artificial necesitan adquirir su propio conocimiento, esta capacidad se denomina machine learning (Gollapudi, 2016).

2.3.1 Machine Learning

(Samuel, 1959) define al machine learning como el campo de estudio que brinda a las máquinas la habilidad de aprender sin ser explícitamente programadas por los humanos. Dentro del machine learning existen los siguientes tipos de aprendizaje:

Ilustración 4 *Tipos de aprendizaje machine learning*



Fuente: (Gollapudi, 2016)

2.3.2 Aprendizaje no supervisado

En el aprendizaje no supervisado no se requiere de una variable objetivo para predecir resultados, ya que busca relacionar elementos similares entre sí y distintos del resto. Entre las técnicas del aprendizaje no supervisado se encuentran: i) clustering; ii) asociación; iii) reducción de dimensionalidad. Para el proyecto en estudio se emplea la metodología clúster que se detallan en el capítulo III.

2.4 Métodos de aprendizaje no supervisado

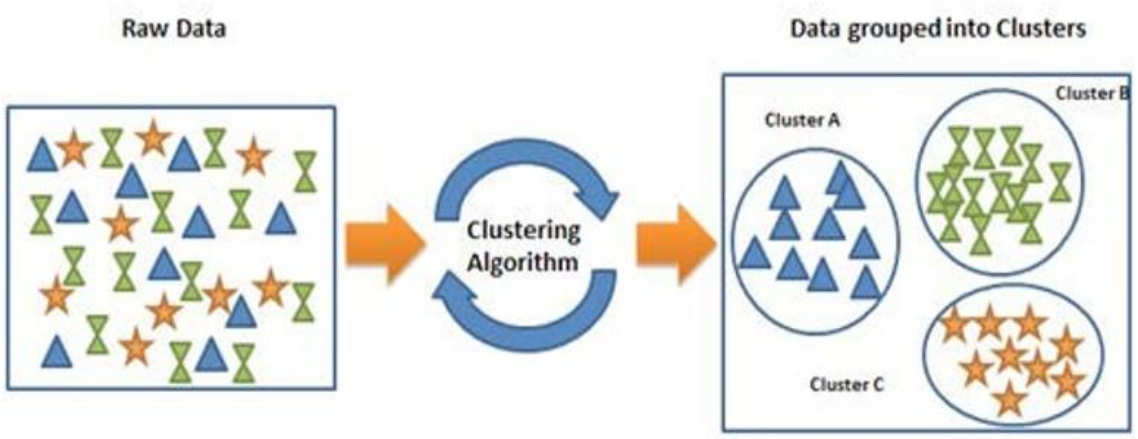
A continuación, se explica las metodologías de aprendizaje no supervisado, clúster, que es la base para el modelamiento del presente estudio.

2.4.1 Metodología clúster

El clúster es una técnica de clasificación, que analiza un conjunto de relaciones interdependientes, no diferencia entre variable explicativa o predictoras (método no supervisado). El objetivo es agrupar a observaciones con más similitud entre ellas y distintas a las observaciones de otros grupos (Naresh K. Malhotra, 2008). A priori no se conoce a que grupo pertenece realmente cada observación, esta característica es la principal diferencia entre los métodos de clustering de los métodos de clasificación, donde se conoce la clasificación real.

Algunas disciplinas donde se puede aplicar las técnicas de agrupación son: segmentación de clientes, segmentación de perfiles de tarjetas de crédito, detección de fraudes, entre otros. La siguiente figura representa el proceso de agrupamiento.

Ilustración 5 *Proceso de agrupamiento*



Fuente: (Gollapudi, 2016)

2.4.1.1 Medidas de distancia

Para llevar a cabo las agrupaciones se requiere una medida que permita evaluar la semejanza y diferencia de las observaciones, comúnmente se mide la semejanza por la distancia entre las observaciones, así, aquellas observaciones con una menor distancia serán más similares que otros con mayor distancia. A continuación, se presentan algunas de las medidas de distancia más utilizadas:

- ***Distancia euclidiana:*** se calcula empleando el teorema de Pitágoras, donde la distancia se presenta mediante la siguiente ecuación, que es la raíz cuadrada de la sumatoria de diferencias de los valores elevadas al cuadrado de cada variable.

$$\boldsymbol{d_{euc}(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}} \qquad (1)$$

- ***Distancia Manhattan:*** define la distancia como la sumatoria de las diferencias absolutas de los valores para cada variable o dimensión.

$$\boldsymbol{d_{man}(p,q) = \sum_{i=1}^{n} |(p_i - q_i)|} \qquad (2)$$

- ***Correlación:*** la medida de distancia de la correlación mide la relación entre dos observaciones. Se puede utilizar la correlación de Pearson, Spearman, Kendall, entre otros.

$$\boldsymbol{d_{cor}(p,q) = 1 - correlación\ (p,q)} \qquad (3)$$

2.4.1.2 Escalado de variables

Generalmente las variables se encuentran en escalas diferentes que podría afectar al agrupamiento, dando mayor peso a aquellas variables que tengan una mayor escala, por lo que es necesario estandarizarlas, a través del escalamiento de variables, a fin de evitar la influencia de las unidades de medida.

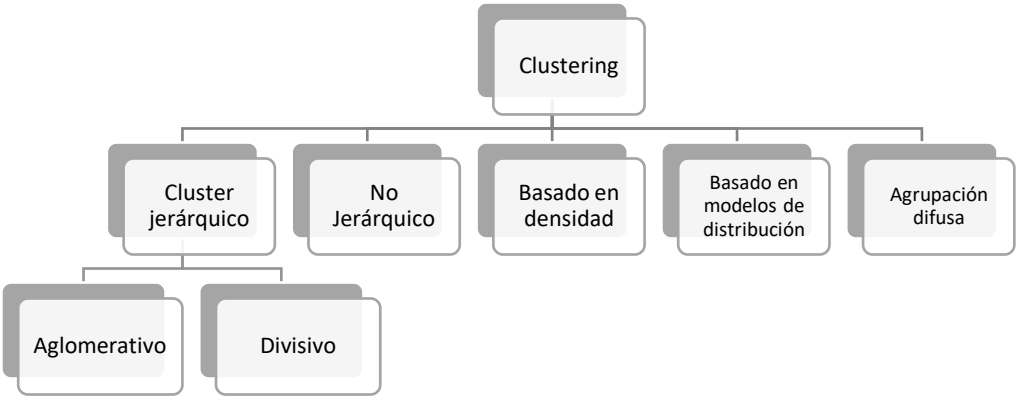
El objetivo es escalar y centrar las variables antes de calcular la matriz de distancias, para que tengan media 0 y desviación estándar 1, esto asegura que las variables tengan el mismo peso al momento de realizar el agrupamiento. La siguiente fórmula representa un tipo de escalamiento, mediante la estandarización de las variables.

$$\frac{x_i - mean(x)}{sd(x)} \qquad (4)$$

2.4.1.3 Tipos de agrupación:

La figura muestra los tipos de agrupación, comúnmente se encuentran los jerárquicos, y no jerárquicos.

Ilustración 6 *Tipos de agrupación*



Fuente: (Gollapudi, 2016)

Agrupación basada en densidad: conecta las áreas con alta densidad en grupos, y las diferencia de las de baja densidad, realiza distribuciones de forma arbitraria (Gollapudi, 2016).

Agrupación basada en modelos de distribución: los datos se agrupan en función de la probabilidad de que un conjunto de datos pertenezca a una distribución particular, generalmente la distribución gaussiana (Gollapudi, 2016).

Agrupación difusa: es un método flexible, donde cada observación puede pertenecer a varios clústeres, es decir, cada observación tiene un coeficiente o grado de pertenencia de cada uno de los clústeres donde se encuentra agrupado (Gollapudi, 2016) .

Clúster Jerárquico: como su nombre lo indica, el agrupamiento se realiza mediante una jerarquía de mayor a menor número de grupos o viceversa, puede ser aglomerativo o divisivo.

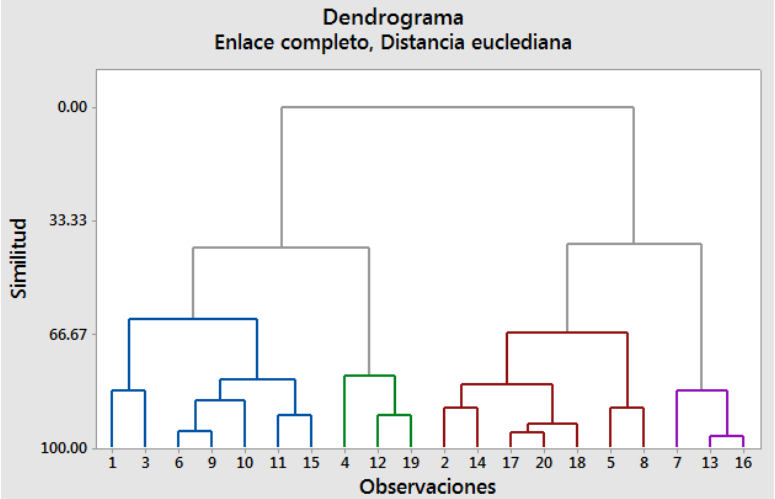
- ***Aglomerativo****:* inicia con todas las observaciones separadas, de este modo cada observación va formando un clúster individual y se van combinando a medida que la estructura crece hasta converger en uno solo, es decir, utiliza el enfoque de abajo hacia arriba o de lo micro a lo macro.
- ***Divisivo****:* inicia con todas las observaciones agrupadas en un mismo clúster y posteriormente se va particionando, es este caso, emplea el enfoque de arriba hacia abajo o de lo macro a lo micro.

Este tipo de algoritmos no requieren que el analista especifique de antemano el número

de clústeres. El número de clúster se puede obtener mediante una representación de un árbol jerárquico, conocido como dendograma.

- **Dendograma:** el dendograma es una representación gráfica de jerarquía, se lee de izquierda a derecha, donde las líneas verticales representan los conglomerados que se unen.

Ilustración 7 *Dendograma*



Fuente: (Gollapudi, 2016)

Clúster no jerárquico: este tipo de algoritmos requieren que el analista defina de antemano el número de agrupaciones k que espera, a este método se lo conoce como k-means.

El método k-means, k se refiere a la cantidad de clústeres (agrupaciones) en los que se van agrupar los datos, y means al promedio del conjunto de datos para encontrar el centroide. Este algoritmo de agrupamiento depende del centroide, donde cada punto de datos se asigna al centroide más cercado formando los k grupos. De ahí, que el método de k-means agrupa el conjunto de datos en k mejores clústeres, donde el mejor clúster es aquel cuya varianza es mínima.

Pasos del algoritmo (K-means)

1. Se debe especificar a priori el número k de clústeres, para ellos se puede emplear el método de elbow (codo), método de average silhouette o gap, que se explicará más adelante.
2. El algoritmo k-means selecciona aleatoriamente k observaciones del conjunto de datos como centroides de inicio.

3. Posteriormente se asigna a cada una de las observaciones al centroide que se encuentra más cerca.
4. Para los k clústeres generados en el paso anterior se vuelve a calcular su centroide.
5. Finalmente, el algoritmo repite los dos últimos pasos, a fin de que las asignaciones sean robustas y permita alcanzar el número máximo de iteraciones prestablecido.

Cabe señalar que los procedimientos no jerárquicos son más rápidos en comparación con los jerárquicos, siendo una ventaja al aplicar la técnica a un número de grande de observaciones. A continuación, se muestra una representación gráfica de la agrupación de observación, mediante k-means.

Ilustración 8 *Agrupación clúster*

Fuente: (Amat, 2017)

2.4.1.4 Número óptimo de clústeres

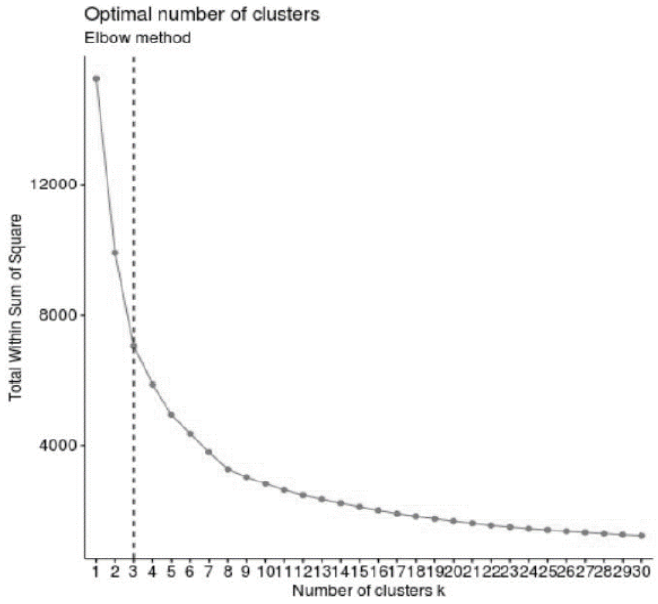
En número de clúster es un tema importante dentro del análisis, entre algunos lineamientos para decidir su número están: la teoría o experiencia, mientras que en el método jerárquico se suele utilizar la técnica del dendograma, mientras que para el método no jerárquico se mide mediante la proporción de la varianza, y se puede emplear el método de elbow (codo) o método de average silhouette.

Método elbow (codo):

El método Elbow permite escoger aquel número de clúster donde minimice la varianza total del clúster, es decir, calcula la varianza dentro de cada clúster simulado en función al número asignado previamente (hiperparámetro) y permite al analista escoger el número de clúster óptimo en función de aquel valor donde añadir mayor número de

clústeres apenas consigue minimizar la varianza (Hung, Phan Duy; Thi, Nguyen; Duc, 2019).

Ilustración 9 *Número óptimo de conglomerado método codo*

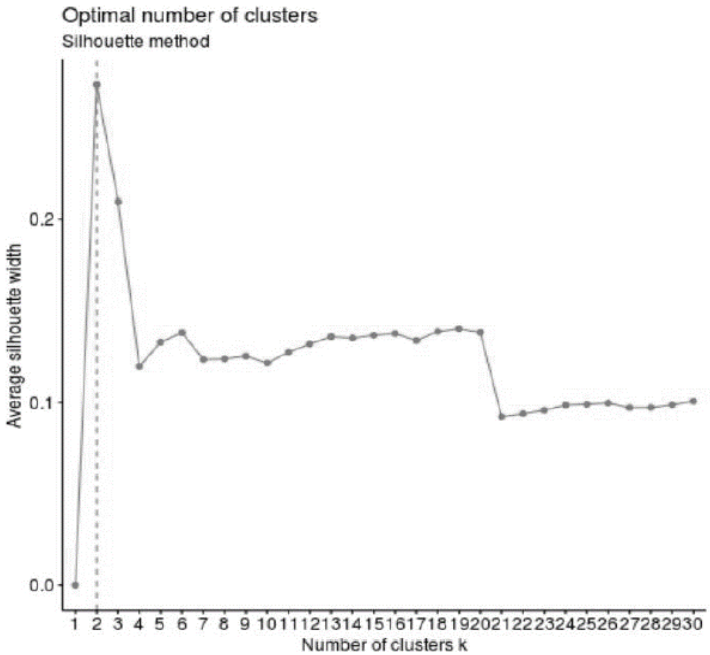


Fuente: (Hung, Phan Duy; Thi, Nguyen; Duc, 2019)

Método average silhouette:

Este método es similar al método elbow, la diferencia radica en que máxima el índice de silueta (average silhouette). Su valor se encuentra en un rango de -1 y 1, donde 1 representa que se ha asignado correctamente la observación al clúster (Ibidem).

Ilustración 10 *Número óptimo de conglomerado método de Silhouette*

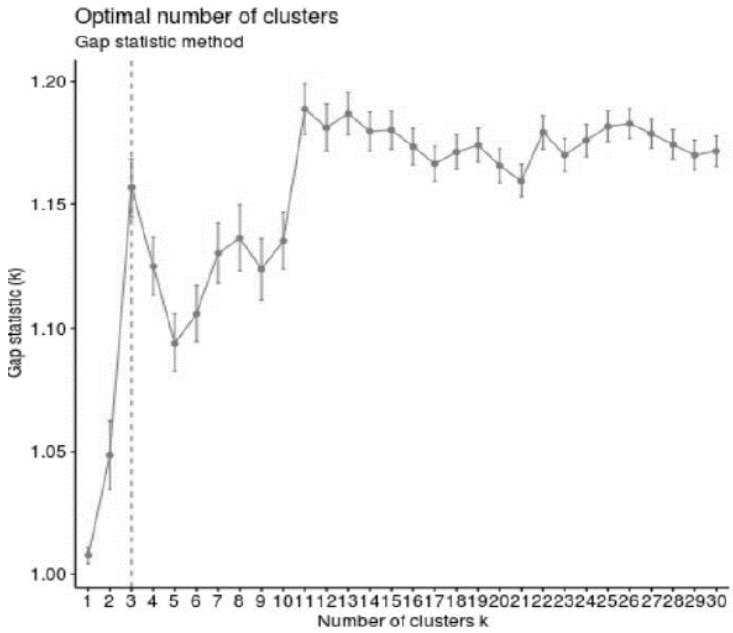


Fuente: (Hung, Phan Duy; Thi, Nguyen; Duc, 2019)

Método GAP:

Este método publicado por (Tibshirani et al., 2001), compara la variación intra-clúster total para diferentes valores de k con sus valores esperados bajo una distribución uniforme de referencia. El número óptimo de k, es aquel donde se consigue maximizar el valor del gap.

Ilustración 11 *Número óptimo de conglomerado método GAP*



Fuente: (Hung, Phan Duy; Thi, Nguyen; Duc, 2019)

2.4.1.5 Validación del clúster:

En la metodología de clúster no es fácil validar los resultados, debido a que no se conoce la agrupación real de los valores. Una forma de validar los resultados es aplicar el mismo método con diferentes muestras de la población, a fin de verificar si la asignación de los clústeres es similar, además, se pueden emplear distintas medidas de distancia.

3. Capítulo III: Metodología

En la presente sección se detalla la metodología empleada, en base a los descrito anteriormente, así como las fuentes de información y las herramientas empleadas.

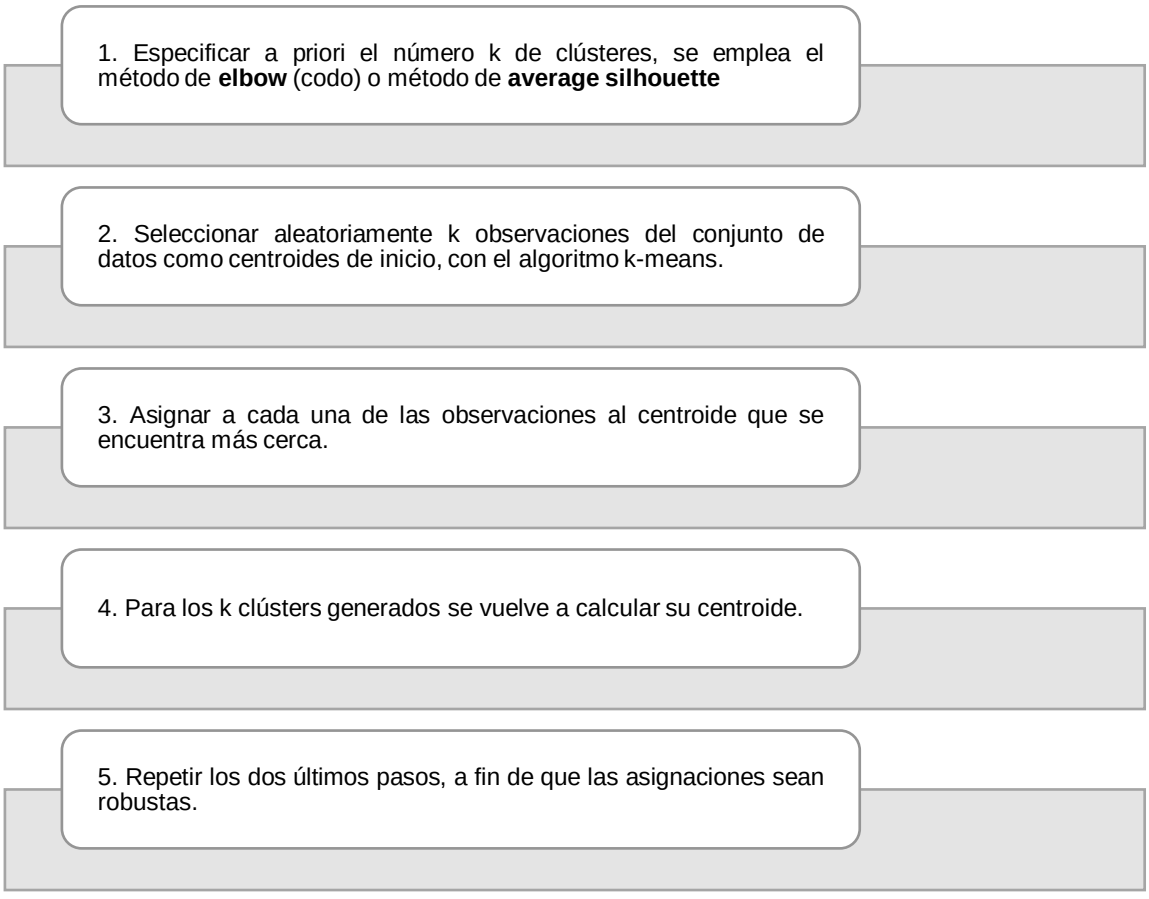
3.1 Métodos de aprendizaje no supervisado

3.1.1 Metodología clúster

Con el objetivo de agrupar observaciones con más similitud entre ellas y distintas a las observaciones de otros grupos, se emplea el método de clúster no jerárquico.

Al aplicar el método de **Clúster no jerárquico,** se requiere definir de antemano el número de agrupaciones k que espera, conocido como k-means. El k-means, al ser un algoritmo de agrupamiento depende del centroide, donde cada punto de datos se asigna al centroide más cercado formando los k grupos, el mejor clúster es aquel cuya varianza es mínima. A continuación, se definen los pasos a seguir en la etapa del modelamiento.

Ilustración 12 *Pasos del algoritmo (K-means)*

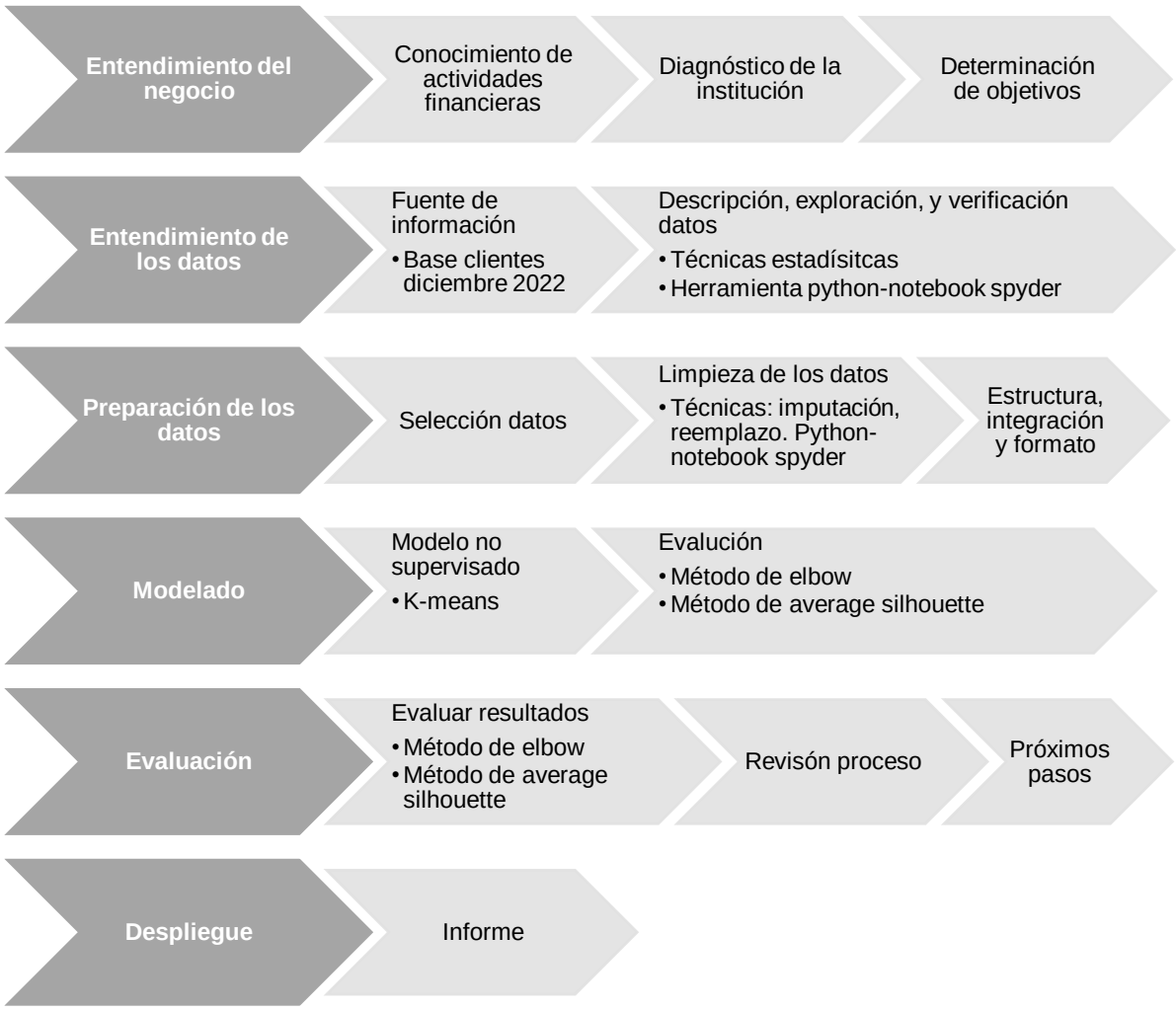


Fuente: (Naresh K. Malhotra, 2008)

Finalmente, para la **validación de los resultados**, al ser un aprendizaje no supervisado se aplica el método elbow, average silhouette y gap, detallado en el capítulo anterior.

Al emplear la metodología de clúster, no es fácil validar los resultados, debido a que no se conoce la agrupación real de los valores. Por lo que se aplica el mismo método con diferentes muestras aleatorias de la población[1], además se emplea el método elbow, average silhouette y gap, a fin de validar si el número de clústeres son óptimos.

Ilustración 13 *Proceso de desarrollo modelo CRISP-DM*



3.2 Fuente de información

La fuente de información del presente estudio es la base de clientes de la institución financiera del período diciembre del 2022. La base de datos consta de clientes con

[1] Muestra aleatoria: es una técnica de selección de los elementos de una población donde cada uno tiene la misma probabilidad de ser elegidos.

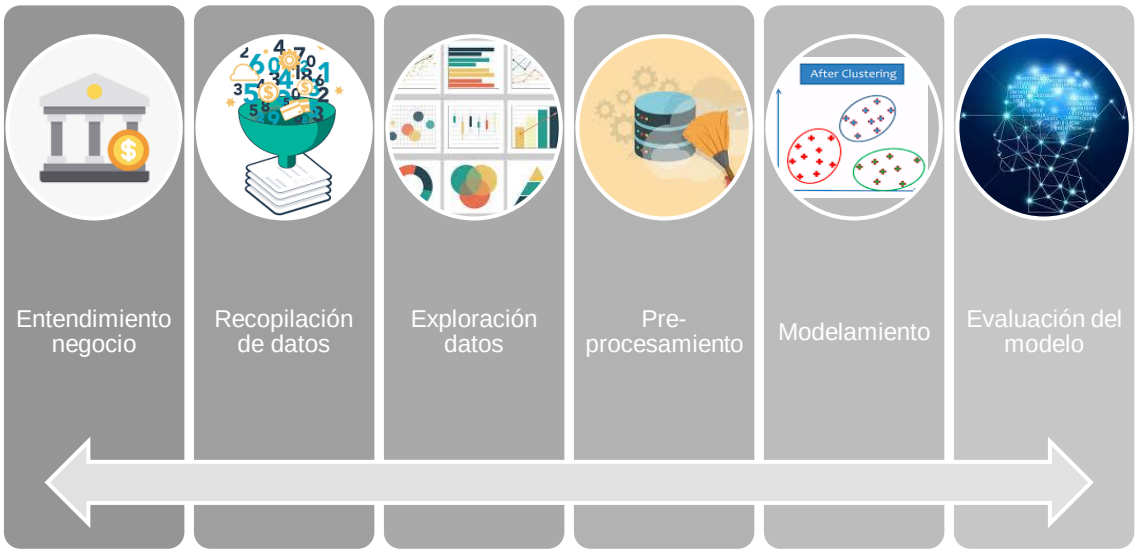
personalidad jurídica y natural, de acuerdo a los objetivos el análisis se centra en el segundo.

El método de investigación es de corte transversal utilizando metodologías de machine learning. Para ello se emplea la biblioteca de scikit learn de Python para el análisis descriptivo y aplicación del método de inteligencia artificial.

4. Capítulo IV: Resultados

En la presente sección se presentan los resultados obtenidos de la Metodología CRISP-DM aplicando técnicas de machine learning como es el análisis de clúster kmeans.

Ilustración 14 *Proceso de desarrollo resultados modelo CRISP-DM*



4.1 Comprensión del negocio

4.1.1 Determinar el Objetivo del Negocio

Contexto

Con la era digital y la incorporación de inteligencia artificial en distintos sectores, áreas y procesos productivos de las empresas, es importante aplicar técnicas de machine learning, a fin de tomar decisiones en base la información que dispone la organización, y ser resilientes ante los cambios tecnológicos.

La institución de análisis al no disponer de técnicas para segmentar a sus clientes se le dificulta llegar a sus objetivos, asumiendo mayor riesgo en nichos de mercados desconocidos al no disponer de información acerca de las características de los clientes. Actualmente la institución financiera registra una participación en el mercado ecuatoriano del al 4,31% en captaciones y 3,60% en colocaciones a nivel nacional. [2] En la siguiente tabla se presenta los objetivos del negocio y los criterios de éxito.

[2] Participación de captaciones y colocaciones de la institución financiera frente a los bancos, mutualistas y cooperativas del segmento 1, 2 y 3.

Tabla 3 *Objetivo del negocio y criterios de éxito*

Objetivo del Negocio	Incrementar la cuota de mercado en captaciones y colocaciones a nivel nacional. Recuperar el mercado perdido. Retener y fidelizar a los clientes
Criterios de éxito	Incrementar la participación en las captaciones a nivel nacional en un 7%, y un 6% en colocaciones.

4.1.2 Evaluación de la situación actual

Inventario de recursos

La institución financiera cuenta con un departamento de tecnología y un área de producción que se encarga de la recopilación y almacenamiento de la información.

A continuación, se presenta los requerimientos, suposiciones que se establecen para el tratamiento de la información y las restricciones a los cuales se podría enfrentar al momento de ejecutar el proceso de minería de datos.

Tabla 4 *Requerimientos, suposiciones y restricciones*

Requisitos	Se requiere disponer de programas de procesamiento de datos, además, de la autorización por parte de la entidad para utilizar las bases de datos de los clientes.
Suposiciones	La información es verídica. Los atributos son completos. La muestra de información es representativa.
Restricciones	Políticas gubernamentales como la ley de protección de datos para el acceso a la información. No disponibilidad de información completa de los clientes.

Riesgos y contingencias

En el transcurso del proceso se puede encontrar con varios riesgos que se pueden materializar, para lo cual se plantean contingentes, a fin de evitar y mitigar las amenazas.

Tabla 5 *Riesgos y Contingentes*

Riesgos	Dificultades de cumplimiento del cronograma por inconvenientes en el tratamiento de los datos. Obtener mala calidad de información. Ataques informáticos a las bases de datos.
Contingencias	Construcción de un cronograma flexible con holguras para garantizar la validación de convenios. Validar la información proporcionada mediante técnicas estadísticas. Respaldo de información en la nube

Terminología

- *Segmentación clientes:* proceso mediante el cual se agrupan clientes con características similares en un grupo y diferentes características a las de otro grupo.
- *Mercado objetivo:* son los clientes potenciales identificados por una entidad
- *Estrategia:* es un proceso mediante el cual se planifican actividades para alcanzar un objetivo.
- *Machine learning:* es una disciplina dentro de la inteligencia artificial, el cual crea sistemas para el aprendizaje automático mediante patrones.
- *Institución financiera:* es una entidad cuya actividad principal es la intermediación financiera, es decir captar recursos para colocarlos obteniendo un margen de rendimiento.
- *Minería de datos:* la minería de datos es la exploración de grandes cantidades de datos.
- *K-means:* es una técnica de machine learning, que agrupa objetos en grupos basándose en las características de los datos.
- *Información:* conjunto de datos organizados.

Costos y beneficios

Respecto al costo y beneficio del proyecto de minería de datos se plantean los siguientes:

Tabla 6 *Costos y beneficios*

Costos	Los costos están cubiertos por la entidad financiera, respecto a la disponibilidad de datos, y a la ejecución de estrategias que se establezcan a raíz del análisis del presente estudio. Respecto a la ejecución y desarrollo del proyecto de minería de datos está cubierto por los recursos del estudiante.
Beneficios	Permitirá incrementar la cuota de mercado de la institución financiera desde la perspectiva de las captaciones y colocaciones.

4.1.3 Determinar los objetivos de la minería de datos

En base al diagnóstico realizado de la institución y el entendimiento del negocio, se plantean los siguientes objetivos y criterios de éxito.

Tabla 7 *Objetivos minería de datos*

Objetivo de minería de datos	Aplicar técnicas de machine learning para la segmentación de clientes. Recomendar productos financieros a clientes potenciales a la entidad financiera.
Criterios de éxito de minería de datos	Técnicas de minería de datos que permita identificar atributos significativos para seleccionar un modelo óptimo. Modelo con un buen ajuste de precisión.

4.1.4 Producir el Plan de Proyecto

Para el cumplimiento de las actividades se construye el siguiente plan de proyecto

Tabla 8 *Plan de proyecto*

Fase	**N° semanas**													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Comprensión del Negocio	x	x												
Comprensión de datos			x	x										
Preparación de datos					x	x	x							
Modelado								x	x					
Evaluación										x	x			
Despliegue												x		
Presentación de resultados													x	x

Evaluación inicial de herramientas y técnicas

Las herramientas a utilizar en el presente proyecto es la biblioteca de scikit learn de Python para el análisis descriptivo y aplicación del método de inteligencia artificial.

4.2 Comprensión de los datos

4.2.1 Recolectar datos iniciales

La fuente de información de este estudio es la base de clientes de la institución financiera del período diciembre del 2022. La base de datos consta de clientes con personalidad jurídica y natural, de acuerdo a los objetivos el análisis se centra en el segundo. Con un total de 124,682 observaciones con 22 variables.

4.2.2 Descripción de los datos

A continuación, se presenta el conjunto de variables de la base de datos. Los datos corresponden a las personas naturales (clientes de la institución financiera). La mayoría de tipo de valores son de origen float y string, con 124,682 filas y 22 columnas. De forma general se observa que existen valores nulos, como en la variable de instrucción, número de cargas familiares, entre otros. Este conjunto de variables posteriormente será seleccionado para su tratamiento y construcción de nuevas variables, así como en la inclusión para el modelo de machine learning.

Tabla 9 *Descripción de la base de datos*

Variable	Detalle	Número de registros	Tipo
fechaapertura	Fecha de apertura de la cuenta de ahorro	2,523,644.00	Date
fechanacimiento	Fecha de nacimiento	2,523,644.00	Date
sector	Sector de residencia del cliente	123,579.00	String
instrucción	Nivel de Instrucción del cliente	124,243.00	String
estadocivil	Estado civil	124,682.00	String
numerocargasfamiliares	Número de cargas familiares	123,371.00	Int
profesion	Profesión	124,225.00	String
ingresostotales	Ingresos totales	124,682.00	Float
gastosfamiliares	Gastos familiares	124,682.00	Float
activostotales	Activos totales	124,682.00	Float
pasivostotales	Pasivos totales	124,682.00	Float
patrimonio	Patrimonio	124,682.00	Float
numeroinmuebles	Número de inmuebles	31,735.00	Int
numeroterrenos	Número de terrenos	4,796.00	Int
numerovehiculos	Número de vehículos	19,660.00	Int
numerocreditosconcedidos	Número de créditos concedidos	29,997.00	Int
aniostrabajo	Años de trabajo	111,796.00	Float
segmento_economico	Segmento económico	112,136.00	String
saldoahorro	Saldo de ahorro	122,942.00	Float
saldodpf	Saldo de depósitos a plazo	6,882.00	Float
saldocreditos	Saldo de crédito	18,288.00	Float
saldotc	Saldo de tarjeta de crédito	23,269.00	Float

4.2.3 Exploración de los datos

Análisis descriptivo de los datos:

En las dos siguientes tablas se muestra el análisis descriptivo de las variables numéricas y cualitativas de la base de datos. A priori se observa que existen el 67% de variables numéricas con valores nulos (10 de 15). Además, existen variables como: "activostotales" y "pasivostotales" que registran un alto valor en su desviación estándar, lo cual sería un indicativo de valores atípicos, para el primer caso se observa un promedio de $50,855, mientras que su mediana es de $7,500, y su valor máximo de $800,062,300. Por su parte la variable número de cargas familiares representa un valor máximo de 30 cargas cuando su promedio es de 1 carga familiar.

Tabla 10 *Descriptivos variables numéricas*

Variables	count	mean	std	min	25%	50%	75%	max
numerocargasfamiliares	123,371	1	1	0	0	0	1	30
ingresostotales	124,682	1,816	42,869	0	450	733	1,500	10,512,624
gastosfamiliares	124,682	1,115	42,263	0	159	290	608	10,412,212
activostotales	124,682	50,855	2,312,298	0	3,300	7,500	39,399	800,062,300
pasivostotales	124,682	4,562	23,103	-205	0	0	984	3,108,952
patrimonio	124,682	46,292	2,311,714	-138,473	3,000	6,200	33,300	800,046,649
numeroinmuebles	31,735	1	0	1	1	1	1	9
numeroterrenos	4,796	1	1	1	1	1	1	9
numerovehiculos	19,660	1	1	1	1	1	1	10
numerocreditosconcedidos	29,997	2	1	1	1	1	2	16
aniostrabajo	111,796	10	8	-45	4	8	13	112
saldoahorro	122,942	466	3,037	0	0	4	73	462,433
saldodpf	6,882	15,097	26,407	100	2,172	7,000	18,000	600,000
saldocreditos	18,288	11,814	20,153	1	2,181	5,286	13,334	493,545
saldotc	23,269	869	1,378	-876	20	499	1,072	38,739

Respecto a las variables cualitativas, de igual forma se observa que existen variables con valores nulos 4 de 5, que representa el 80%, además, se muestra el top o la moda de cada variable. La columna denominada "unique" representa el número de categorías de cada variable, es así que para la variable sector indica que existen 117,027 categorías, lo cual es un número muy amplio para la selección dentro del modelo.

Tabla 11 *Descriptivos variables cualitativas*

Variables	count	unique	top	freq
sector	123,579.00	117,027.00	independiente	75.00
instrucción	124,243.00	7.00	secundaria	63,260.00
estadocivil	124,682.00	5.00	soltero	64,449.00
profesion	124,225.00	73.00	ninguna	83,498.00
segmento_economico	112,136.00	7.00	empleado privado	45,946.00
fechaapertura	124682	8197	22/06/2019	369
fechanacimiento	124682	21720	01/01/1990	29

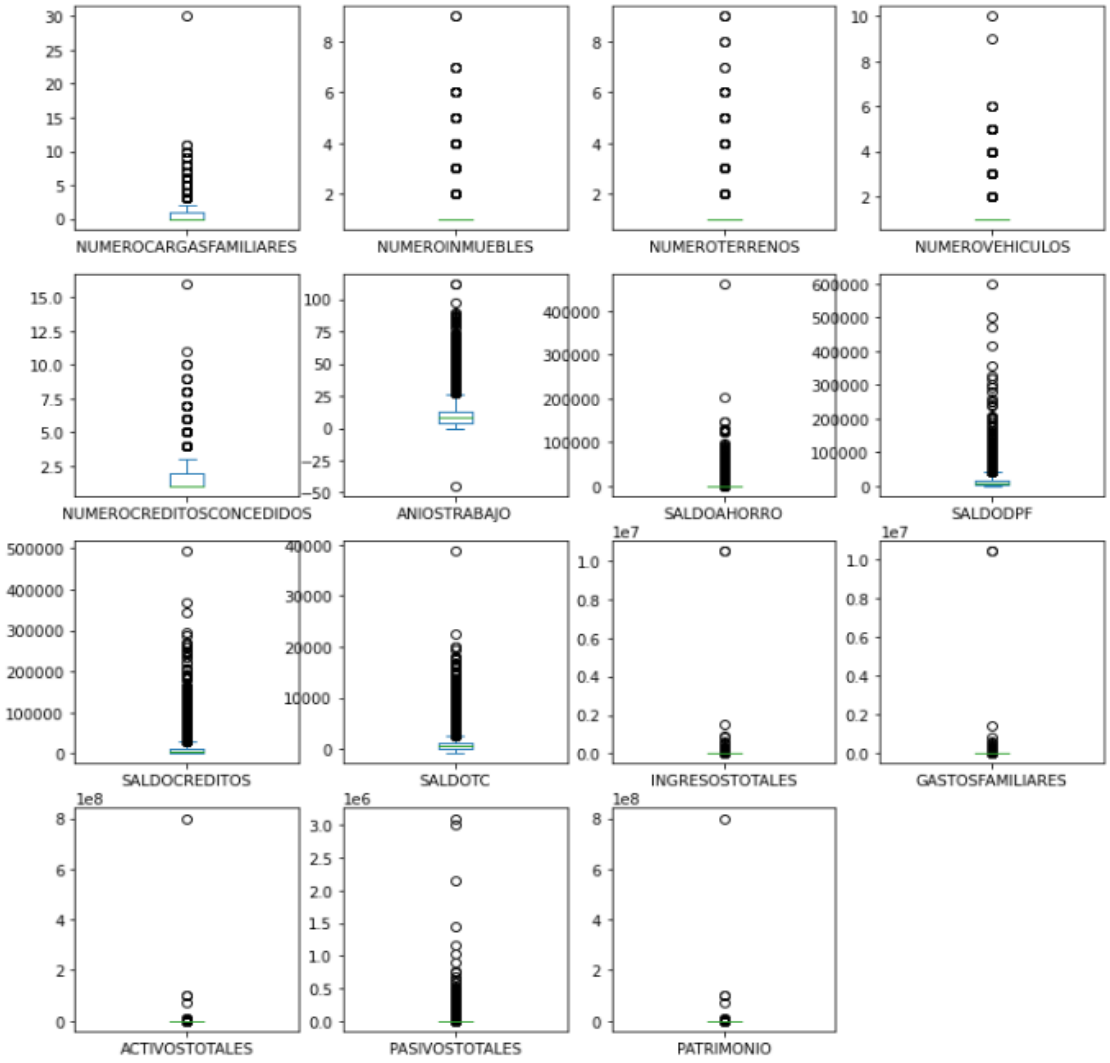
Si bien existen variables catalogadas como numéricas y categóricas no todas son insumos y apropiadas para el modelo. Además, se observa que es necesario una preparación de los datos antes de aplicar el modelo, ya que existen variables con alta desviación estándar lo cual podría dar indicio de valores atípicos.

Análisis gráfico de los datos:

A continuación, se presenta un análisis gráfico mediante diagramas de cajas de las variables cuantitativas, el cual permite representar los datos a través de cuartiles, donde se muestra la mediana, el percentil del 25% y 75%, y los valores extremos inferior y superior, además, se puede visualizar los valores atípicos, aquellos que estén por fuera de los extremos.

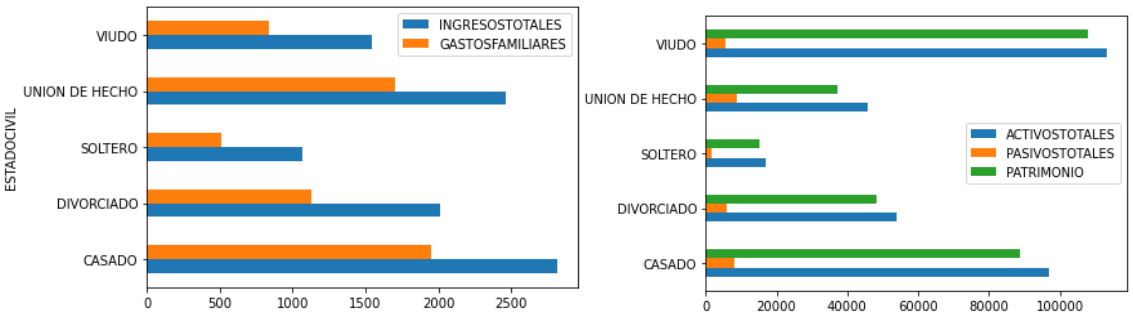
De la gráfica se observa que todas las variables registran valores superiores al valor extremo, principalmente en Saldos DPF, Saldo de Créditos, Pasivos, además se aprecia que existen valores negativos en variables como los Años de Trabajo, Patrimonio.

Ilustración 15 *Diagrama de caja*



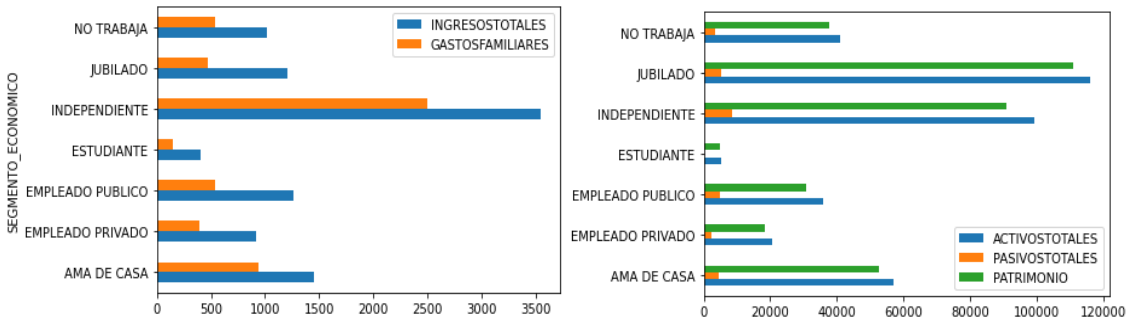
Por otro lado, se realizan gráficos de barras entre variable, en el primer caso se grafica el estado civil frente al nivel de ingresos y gastos, y en la gráfica de la derecha se presenta frente al nivel de activos, pasivos y patrimonio. Resaltan los casados y viudos como con el mayor saldo promedio por cada categoría analizada.

Ilustración 16 *Interacción variables con estado civil*



Al analizar por segmento económico para el gráfico de la izquierda resaltan los independientes con mayores saldos promedios de ingresos y gastos, mientras que en el gráfico de la derecha los jubilados e independientes son los que registran mayores saldos en activos, y patrimonio.

Ilustración 17 *Interacción de variables con segmento económico*



Análisis tablas cruzadas:

Se realiza un análisis del segmento económico y estado civil por saldo promedio de los productos de la institución financiera, saldo ahorro; saldo dpf; saldo crédito y saldo tc.

Respecto al segmento económico se aprecia que en promedio los jubilados registran mayores niveles de saldo de ahorro y dpf, seguido de los no trabaja, este último grupo puede deberse al registro de remesas que recibe. En cuanto a los productos de colocación saldo de créditos y tarjetas de crédito, los independientes registran en promedio el mayor saldo en los dos productos, seguido de los no trabaja en crédito y empleado público en tarjetas de crédito.

Tabla 12 *Segmento económico por saldo promedio de producto*

Segmento económico / Productos	saldoahorro	saldodpf	saldocreditos	saldotc
ama de casa	573.00	14,549.00	11,196.00	572.00
empleado privado	331.00	11,410.00	8,048.00	730.00
empleado público	536.00	10,516.00	10,031.00	954.00
estudiante	275.00	9,618.00	6,276.00	677.00
independiente	643.00	17,923.00	15,736.00	1,038.00
jubilado	1,312.00	20,253.00	8,586.00	753.00
no trabaja	687.00	20,198.00	11,464.00	664.00

En cuanto al estado civil, los viudos registran el mayor saldo promedio de ahorro, mientras que los casados y viudos el mayor saldo por depósitos a plazo fijo. El saldo de

los viudos puede estar relacionado al segmento económico donde los jubilados registran el mayor saldo en estos dos productos. En relación al producto de crédito, los casados, seguido de unión de hecho en promedio registra el mayor saldo, lo cual estaría relacionado con la estructura familiar. Por saldo de tarjetas de crédito los divorciados registran en promedio el mayor saldo.

Tabla 13 *Estado civil por saldo promedio de producto*

Estado civil / Productos	saldoahorro	saldodpf	saldocreditos	saldotc
casado	614.00	17,476.00	15,627.00	931.00
divorciado	648.00	17,048.00	12,165.00	1,002.00
soltero	327.00	11,949.00	8,199.00	800.00
unión de hecho	420.00	9,026.00	12,487.00	831.00
viudo	859.00	17,144.00	11,916.00	807.00

4.2.4 Verificar la calidad de los datos

En el siguiente apartado se realiza la verificación de los datos, mediante el análisis de valores nulos, valores duplicados y atípicos.

Valores duplicados: Existe una cantidad considerable de duplicados, el cual representa el 2,2%, es decir 2,766.

Tabla 14 *Valores duplicados*

Base original	124,682.00
N° Duplicados	2,766.00
% Duplicados	2.22%

Valores nulos: Respecto a los valores nulos, se realiza el análisis por variables, donde el 67% de variables entre numéricas y categóricas (14 de 22) registran valores nulos, siendo la variable de número de terrenos, saldo de dpf, ingresos del cónyuge los que registran mayor porcentaje.

Tabla 15 *Valores nulos*

Variables	Valores nulos	% Valores nulos
numerocargasfamiliares	1,311	1.05%
ingresostotales	0	0
gastosfamiliares	0	0
activostotales	0	0
pasivostotales	0	0
patrimonio	0	0
numeroinmuebles	92,947	74.55%
numeroterrenos	119,886	96.15%
numerovehiculos	105,022	84.23%
numerocreditosconcedidos	94,685	75.94%
aniostrabajo	12,886	10.34%
saldoahorro	1,740	1.40%
saldodpf	117,800	94.48%
saldocreditos	106,394	85.33%
saldotc	101,413	81.34%
sector	1,103	0.88%
instrucción	439	0.35%
estadocivil	0	0.00%
profesion	457	0.37%
segmento_economico	12,546	10.06%
fechaapertura	0	0.00%
fechanacimiento	0	0.00%

Valores atípicos: de los valores atípicos en la Ilustración 15 se evidencia que en su mayoría requieren de un tratamiento al encontrarse valores dispersos

4.3 Preparación de los datos

En esta fase se realiza la preparación de los datos para su posterior modelamiento.

4.3.1 Seleccionar los datos

En la selección de los datos, de las 22 variables se escogen 19, se excluye el sector y la profesión al tener una gran cantidad de categorías, lo cual podría dificultar el modelamiento posterior.

Tabla 16 *Selección de variables*

Tipo	Variables originales	Variables selección
Numéricas	numerocargasfamiliares	numerocargasfamiliares
	ingresostotales	ingresostotales
	gastosfamiliares	gastosfamiliares
	activostotales	activostotales
	pasivostotales	pasivostotales
	patrimonio	patrimonio
	numeroinmuebles	numeroinmuebles
	numeroterrenos	numeroterrenos
	numerovehiculos	numerovehiculos
	numerocreditosconcedidos	numerocreditosconcedidos
	aniostrabajo	aniostrabajo
	saldoahorro	saldoahorro
	saldodpf	saldodpf
	saldocreditos	saldocreditos
	saldotc	saldotc
Categóricas	sector	
	instrucción	instrucción
	estadocivil	estadocivil
	profesion	
	segmento_economico	segmento_economico
	fechaapertura	fechaapertura
	fechanacimiento	fechanacimiento

4.3.2 Limpiar los datos

En esta etapa se emplean diversas técnicas para preparar y optimizar la calidad de los datos, que serán empleados en la fase de modelamiento. De la fase de comprensión de los datos, se detectó que existen valores duplicados, nulos, y atípicos, para cada uno de estos se realiza el respectivo tratamiento.

Valores duplicados: En relación a los valores duplicados, se limpian el 2,2%, es decir 2,766 observaciones, quedando un total de 124,682.

Tabla 17 *Tratamiento valores duplicados*

	Verificación	Limpieza
Base original	124,682.00	124,682.00
N° Duplicados	2,766.00	0
% Duplicados	2.22%	0%

Valores nulos: De los valores nulos, se decide hacer el siguiente tratamiento, imputación de valores con valor cero, se establece este criterio para las variables que se presentan en la tabla, ya que imputarles mediante la media u otro criterio podría generar sesgo de la información, es así que para el caso de la variable "numerocargasfamiliares" si su valor es nulo se considera que no registra cargas familiares el cliente. Respecto a las variables: instrucción, y segmento económico, se eliminan sus valores nulos, dado que al ser características únicas de los clientes y establecer una imputación mediante un criterio en específico podría sesgar la información.

Tabla 18 *Tratamiento valores nulos*

Variables	Valores nulos	% Valores nulos	Tratamiento
fechaapertura	0	0.00%	
fechanacimiento	0	0.00%	
numerocargasfamiliares	1,311	1.05%	imputado valor cero
ingresostotales	0	0	
gastosfamiliares	0	0	
activostotales	0	0	
pasivostotales	0	0	
patrimonio	0	0	
numeroinmuebles	92,947	74.55%	imputado valor cero
numeroterrenos	119,886	96.15%	imputado valor cero
numerovehiculos	105,022	84.23%	imputado valor cero
numerocreditosconcedidos	94,685	75.94%	imputado valor cero
aniostrabajo	12,886	10.34%	imputado valor cero
saldoahorro	1,740	1.40%	imputado valor cero
saldodpf	117,800	94.48%	imputado valor cero
saldocreditos	106,394	85.33%	imputado valor cero
saldotc	101,413	81.34%	imputado valor cero
instrucción	439	0.35%	Se eliminan los valores
estadocivil	0	0.00%	
segmento_economico	12,546	10.06%	Se eliminan los valores

Además, se analizan los valores negativos, el tipo de tratamiento que se realiza es la imputación con valores cero, por ejemplo, la variable años de trabajo registra valor mínimo de -45, lo cual no guarda consistencia con la realidad, debido a que el mínimo de años trabajado debería ser cero.

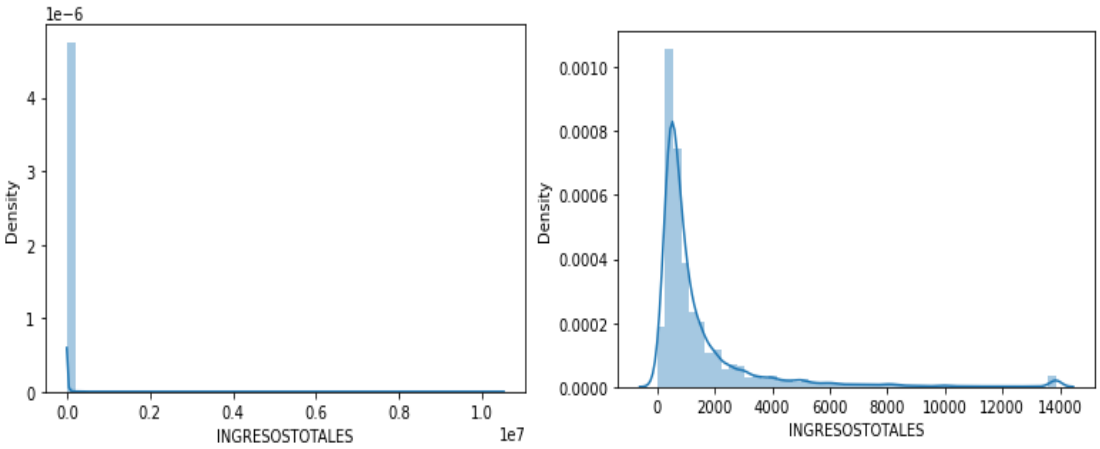
Tabla 19 *Tratamiento valores negativos*

variables	mean	min	tratamiento min
pasivostotales	4,562.00	-205.00	0
patrimonio	46,292.00	-138,473.00	0
saldotc	869.00	-876.00	0
aniostrabajo	10.00	-45.00	0

Valores atípicos:

Para el análisis de valores atípicos se realizan los conocidos gráficos de densidad como se muestra en la siguiente gráfica sobre los ingresos totales, donde se aprecia que existe un sesgo a la derecha, es decir, que existen valores extremos positivos, al realizar el tratamiento de los datos, el gráfico de densidad mejora, e incluso muestra una distribución normal sesgada a la derecha. Se analiza la distribución de cada variable (ver Anexo 1.), y se realiza el respectivo tratamiento mediante la imputación de valores atípicos por el percentil 99.

Ilustración 18 *Densidad - Ingresos totales*



A continuación, se presenta un resumen del tratamiento de los valores atípicos, cuáles fueron sus valores originales, y los actuales después del tratamiento. Donde se evidencia, que en el caso de la variable “numerocargasfamiliares” anteriormente registra un máximo de 30, ahora de 4.

Tabla 20 *Resumen valores originales frente al tratamiento*

	Original			Tratamiento		
variables	**mean**	**75%**	**max**	**mean**	**75%**	**max**
ingresostotales	1,816	1,500	10,512,624	1,472	1,500	13,851
gastosfamiliares	1,115	608	10,412,212	787	635	10,125
pasivostotales	4,562	984	3,108,952	4,102	1,146	78,042
patrimonio	46,292	33,300	800,046,649	36,210	35,500	426,718
aniostrabajo	10	13	112	8	11	37
saldoahorro	466	73	462,433	384	97	9,254
saldoahorro	466	73	462,433	384	97	9,254
saldodpf	15,097	18,000	600,000	634	-	26,105
saldocreditos	11,814	13,334	493,545	1,548	-	38,256
saldotc	869	1,072	38,739	158	-	3,177
edad	42	51	110	41	51	79
antigüedad	8	11	123	7	11	22
Numerocargas familiares	1	1	30	1	1	4

4.3.3 Estructurar los datos

En esta etapa se realiza la generación de nuevos atributos, así como la transformación, a fin de expandir variables y probar el modelo con varios sets. Para las variables instrucción, estado civil, y segmento económico se obtienen dummies.

Tabla 21 *Nuevos atributos*

Variables originales	Nuevos atributos
fechaapertura	Se construye la variable antigüedad
fechanacimiento	Se construye la variable edad
instrucción	Se construye las siguientes variables dummies: Secundaria, Primaria, Superior, Ninguno, Intermedia, Elemental, Postgrado
estadocivil	Se construye las siguientes variables dummies: Soltero; Casado; Divorciado; Unión de hecho; Viudo
segmento_economico	Se construye las siguientes variables dummies: Empleado Privado; Independiente; Empleado Público; Ama de casa; Estudiante; Jubilado; No trabaja

4.3.4 Integrar los datos

En esta etapa se integra los nuevos atributos creados, quedando de esta forma 40 variables de 22 que fueron de forma original.

Tabla 22 *Integración de datos*

Variable original	Variable construida	Detalle	Número de registros	Tipo
fechaapertura	antigüedad	Fecha de apertura de la cuenta de ahorro	109,640.00	Date
fechanacimiento	edad	Fecha de nacimiento	109,640.00	Date
instrucción	d_Secundaria	Variable dummy por nivel de Instrucción del cliente	109,640.00	Dummy
	d_Primaria		109,640.00	Dummy
	d_Superior		109,640.00	Dummy
	d_Ninguno		109,640.00	Dummy
	d_Intermedia		109,640.00	Dummy
	d_Elemental		109,640.00	Dummy
	d_Postgrado		109,640.00	Dummy
estadocivil	d_Soltero	Variable dummy por estado civil	109,640.00	Dummy
	d_Casado		109,640.00	Dummy
	d_Divorciado		109,640.00	Dummy
	d_Unión de hecho		109,640.00	Dummy
	d_Viudo		109,640.00	Dummy
numerocargasfamiliares	numerocargasfamiliares	Número de cargas familiares	109,640.00	Int
ingresostotales	ingresostotales	Ingresos totales	109,640.00	Float
gastosfamiliares	gastosfamiliares	Gastos familiares	109,640.00	Float
activostotales	activostotales	Activos totales	109,640.00	Float
pasivostotales	pasivostotales	Pasivos totales	109,640.00	Float
patrimonio	patrimonio	Patrimonio	109,640.00	Float
numeroinmuebles	numeroinmuebles	Número de inmuebles	109,640.00	Int
numeroterrenos	numeroterrenos	Número de terrenos	109,640.00	Int
numerovehiculos	numerovehiculos	Número de vehículos	109,640.00	Int
numerocreditosconcedidos	numerocreditosconcedidos	Número de créditos concedidos	109,640.00	Int
aniostrabajo	aniostrabajo	Años de trabajo	109,640.00	Float
segmento_economico	d_Empleado Privado	Variable dummy por segmento económico	109,640.00	Dummy
	d_Independiente		109,640.00	Dummy
	d_Empleado Público		109,640.00	Dummy
	d_Ama de casa		109,640.00	Dummy
	d_Estudiante		109,640.00	Dummy
	d_Jubilado		109,640.00	Dummy
	d_No trabaja		109,640.00	Dummy
saldoahorro	saldoahorro	Saldo de ahorro	109,640.00	Float
saldodpf	saldodpf	Saldo de depósitos a plazo	109,640.00	Float
saldocreditos	saldocreditos	Saldo de crédito	109,640.00	Float
saldotc	saldotc	Saldo de tarjeta de crédito	109,640.00	Float
saldoahorro	c_saldoahorro	Variable dummy 1 si el saldo de ahorro es mayor a 0, o 0 caso contrario.	109,640.00	Dummy
saldodpf	c_saldodpf	Variable dummy 1 si el saldo dpf es mayor a 0, o 0 caso contrario.	109,640.00	Dummy
saldocreditos	c_saldocreditos	Variable dummy 1 si el saldo de crédito es mayor a 0, o 0 caso contrario.	109,640.00	Dummy
saldotc	c_saldotc	Variable dummy 1 si el saldo de tarjetas de crédito es mayor a 0, o 0 caso contrario.	109,640.00	Dummy

4.3.5 Formateo de los datos

Respecto al formateo de los datos se realiza la transformación de los mismo sin modificar su significado, es así que se elimina los espacios de los nombres de las variables mediante el siguiente script *columns.str.strip(),* a fin de facilitar el manejo al momento del modelamiento.

4.3.6 Exploración previo el modelamiento de los datos

Después de haber realizar el tratamiento de la información se vuelve a realizar el análisis descriptivo de los datos, donde se compara el comportamiento original frente al tratado. La principal diferencia radica en el número de observaciones de cada variable, y la disminución del grado de dispersión en la mayoría de estas (desviación estándar), lo que implica que la influencia de los valores extremos se encuentra mitigado con el preprocesamiento. En el caso de los ingresos totales, el promedio disminuye de $1,816 a $1,472, con un grado de dispersión de 42,869 a 2,066, lo que significa que los valores atípicos ejercieron influencia en esta (Ver Anexo 2, total de variables).

Tabla 23 *Descriptivos datos con tratamiento*

	Original			Tratamiento		
Variables	**count**	**mean**	**std**	**count**	**mean**	**std**
numerocargasfamiliares	123,371	1	1	109,640	1	1
ingresostotales	124,682	1,816	42,869	109,640	1,472	2,066
gastosfamiliares	124,682	1,115	42,263	109,640	787	1,496
activostotales	124,682	50,855	2,312,298	109,640	54,081	2,465,650
pasivostotales	124,682	4,562	23,103	109,640	4,102	12,020
patrimonio	124,682	46,292	2,311,714	109,640	36,210	70,171
numeroinmuebles	31,735	1	0	109,640	0	1
numeroterrenos	4,796	1	1	109,640	0	0
numerovehiculos	19,660	1	1	109,640	0	1
numerocreditosconcedidos	29,997	2	1	109,640	0	1
aniostrabajo	111,796	10	8	109,640	8	8
saldoahorro	122,942	466	3,037	109,640	384	1,292
saldodpf	6,882	15,097	26,407	109,640	634	3,380
saldocreditos	18,288	11,814	20,153	109,640	1,548	5,548
saldotc	23,269	869	1,378	109,640	158	502

Análisis matriz de correlación:

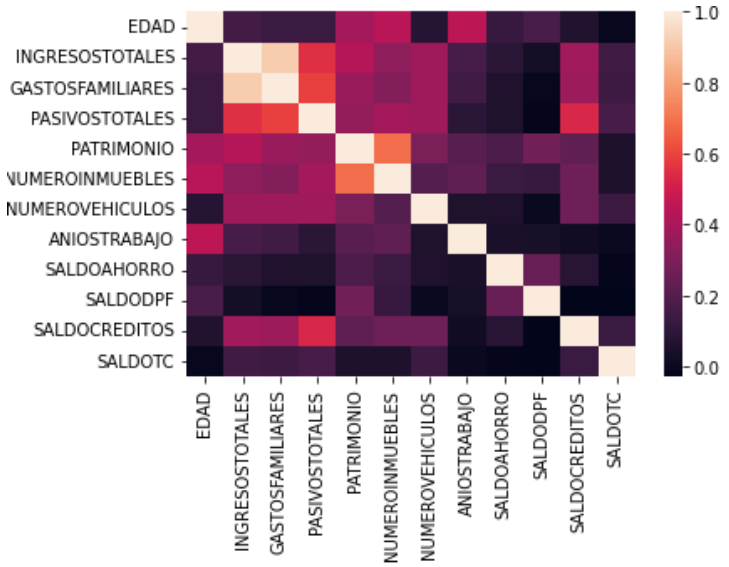
Se realiza la matriz de correlación, para ello se escoge un conjunto de variables a fin de observar la correlación entre las mismas.

Tabla 24 *Matriz de correlación*

	EDAD	INGRESOSTC	GASTOSFAM	PASIVOSTOT	PATRIMONIO	NUMEROINM	NUMEROVEH	ANIOSTRABA	SALDOAHOR	SALDODPF	SALDOCREDI	SALDOTC
EDAD	100%	16%	13%	13%	39%	44%	7%	45%	12%	17%	6%	-1%
INGRESOSTOTALES	16%	100%	91%	55%	43%	34%	38%	16%	9%	3%	39%	15%
GASTOSFAMILIARES	13%	91%	100%	59%	37%	32%	38%	15%	6%	-1%	37%	14%
PASIVOSTOTALES	13%	55%	59%	100%	35%	39%	38%	9%	6%	-2%	53%	17%
PATRIMONIO	39%	43%	37%	35%	100%	68%	30%	21%	18%	27%	23%	5%
NUMEROINMUEBLES	44%	34%	32%	39%	68%	100%	20%	23%	13%	12%	26%	5%
NUMEROVEHICULOS	7%	38%	38%	38%	30%	20%	100%	5%	6%	0%	26%	14%
ANIOSTRABAJO	45%	16%	15%	9%	21%	23%	5%	100%	4%	4%	2%	0%
SALDOAHORRO	12%	9%	6%	6%	18%	13%	6%	4%	100%	25%	8%	-2%
SALDODPF	17%	3%	-1%	-2%	27%	12%	0%	4%	25%	100%	-2%	-3%
SALDOCREDITOS	6%	39%	37%	53%	23%	26%	26%	2%	8%	-2%	100%	13%
SALDOTC	-1%	15%	14%	17%	5%	5%	14%	0%	-2%	-3%	13%	100%

A su vez, se presenta a la matriz de correlación como un mapa de calor, donde valores que se acerquen a 1 registran un color claro, lo que significa que tienen una alta correlación, mientras que valores que tienden a cero el color es negro, con una baja correlación.

Ilustración 19 *Mapa de calor matriz de correlación*



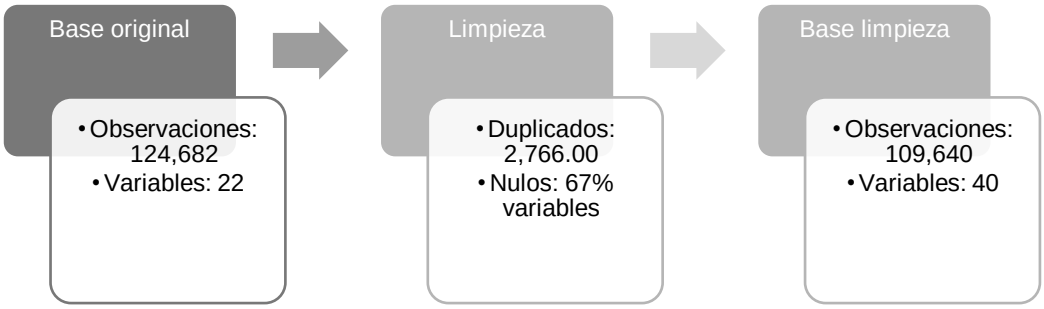
La matriz de correlación permite identificar relaciones entre variables, cabe señalar que la correlación entre dos variables no significa causalidad, es decir que la dirección o comportamiento de una variable sea causa del otro o viceversa. A continuación, se presenta un análisis de los resultados obtenidos de la matriz:

- La variable edad registra una correlación positiva sobre el 39% con el patrimonio, número de inmuebles y años de trabajo, lo que significa que a mayor edad mayor será el valor de estas.
- Respecto al ingreso total, registra una alta correlación positiva con el gasto familiar, lo cual guarda coherencia, ya que a mayores ingresos mayores gastos. Así también, registra una alta correlación con la variable pasivo.

- Respecto a los pasivos registra una correlación positiva con el saldo de créditos, esto guarda relación, dado que el cliente al registrar deudas en la institución se carga en sus pasivos.
- En cuanto a la variable patrimonio tienen una alta correlación con el número de inmuebles con un 68%.

En base al análisis realizado sobre la comprensión de los datos y la preparación de estos para su posterior uso en la fase de modelamiento, se presenta un resumen. A priori la base fue de 124,682 observaciones con 22 variables, al realizar el tratamiento de la información, la base final registra un total de 109,640 con 40 variables.

Ilustración 20 *Resumen desde la verificación hasta la preparación de los datos*



4.4 Modelado

De acuerdo a lo señalada en el apartado 3.1.1 se plantea utilizar la metodología clúster para segmentar a los clientes y recomendar productos de acuerdo a sus características.

Dado que la metodología clúster es un método de aprendizaje no supervisado, no se requiere realizar el particionamiento de la data entre test y train, sin embargo, se realiza la comprobación del modelo mediante varias muestras aleatorias a fin de demostrar la robustes de los resultados.

4.4.2 Construcción del modelo

A continuación, se presenta el procedimiento que se realiza para la construcción del modelo clúster.

Tabla 25 *Pasos para la construcción del modelo clúster (k-means)*

1. Escalamiento variables	• Se realiza el escalamiento de variables mediante la distribución normal.
2. Medida de distancia.	• Euclideana
1. Número k de clústeres	• Se emplea el método de elbow (codo), método de average silhouette y método gap. • Se obtiene 10 clúster
2. Seleccionar aleatoriamente k observaciones del conjunto de datos como centroides de inicio	• Algoritmo k-means realiza esta actividad
3. Asignar a cada una de las observaciones al centroide que se encuentra más cerca.	• Algoritmo k-means realiza esta actividad
4. Para los k clústers generados se vuelve a calcular su centroide.	• Algoritmo k-means realiza esta actividad
5. Repetir los dos últimos pasos, a fin de que las asignaciones sean robustas.	Se repite el proceso con diferentes muestras

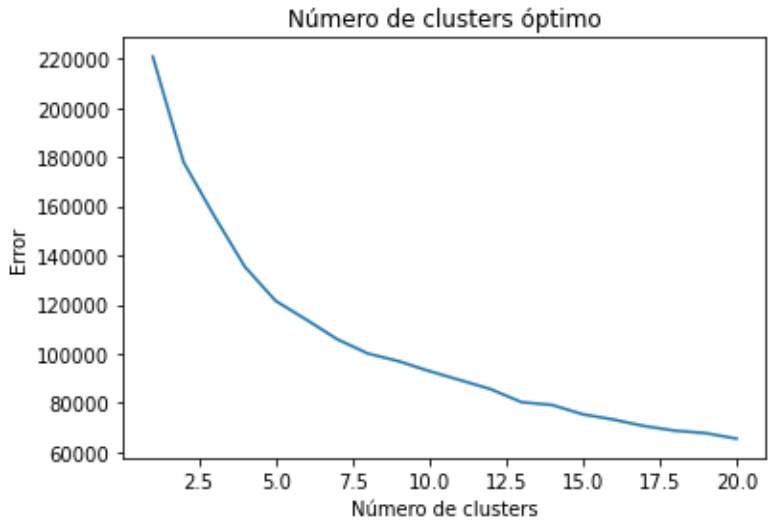
Número de clústeres:

De acuerdo a (Hung, Phan Duy; Thi, Nguyen; Duc, 2019), para establecer el número de clústeres óptimo se consideran distintos métodos: elbow (codo), método de average silhouette, y método gap.

Método elbow (codo):

El método elbow como se menciona en el apartado de la metodología permite escoger el número de clúster donde minimice la varianza total del clúster, de esta forma se escoge el número de clúster 10, ya que al añadir mayor número de clústeres consigue minimizar la varianza, pero no en gran magnitud.

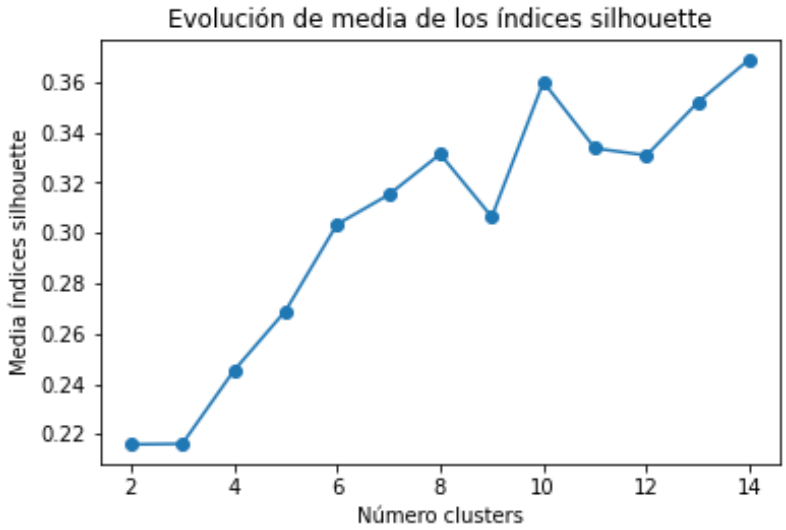
Ilustración 21 *Número de clúster con el método codo*



Método average silhouette:

Por su parte con el método silueta (average silhouette), su máximo valor se encuentra en el clúster 10 y 13, donde representa que se ha asignado correctamente la observación al clúster.

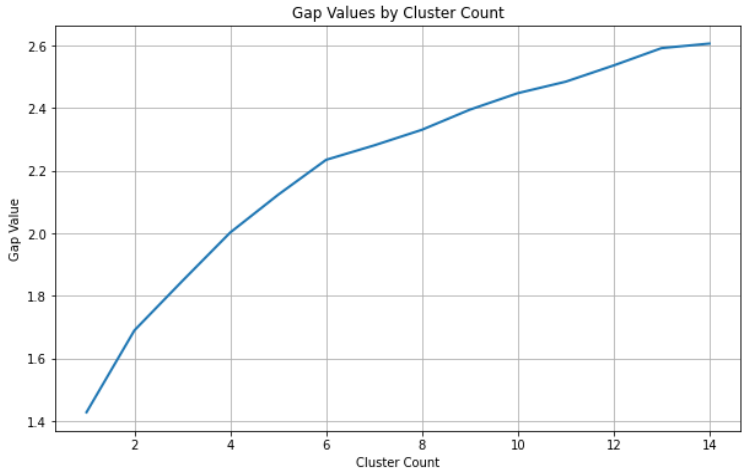
Ilustración 22 *Número de clúster con el método silhouette*



Método GAP:

De acuerdo al método GAP, el número óptimo de clúster, es aquel donde se consigue maximizar el valor del gap, de acuerdo al gráfico existe un punto de quiebre en el clúster número 10 y 13.

Ilustración 23 *Número de clúster con el método GAP*

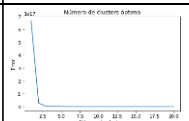
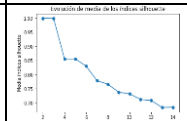
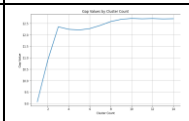
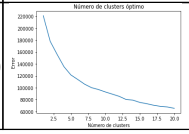
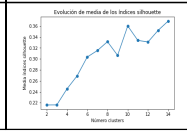
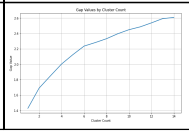


Del análisis realizado con los tres métodos de evaluación para la determinación de número de clúster se decide tomar 10 clústeres, ya que determina una mejor agrupación para la interpretación de resultados. La evaluación se realizará con diferentes muestras más adelante, lo que permitirá comprobar la robustes de los resultados.

Modelado:

En la Tabla 26 se presenta el modelo clúster sin escalamiento y con escalamiento. Se evidencia que, al no realizar un tratamiento previo de los datos (estandarizarlos), y considerarlos en su unidad de medida original los métodos codo, silhouette y gap recomiendan que los datos se agrupen en 3 segmentos, concentrando prácticamente el 100% de las observaciones en un solo clúster. Por el contrario, al realizar el escalamiento de las variables, los métodos codo, silhouette y gap recomiendan agrupar las observaciones en 10 segmentos, donde la concentración para cada clúster es más equitativa, como se muestra en la última columna de la tabla.

Tabla 26 *Modelo clúster sin escalamiento y con escalamiento*

Detalle		N° variables	N° clúster	Codo	Silhouette	Gap	Observaciones por clúster	
Modelo Original	Sin escalamiento	34	3				0 109636 2 3 1 1	
Modelo Original	Con escalamiento	34	10				4 20807 3 20219 6 15320 0 15109 1 11233	7 6852 8 6598 2 4825 9 4355 5 4322

La Tabla 27 muestra los resultados de la metodología, las filas corresponden a las características de los clientes, mientras que las columnas son el número de clústeres al que está agrupado.

Para el modelamiento se prueba la metodología clúster con varios sets de variables, como se muestra en el Anexo 3. Se decide optar por el modelo presentado a continuación, ya que se incluyen variables como el nivel de instrucción y el segmento económico, que permite conocer las características de los clientes.

Tabla 27 *Características del clúster*

clúster	0.00	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00
edad	42.00	50.00	36.00	35.00	43.00	45.00	39.00	44.00	45.00	40.00
antigüedad	8.00	8.00	8.00	6.00	7.00	7.00	7.00	7.00	8.00	7.00
numerocargasfamiliares	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ingresostotales	801.00	1,983.00	984.00	818.00	2,032.00	3,444.00	852.00	4,518.00	1,493.00	1,339.00
gastosfamiliares	370.00	1,211.00	418.00	336.00	1,088.00	2,020.00	383.00	3,241.00	636.00	558.00
activostotales	25,798.00	72,436.00	24,196.00	17,282.00	114,491.00	128,665.00	30,816.00	123,134.00	59,070.00	41,569.00
pasivostotales	1,746.00	5,029.00	3,393.00	1,453.00	3,299.00	11,718.00	1,868.00	21,770.00	6,793.00	4,828.00
patrimonio	23,214.00	52,631.00	20,372.00	14,846.00	46,446.00	86,480.00	27,118.00	83,473.00	50,260.00	34,597.00
numeroinmuebles	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
numeroterrenos	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
numerovehiculos	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
numerocreditosconcedidos	0.00	0.00	1.00	0.00	0.00	1.00	0.00	2.00	1.00	0.00
aniostrabajo	7.00	14.00	6.00	7.00	10.00	10.00	5.00	9.00	9.00	8.00
saldoahorro	325.00	441.00	283.00	257.00	386.00	684.00	369.00	540.00	599.00	473.00
saldodpf	518.00	699.00	314.00	304.00	625.00	1,493.00	852.00	303.00	1,235.00	999.00
saldocreditos	910.00	1,909.00	1,365.00	731.00	0.00	3,764.00	413.00	13,047.00	1,912.00	1,603.00
saldotc	86.00	105.00	909.00	0.00	130.00	272.00	0.00	444.00	255.00	199.00
d_elemental	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

d_intermedia (técnica)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
d_ninguno	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
d_postgrado	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
d_primaria	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
d_secundaria	0.00	0.00	1.00	1.00	1.00	0.00	1.00	1.00	0.00	0.00
d_superior	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	1.00	1.00
d_ama de casa	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
d_empleado privado	1.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00
d_empleado publico	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
d_estudiante	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
d_independiente	0.00	1.00	0.00	0.00	1.00	1.00	0.00	1.00	0.00	0.00
d_jubilado	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
d_no trabaja	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
c_saldodpf	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
c_saldocreditos	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
c_saldotc	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

En términos generales, aquellos clientes con mayores saldos de crédito se agrupan en el clúster 7. Los clientes agrupados en este clúster registran las siguientes características, adulto que en promedio tiene 44 años, con ingresos en promedio altos, con niveles de activos en promedio altos en relación a otros grupos, así como el patrimonio, con bienes inmuebles y vehículo, con crédito, con 9 años promedio de experiencia, con un nivel de instrucción secundaria y que pertenecen al segmento independiente.

En cuanto a los clientes con mayores saldos en tarjetas de crédito, se encuentra en el clúster 2, con las siguientes características: adulto con 36 años en promedio, con ingresos en promedio medios en relación a otros grupos, con niveles de activos en promedio medios, así como el patrimonio, sin bienes muebles ni inmuebles, con crédito, con 6 años promedio de experiencia, con nivel de instrucción secundaria y empleado.

Por su parte, aquellos clientes con mayores saldos en el producto depósitos a plazo fijo (DPF) y saldo de ahorro se encuentran en el clúster 5, y registran las siguientes características: adulto con edad promedio de 45 años, con ingresos promedio altos, con niveles de activos en promedio altos en relación a otros grupos, así como el patrimonio, con bienes inmuebles y crédito, con 10 años de experiencia, con educación superior e independiente. El detalle se presenta en la Tabla 27.

Tabla 28 *Recomendaciones de productos por clúster*

Clúster	Característica	Recomendación
0	Adulto 42 años, con ingresos medios, con niveles de activos medios en relación a otros grupos, así como el patrimonio, sin bienes muebles ni inmuebles, con 7 años de experiencia, primaria y empleado	No resalta ningún producto en particular
1	Adulto 50 años, con ingresos medio alto, con niveles de activos medio alto en relación a otros grupos, así como el patrimonio, sin bienes muebles ni inmuebles, con 14 años de experiencia, primaria e independiente	Resalta el producto de crédito
2	Adulto 36 años, con ingresos medios, con niveles de activos medios en relación a otros grupos, así como el patrimonio, sin bienes muebles ni inmuebles, con crédito, con 6 años de experiencia, secundaria y empleado	Resalta de tarjeta de crédito
3	Adulto 35 años, con ingresos medios, con niveles de activos medio bajo en relación a otros grupos, así como el patrimonio, con 7 años de experiencia, secundaria y empleado	No resalta ningún producto en particular
4	Adulto 43 años, con ingresos altos, con niveles de activos altos en relación a otros grupos, así como el patrimonio, sin bienes muebles	No resalta ningún producto en particular

	ni inmuebles, con 10 años de experiencia, secundaria e independiente	
5	Adulto 45 años, con ingresos altos, con niveles de activos altos en relación a otros grupos, así como el patrimonio, con bienes inmuebles y crédito, con 10 años de experiencia, superior e independiente	Resalta el producto de ahorro y de DPF
6	Adulto 39 años, con ingresos medios, con niveles de activos medios en relación a otros grupos, así como el patrimonio, sin bienes muebles ni inmuebles, con 5 años de experiencia, secundaria	No resalta ningún producto en particular
7	Adulto 44 años, con ingresos altos, con niveles de activos altos en relación a otros grupos, así como el patrimonio, con bienes inmuebles y vehículo, con crédito, con 9 años de experiencia, secundaria e independiente	Resalta el producto de crédito con el mayor monto
8	Adulto 45 años, con ingresos medios, con niveles de activos medios en relación a otros grupos, así como el patrimonio, sin bienes muebles ni inmuebles, con crédito, con 9 años de experiencia, superior y empleado	Resalta el producto de ahorro y de DPF
9	Adulto 40 años, con ingresos medios, con niveles de activos medios en relación a otros grupos, así como el patrimonio, sin bienes muebles ni inmuebles, con 8 años de experiencia, superior y empleado	Resalta el producto de ahorro y de DPF

*Los valores son promedio de las características presentadas en cada clúster.

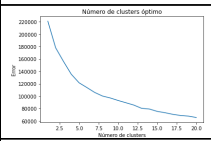
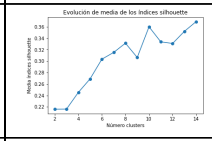
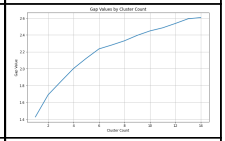
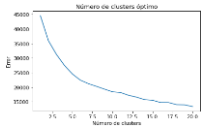
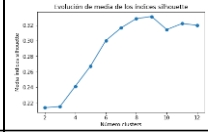
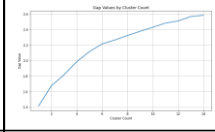
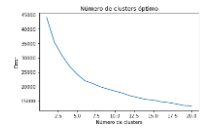
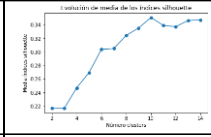
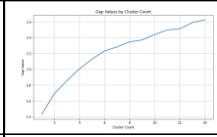
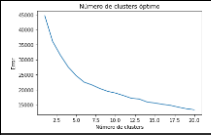
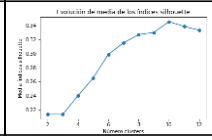
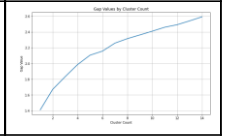
4.4.3 Evaluación del modelo

Para la evaluación del modelo se aplica la misma metodología y set de variable en diferentes muestras aleatorias, con el fin de verificar si la asignación del número de clústeres con el método elbow, average silhouette y gap son similares. Del Dataset principal se particiona en tres grupos, y se vuelve a correr el modelo k-means.

La Tabla 29 presenta un resumen del modelo original y las tres muestras con el método elbow, silhouette y gap. En términos generales, se observa que en su mayoría recomiendan 10 agrupaciones como número de clúster óptimo. Cabe señalar que estas muestras al igual que el Dataset original son con escalamiento.

De esta forma, los resultados obtenidos con el método k-means para la segmentación de clientes es robusto, con una precisión aceptable.

Tabla 29 *Número de clúster con diferentes muestras*

Muestras	Detalle	N° variables	N° clúster	Codo	Silhouette	Gap
Modelo Original	Con escalamiento	34	10			
Muestra 1	Con escalamiento	34	9			
Muestra 2	Con escalamiento	34	10			
Muestra 2	Con escalamiento	34	10			

5. Capítulo V: Análisis de resultados

Los resultados muestran que las características sociodemográficas de los clientes potenciales de la institución financiera se diferencian de acuerdo al producto que demanden. Si bien la metodología segmenta en 10 grupos a los clientes de acuerdo a sus características, para el análisis se considera a los grupos potenciales (mercado objetivo) para la institución, ya que registran en promedio un mayor saldo en los productos que demandan.

Al tener identificado las características de los potenciales clientes para la institución, se deben establecer estrategias de ventas y marketing para capturar y mantener este tipo de clientes y de esta forma incrementar la demanda de los productos, lo que aportaría al incremento de la cuota de mercado a nivel nacional y fidelización de los clientes.

Es así que para los ***productos de crédito*** resaltan los adultos con edades promedio de 44 años, con niveles de ingresos, activos y patrimonio altos, con bienes y con una experiencia promedio de 9 años, que pertenezcan al segmento independiente y con niveles de instrucción de secundaria. Las estrategias que se deben establecer para este grupo son los relacionados con la comunicación, los cuales debe ser mediante canales que estén a la mano de este segmento, como publicidad en televisión, radio, redes sociales, con el objetivo de que se adapten a este nicho de mercado. Así también, destaca la característica del segmento independiente, por lo que la estrategia de marketing podría estar encaminada a potenciar sus empresas, mediante el acceso a créditos.

Respecto a las características de los clientes que registran en promedio mayores niveles de ***ahorro y depósitos a plazo fijo DPF***, destacan adultos de edades mayor o igual a 45 años en promedio, con altos niveles de ingresos y activos, con nivel de educación superior, por lo que, este segmento podría tener una mejor educación financiera, al destacar sus niveles de ahorro. Una estrategia para fidelizar a estos clientes, y retenerlos es ofrecerles mejores tasas de interés, a mejores plazos. Así, como el establecimiento de estrategias dirigidas de marketing a este nicho de mercado de acuerdo a sus características.

En cuanto a las características de los clientes con mayores saldos en ***tarjetas de crédito*** se encuentran aquellos con edades promedio de 36 años, más jóvenes que los anteriores segmentos, con niveles de ingresos, activos y patrimonio medios en relación

a otros grupos, sin bienes muebles ni inmuebles, con niveles de instrucción secundaria y que forman parte del segmento de empleados. Este segmento de clientes puede ser la entrada para la inclusión financiera de nuevos clientes, una estrategia es inducirlos en educación financiera, mediante una colocación de tarjetas de crédito con bajos cupos de crédito, evitando el sobreendeudamiento.

Estos grupos hacen referencia a las características de los clientes que mayor demanda registran en créditos, ahorro, depósitos a plazo fijo DPF, y tarjetas de crédito. En la siguiente tabla se muestra las características de estos y otros grupos potenciales para la recomendación de productos.

Tabla 30 *Características clientes potenciales*

Clúster	Característica	Recomendación
1	Adulto 50 años, con ingresos medio alto, con niveles de activos medio alto en relación a otros grupos, así como el patrimonio, sin bienes muebles ni inmuebles, con 14 años de experiencia, primaria e independiente	Resalta el producto de crédito
2	Adulto 36 años, con ingresos medios, con niveles de activos medios en relación a otros grupos, así como el patrimonio, sin bienes muebles ni inmuebles, con crédito, con 6 años de experiencia, secundaria y empleado	Resalta tarjeta de crédito
5	Adulto 45 años, con ingresos altos, con niveles de activos altos en relación a otros grupos, así como el patrimonio, con bienes inmuebles y crédito, con 10 años de experiencia, superior e independiente	Resalta el producto de ahorro y de DPF
7	Adulto 44 años, con ingresos altos, con niveles de activos altos en relación a otros grupos, así como el patrimonio, con bienes inmuebles y vehículo, con crédito, con 9 años de experiencia, secundaria e independiente	Resalta el producto de crédito con el mayor monto
8	Adulto 45 años, con ingresos medios, con niveles de activos medios en relación a otros grupos, así como el patrimonio, sin bienes muebles ni inmuebles, con crédito, con 9 años de experiencia, superior y empleado	Resalta el producto de ahorro y de DPF
9	Adulto 40 años, con ingresos medios, con niveles de activos medios en relación a otros grupos, así como el patrimonio, sin bienes muebles ni inmuebles, con 8 años de experiencia, superior y empleado	Resalta el producto de ahorro y de DPF

Respecto a la evaluación del ***proceso CRISP-DM***, durante la aplicación del modelo se ha retroalimentado cada paso, al ser un proceso interactivo e iterativo permite ajustar pasos anteriores a fin de obtener mejores resultados. Así, por ejemplo, ha sido necesario volver a la fase de preparación de datos, a fin de construir nuevas variables que permitan mejorar el modelo. En definitiva, mediante el proceso CRISP-DM se cumple con los objetivos del negocio y de minería de datos.

Con el fin de mejorar los resultados de la segmentación de clientes, es necesario establecer ***líneas de continuidad*** con el estudio las características psicográficas, de comportamiento, y geográfica de los clientes, lo cual agregaría valor a las estrategias planteadas. Es importante señalar que, si bien el análisis se relaciona al nivel de demanda que tienen estos clientes con los productos de la institución, se debe realizar un análisis sobre el nivel de rentabilidad que dan a la institución. De esta forma se debería incluir variables por tipo de producto, así, por ejemplo, para el caso de los créditos y tarjetas de crédito se debe incorporar variables de comportamiento del cliente, como el score de crédito o los niveles promedio de morosidad.

Finalmente, el ***despliegue*** de este proyecto se vincula con las recomendaciones que se realiza al negocio sobre los resultados obtenidos en el proceso CRISP-DM, lo que se transforma en estrategias para este.

6. Capítulo VI: Conclusiones y Recomendaciones

6.1 Conclusiones

El objetivo del presente estudio es recomendar productos financieros a clientes potenciales de una institución financiera del Ecuador mediante la segmentación de clientes. Para el análisis de minería de datos se emplea el proceso CRISP-DM, donde se incorpora los objetivos del negocio. Para la fase del modelado se aplica el método de aprendizaje no supervisado, metodología clúster mediante k-means, con información de los clientes corte diciembre del 2022.

De los resultados obtenidos, el modelo k-means segmenta a los clientes en diez grupos, de los cuales se escoge tres como los potenciales para la institución, al registrar un mayor nivel de saldos en promedio en los productos de ahorro, depósitos a plazo fijo, créditos y tarjetas de crédito.

En los productos de crédito, ahorro y depósitos a plazo fijo, en términos generales destacan aquellos clientes con mayores niveles de ingresos, activos y patrimonio, así como personas en edades promedio de 44 años, que pertenecen al segmento independiente, entre otras características. Por lo que, el target para la recomendación de productos debe estar encaminado a este tipo de segmento que demandan en mayor medida estos, siendo considerados como de estrato alto, en comparación a otros segmentos de clientes analizados.

Respecto a las características de los clientes con mayores saldos en tarjetas de crédito se encuentran las personas con edades promedio de 36 años, más jóvenes que los anteriores segmentos, con niveles de ingresos medios, así mismo registran niveles medios en sus activos y patrimonio, medido en relación a otros grupos, y forman parte del segmento de empleados, entre otras características. Siendo el target de clientes que mayor demanda de tarjeta de créditos registra, considera de estrato medio en relación a otros segmentos. De la misma forma, las estrategias deben estar encaminadas a fortalecer este segmento.

En esta medida, los recursos y esfuerzos para el establecimiento de estrategias del negocio se vuelven importante y deben estar dirigidas de acuerdo a las características de cada segmento económico y el mercado objetivo de la institución, lo que permitirá la fidelización de los clientes, e incremento de la cuota de mercado.

6.2 Recomendaciones

Del análisis planteado, se recomienda:

Analizar las características psicográficas, y geográfica de los clientes, a fin de fortalecer las estrategias del negocio. Como menciona (Harrison, 2006), la segmentación de clientes bajo la perspectiva del uso del producto, potencia el análisis de las actitudes y motivaciones que tienen los clientes hacia el producto. Esto enriquecería el análisis, ya que las decisiones dependen de las aspiraciones y ambiente de las personas.

Además, se recomienda incluir variables de comportamiento del socio sobre el uso del producto, esto a fin de determinar el nivel de rentabilidad y beneficio que dan a la institución. Tal es el caso para los productos de créditos y tarjetas de créditos, cuyo comportamiento de los clientes depende de la puntualidad de sus pagos, para lo cual se pueden incluir variables como: score de crédito o los niveles promedio de morosidad.

Finalmente, se recomienda realizar una evaluación periódica sobre el modelo para asegurar la eficacia y mejoras continuas del modelo.

7. Referencias

Amat, J. (2017). *Clustering y heatmaps: aprendizaje no supervisado*. https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps

Dáderman, A., & Rosander, S. (2018). *Evaluating Frameworks for Implementing Machine Learning in Signal Processing and KDD*.

Gallardo, J. (2010). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM*.

Gollapudi, S. (2016). *Practical Machine Learning*.

Harrison, T. S. (2006). *Mapping Customer Segments for Personal Financial Services*.

Hung, Phan Duy; Thi, Nguyen; Duc, N. (2019). *Customer Segmentation Using Hierarchical Agglomerative Clustering*. 33–37.

Jayant, Tikmani; Sudhanshu, Tiwari; Sujata, K. (2015). Telecom customer segmentation based on cluster analysisAn Approach to Customer Classification using k-means. *IJIRCCE*.

Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018). Customer Segmentation using K-means Clustering. *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 135–139.

Monil, P., Darshan, P., Jecky, R., Vimarsh, C., & Bhatt, P. B. R. (2020). *Customer Segmentation using Machine Learnin*. *June*.

Naresh K. Malhotra. (2008). *Naresh K. Malhotra*. https://doi.org/10.1017/CBO9781107415324.004

Piatetsky-Shapito, G. (1991). *Knowledge Discovery in Databases*.

Samuel, A. (1959). *Some Studies in Machine Learning*.

Sulekha, G. (2011). *The basis of market segmentation : a critical review of literature*. 45–55.

Tibshirani, R., Walther, G., & Hastie, T. (2001). *Estimating the number of clusters in a data set via the gap statistic*. 411–423.

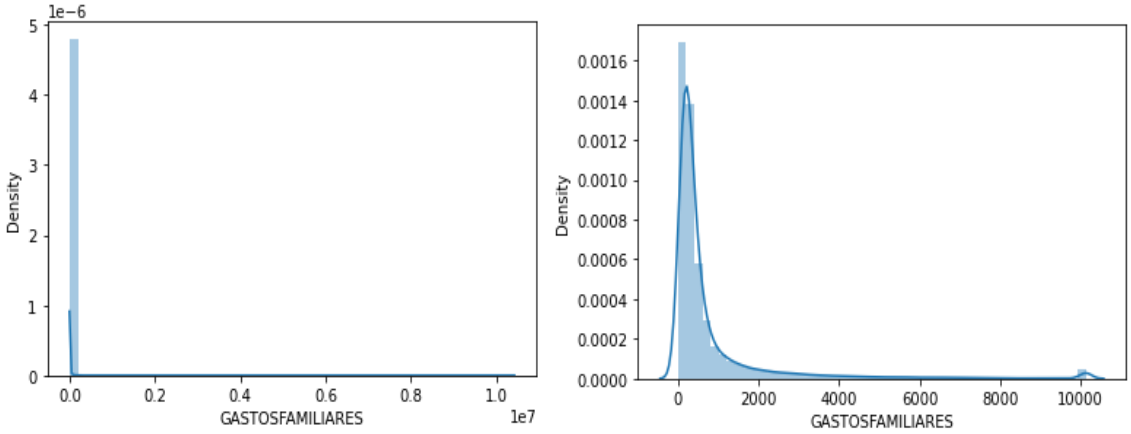
Yang, X., Chen, J., Hao, P., & Wang, Y. J. (2015). *Application of Clustering for Customer Segmentation in Private Banking*. *9631*, 1–6. https://doi.org/10.1117/12.2197182

Yefta, Christian; Oktaviani, K. (2022). *Penerapan K-Means pada Segmentasi Pasar untuk Riset Pemasaran pada*. *9*(4), 966–973. https://doi.org/10.30865/jurikom.v9i4.4486

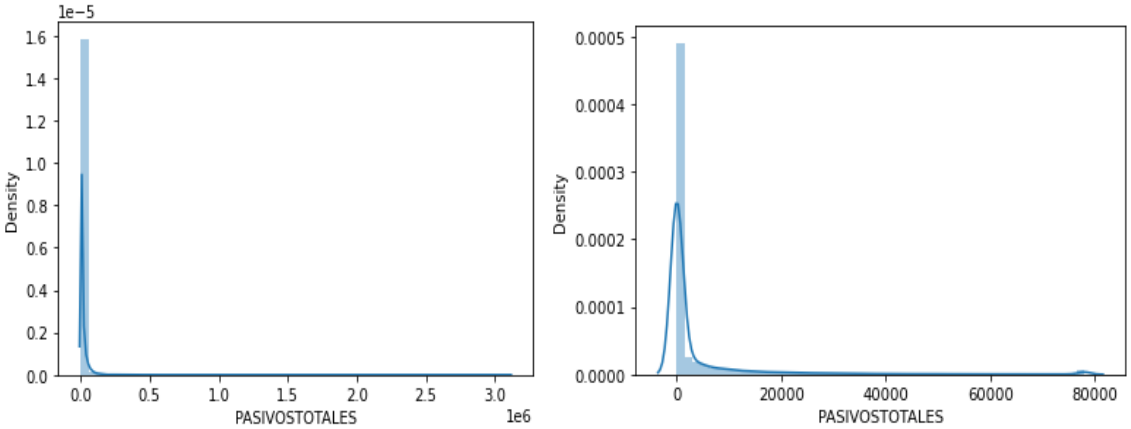
8. Anexos

Anexo 1 *Limpieza de valores atípicos*

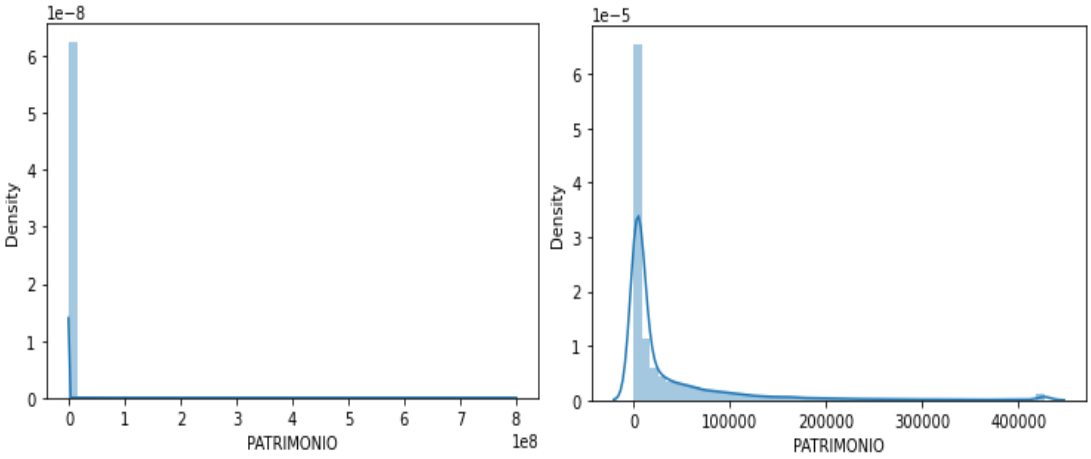
Gastos familiares



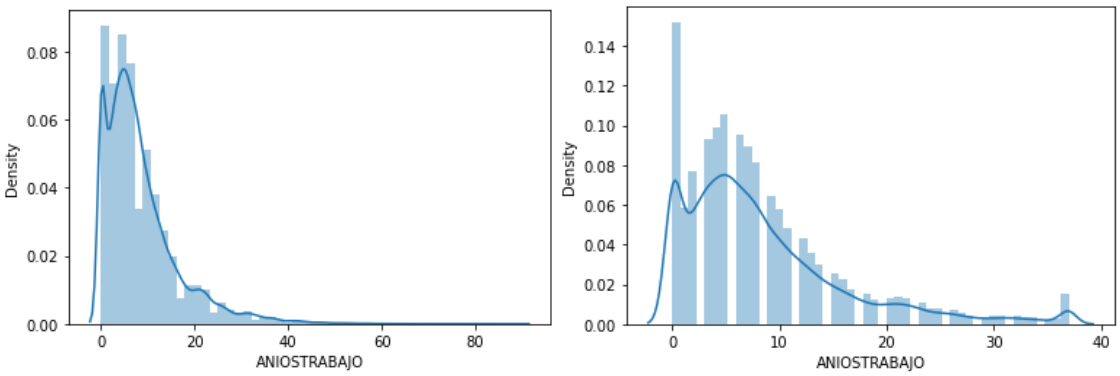
Pasivos totales



Patrimonio



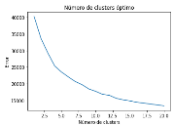
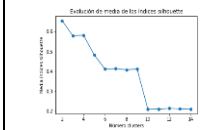
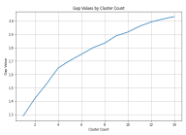
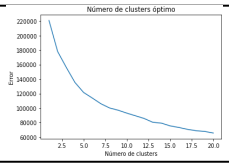
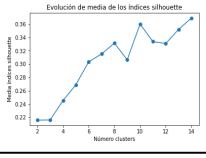
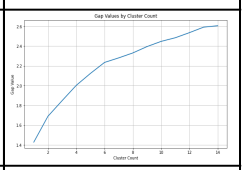
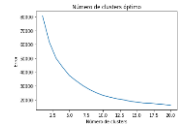
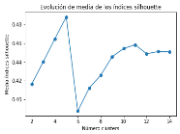
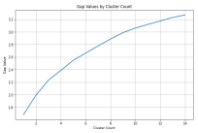
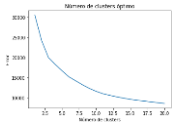
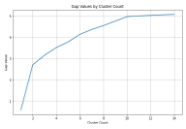
Años de trabajo



Anexo 2 *Exploración datos*

	Tratamiento		
Variables	**count**	**mean**	**std**
numerocargasfamiliares	109,640.00	1.00	1.00
ingresostotales	109,640.00	1472.00	2066.00
ingresosconyugetotales	109,640.00	142.00	563.00
gastosfamiliares	109,640.00	787.00	1496.00
activostotales	109,640.00	54081.00	2465650.00
pasivostotales	109,640.00	4102.00	12020.00
patrimonio	109,640.00	36210.00	70171.00
numeroinmuebles	109,640.00	0.00	1.00
numeroterrenos	109,640.00	0.00	0.00
numerovehiculos	109,640.00	0.00	1.00
numerocreditosconcedidos	109,640.00	0.00	1.00
aniostrabajo	109,640.00	8.00	8.00
saldoahorro	109,640.00	384.00	1292.00
saldodpf	109,640.00	634.00	3380.00
saldocreditos	109,640.00	1548.00	5548.00
saldotc	109,640.00	158.00	502.00
antiguedad	109,640.00	7.00	5.00
edad	109,640.00	41.00	14.00
d_elemental	109,640.00	0.00	0.00
d_intermedia (técnica)	109,640.00	0.00	0.00
d_ninguno	109,640.00	0.00	0.00
d_postgrado	109,640.00	0.00	0.00
d_primaria	109,640.00	0.00	0.00
d_secundaria	109,640.00	1.00	0.00
d_superior	109,640.00	0.00	0.00
d_ama de casa	109,640.00	0.00	0.00
d_empleado privado	109,640.00	0.00	0.00
d_empleado publico	109,640.00	0.00	0.00
d_estudiante	109,640.00	0.00	0.00
d_independiente	109,640.00	0.00	0.00
d_jubilado	109,640.00	0.00	0.00
d_no trabaja	109,640.00	0.00	0.00
d_casado	109,640.00	0.00	0.00
d_divorciado	109,640.00	0.00	0.00
d_soltero	109,640.00	1.00	0.00
d_union de hecho	109,640.00	0.00	0.00
d_viudo	109,640.00	0.00	0.00
c_saldoahorro	109,640.00	1.00	0.00
c_saldodpf	109,640.00	0.00	0.00
c_saldocreditos	109,640.00	0.00	0.00
c_saldotc	109,640.00	0.00	0.00

Anexo 3 *Resultados diferentes modelos*

	Detalle	N° variables	N° clúster	Codo	Silhouette	Gap
Modelo 1	Con escalamiento	16	4			
Modelo Original	Con escalamiento	34	10			
Modelo 3	Con escalamiento	16	5			
Modelo 4	Con escalamiento	12	8			

Modelo de prueba A

Modelo 1				
clúster	**0**	**1**	**2**	**3**
edad	32	58	41	46
antigüedad	5	9	9	9
numerocargasfamiliares	0	0	2	1
ingresostotales	870	1,352	1,363	7,620
gastosfamiliares	339	607	787	5,456
activostotales	15,119	108,344	34,397	209,444
pasivostotales	1,106	3,236	2,782	39,101
numeroinmuebles	0	1	0	1
numeroterrenos	0	0	0	0
numerovehiculos	0	0	0	1
numerocreditosconcedidos	0	0	0	1
aniostrabajo	5	14	9	11
saldoahorro	241	637	311	708
saldodpf	259	1,597	350	437
saldocreditos	656	989	888	14,475
saldotc	144	121	150	481

clúster	Característica	Recomendación
0	Jóvenes, con ingresos más bajo, no registra activos ni pasivos significativos, no registra inmuebles ni bienes, con 5 años laborales.	No resaltan en ningún producto
1	Personas adultas, ingresos medios, con activos significativos, y bienes inmuebles, con amplia experiencia.	Resalta producto de ahorro
2	Personas adultas, ingresos medios, con bajos activos, sin bienes inmuebles, con amplia experiencia.	Resalta producto de ahorro
3	Personas adultas con ingresos altos, y altos niveles de activos, registra bienes inmuebles, vehículos y créditos, 11 años de trabajo.	Registra altos valores en créditos y tarjetas de crédito

Modelo de prueba B

Modelo 3							
clúster	**0**	**1**	**2**	**3**	**4**	**5**	**6**
edad		56	43	33	43	39	39
ingresostotales	3,819	1,398	2,092	834	1,315	1,954	1,522
gastosfamiliares	2,446	647	1,279	363	737	1,154	756
pasivostotales	13,003	2,585	10,347	727	3,231	8,738	4,099
patrimonio	119,607	53,174	48,383	10,078	30,967	35,804	36,000
numeroinmuebles	1	-	-	-	-	-	-
numerovehiculos	-	-	-	-	-	-	-
aniostrabajo	10	14	8	5	10	8	8
saldoahorro	1,405	468	425	220	-	205	534
saldodpf	8,467	-	54	-	19	145	98
saldocreditos	2,931	-	8,684	-	-	4,527	4
saldotc	55	-	-	-	-	1,063	897
c_saldoahorro	1	1	1	1	-	-	1
c_saldodpf	1	-	-	-	-	-	-
c_saldocreditos	-	-	1	-	-	-	-
c_saldotc	-	-	-	-	-	1	1

clúster	Característica	Recomendación
0	Adulto, ingresos medios bajos, bajos niveles patrimoniales, sin bienes, 8 años laborables.	No resalta ningún producto
1	Adulto, ingresos medios bajos, medio altos niveles patrimoniales, un bien inmueble, 9 años laborables.	Resalta productos de ahorro y DPF
2	Adulto, ingresos medios bajos, bajos niveles patrimoniales, 8 años laborables.	Resalta TC
3	Adulto, ingresos medios alto, medio altos niveles patrimoniales, un bien inmueble y vehículo, 8 años laborables.	Resalta créditos
4	Adulto, ingresos medios alto, bajos niveles patrimoniales, sin bienes, 10 años laborales.	No resalta ningún producto

Modelo de prueba C

Modelo 4										
clúster	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
edad	29	40	57	61	41	47	45	56	57	46
ingresostotales	732	1,889	1,494	1,033	1,274	10,817	4,128	1,762	3,089	1,684
gastosfamiliares	302	995	761	473	645	8,145	2,535	626	1,521	770
pasivostotales	783	6,866	3,246	2,271	2,199	37,684	38,587	2,292	10,078	3,694
patrimonio	7,423	32,696	47,089	38,717	20,034	142,826	94,785	160,086	287,274	56,718
numeroinmuebles	-	-	1	-	-	1	1	1	1	-
numerovehiculos	-	-	-	-	-	1	1	-	1	-
aniostrabajo	3	8	26	6	10	12	8	10	12	9
saldoahorro	147	174	291	256	153	596	545	2,207	484	7,256
saldodpf	110	121	393	398	126	348	213	22,813	580	1,323
saldocreditos	494	2,153	662	709	607	7,872	24,616	533	1,577	2,293
saldotc	77	2,089	45	46	54	367	285	66	119	76

clúster	Característica	Recomendación
0	Adultos ingresos medio bajo patrimonio medio bajo, con 7 años de experiencia.	No resaltan productos
1	Adultos ingresos medio alto patrimonio alto, con bien inmueble, con 10 años de experiencia.	Resalta producto DPF
2	Joven ingresos bajo, patrimonio bajo, sin bien inmueble, con 5 años de experiencia.	No resalta producto
3	Adultos ingresos medio alto, patrimonio medio, sin bien inmueble, con 8 años de experiencia.	Resalta producto tarjetas de crédito
4	Adultos ingresos alto, patrimonio alto, con bienes muebles e inmueble, con 12 años de experiencia.	Resalta producto crédito
5	Adultos ingresos medio bajo, patrimonio medio, con bien inmueble, con 24 años de experiencia.	No resalta producto
6	Adultos ingresos alto, patrimonio alto, con bien inmueble, con 8 años de experiencia.	Resalta producto de crédito
7	Adultos ingresos medio bajo, patrimonio medio alto, sin bien mueble e inmueble, con 9 años de experiencia.	Resalta producto de ahorro