GOBIERNO DE MÉXICO | CONAHCYT CONSEJO NACIONAL DE HUMANIDADES CIENCIAS Y TECNOLOGÍAS | INFOTEC

BIBLIOTECA INFOTEC
VISTO BUENO DE TRABAJO TERMINAL

Maestría en Ciencia de Datos e Información
(MCDI)

Ciudad de México, a 13 de febrero de 2024

UNIDAD DE POSGRADOS
PRESENTE

Por medio de la presente se hace constar que el trabajo de titulación:

"Segmentación de clientes bancarios con K-MEANS"

Desarrollado por el alumno: **Alfonso de Jesús Granados Álvarez**, bajo la asesoría del **Dr. Daniel Alejandro Cervantes Cabrera** cumple con el formato de Biblioteca, así mismo, se ha verificado la correcta citación para la prevención del plagio; por lo cual, se expide la presente autorización para entrega en digital del proyecto terminal al que se ha hecho mención. Se hace constar que el alumno no adeuda materiales de la biblioteca de INFOTEC.

No omito mencionar, que se deberá anexar la presente autorización al inicio de la versión digital del trabajo referido, con el fin de amparar la misma.

Sin más por el momento, aprovecho la ocasión para enviar un cordial saludo.

Mtro. Carlos Josué Lavandeira Portillo
Director Adjunto de Innovación y Conocimiento

CJLP/jah

C.c.p. Felipe Alfonso Delgado Castillo.- Gerente de Capital Humano.- Para su conocimiento.
Alfonso de Jesús Granados Álvarez.- Alumno de la Maestría en Ciencia de Datos e Información.- Para su conocimiento.

MCDI

MAESTRÍA EN CIENCIA DE DATOS E INFORMACIÓN

INFOTEC CENTRO DE INVESTIGACIÓN E INNOVACIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN

DIRECCIÓN ADJUNTA DE INNOVACIÓN Y CONOCIMIENTO

GERENCIA DE CAPITAL HUMANO

POSGRADOS

SEGMENTACIÓN DE CLIENTES BANCARIOS CON K-MEANS

Solución estratégica
Que para obtener el grado de MAESTRO EN CIENCIA DE DATOS E INFORMACIÓN

Presenta:

Alfonso de Jesús Granados Álvarez

Asesor:

Daniel Alejandro Cervantes Cabrera

Ciudad de México, agosto, 2023

CONACYT

INFOTEC POSGRADOS

Tabla de contenido

Índice de figuras

Índice de cuadros

Resumen

En el presente trabajo se implementa una segmentación de clientes bancarios, con la finalidad de encontrar y describir grupos cuyas características permitan proponer estrategias que sustenten la toma de decisiones del negocio, seguido de la implementación de un modelo de aprendizaje supervisado que clasifique los nuevos clientes al grupo correspondiente y de esta forma poder asociarles un conjunto de productos o servicios adecuados a sus características.

Lo anterior se lleva a cabo siguiendo las etapas que se pueden encontrar en la mayoría de los proyectos de ciencia de datos, que son:

- *Adquisición de los datos*: Una base de datos de acceso público a través de la plataforma *Kaggle* que cuenta con 12 variables que incluyen información demográfica, calificación en el sistema financiero y línea de crédito, por mencionar algunas [5].
- *Análisis exploratorio*: Se realiza un análisis cuantitativo y visual (con gráficas) de las variables para eliminar correlaciones, identificación y tratamiento de datos faltantes y atípicos, así como la identificación de variables que no aporten información relevante.
- *Preprocesamiento*: Dada la naturaleza de la información, es necesario realizar las transformaciones y normalizaciones correspondientes, que permitan una vectorización adecuada para su posterior uso en el modelo.
- *Implementación del algoritmo de segmentación*: Se inicia con la implementación del modelo llamado *K-Medias* (*K-Means*), utilizando diferentes combinaciones de las variables en busca de una segmentación adecuada. Al finalizar esta etapa, se analizan los grupos obtenidos para darles una descripción que tenga sentido a nivel negocio.
- *Implementación del algoritmo de clasificación*: Utilizando la clasificación que obtenemos con la configuración elegida del algoritmo *K-Means*, se entrena y

valida un Árbol de decisiones para utilizarlo en la clasificación de nuevos clientes del banco.

Finalmente, describiremos algunas estrategias basadas en los resultados de este trabajo, que nos permiten proponer soluciones “inteligentes” o informadas en el problema de segmentación de clientes y asignación automática de productos y servicios.

Introducción

Sin importar el giro de una empresa, es fundamental tener un amplio conocimiento de sus clientes con la finalidad de adaptar y mejorar su oferta de productos y servicios que permita generar un impacto positivo en los ingresos. Las compañías del sector financiero como los bancos, no son la excepción, ya que por cuestiones regulatorias y del comportamiento del negocio se necesita conocer mucho mejor a sus clientes. Esta situación **plantea el problema** de establecer metodologías de agrupación o *clustering* de usuarios de acuerdo a sus características particulares (por ejemplo nivel de ingreso, escolaridad, etc.) y consumo o uso de instrumentos bancarios-financieros, contribuyendo así a la toma de acciones estratégicas y pertinentes de acuerdo a cada uno de estos grupos.

Este trabajo tiene el **objetivo** de desarrollar una metodología para la segmentación de usuarios de entidades bancarias en función de características específicas que permita identificar grupos o cúmulos de clientes con rasgos similares que contribuya a la definición de estrategias comerciales adaptadas a las necesidades individuales.

Su **justificación** se basa, en el hecho de que actualmente la tecnología permite que una parte importante de las operaciones bancarias se realicen en línea desde una computadora o a través de dispositivos inteligentes como los smartphones, lo cual brinda a las entidades financieras, en el mejor de los casos, el acceso a la información de los usuarios prácticamente en tiempo real, dando origen a la necesidad de crear procesos automáticos que permitan clasificar y organizar la información de forma eficiente, y que contribuya en la toma de decisiones del negocio.

Así, podemos decir que el **alcance** de esta tesis, consiste en establecer una metodología básica del proceso de construcción de cúmulos y asignación de servicios y/o productos bancarios-financieros para bases de datos de clientes bancarios similares, teniendo en consideración que el acceso a esta información no es sencilla debido a las estrategias institucionales y legislaciones actuales.

Por ejemplo, una institución bancaria necesita identificar las características y comportamientos de los clientes para evitar en medida de los posible delitos como el lavado de dinero o financiamiento al terrorismo, al mismo tiempo debe identificar aquellos que tienen probabilidad de incumplir con sus pagos (caer en mora) o identificar clientes con potencial para generar un mayor ingreso. Una vez que conocen las características de los clientes, es más sencillo implementar campañas focalizadas o estrategias preventivas según sea el caso.

Todo lo anterior se puede abordar con la *segmentación de clientes*, ya que es una tarea que permite separar a los clientes de tal forma que se puedan crear distintos grupos en los cuales éstos tengan características similares, lo cual permite adaptar la estrategia organizacional a las características de cada grupo, facilitando el éxito de las empresas [6]. Una de las ventajas que se obtienen gracias a una segmentación creativa, es que las empresas pueden conseguir ventajas competitivas, ya que para cada segmento identificado se puede adaptar la mejor oferta de producto o servicio, política de precios y medios de distribución [6]. La segmentación se puede realizar en diferentes sectores y el bancario no es la excepción, esta tarea permite conocer a los clientes de cada segmento de una forma generalizada, lo cual da pie a la construcción e implementación de estrategias mejor enfocadas y con los recursos necesarios.

Distintos autores concuerdan con las características que pueden ser utilizadas para la elaboración de una segmentación, tales como el uso de productos o servicios, la demografía, ubicación geográfica o estilo de vida y definen cuatro categorías de segmentación: conductual, demográfica, psicográfica y geográfica. Dichas categorías no son necesariamente excluyentes, ya que es posible utilizar una combinación de ellas independientemente del tipo de segmentación seleccionado [6].

De acuerdo con la recomendación que hace Bautista (2021) [6] donde menciona que se deben realizar segmentaciones utilizando varios tipos de información de los clientes y no de un solo tipo, es por lo que en este trabajo se implementará una segmentación de clientes considerando variables económicas y geográficas.

Capítulo 1

Estado del arte

1 Estado del arte

El punto de partida por excelencia para este tipo de proyectos consiste en estudiar diversos trabajos que aborden problemas similares al que se pretende analizar, ya que este proceso permite identificar un diverso número de estrategias que se pueden replicar o evitar si no son convenientes, logrando enriquecer el trabajo en cuestión, porque su elaboración deja de apoyarse solo en la experiencia del autor, dado que se revisaron distintas perspectivas de un objeto de estudio con similitudes significativas, como las que se presentan a continuación.

En el trabajo de Didik, Mediana y Miranda (2018) [4] se realizó una segmentación de clientes de un banco con base en información referente a las transacciones realizadas a través de internet, como la fecha, el monto de la transacción y el saldo del cliente, para posteriormente aplicar los algoritmos K-Means y K-Medoids, para elegir el mejor algoritmo utilizaron el índice Davies-Bouldin, es decir, eligieron el número de clusters que tuviera menor índice, en este caso, los dos algoritmos consiguieron desempeños muy similares, siendo K-Medoids ligeramente mejor que K-Means.

Dong, Zhang y Ye (2017) [7] realizaron una segmentación de clientes para un banco comercial chino, en la cual consideraron variables como la edad, estatus marital, nivel de educación, ingresos anuales, total de depósitos y monto total de consumo, dichas variables fueron utilizadas para aplicar el algoritmo K-Means con 5 clusters. Como resultado obtuvieron 5 grupos de clientes (Con potencial de crecimiento, Generales, Intermedios, Senior y VIP) que definieron de acuerdo con las características de cada grupo lo cual les ayudó a definir la estrategia de administración para cada cluster, como promociones de bajo costo, descuentos, incremento de incentivos y servicios personalizados.

También se revisó el trabajo de Mahova y Pavlov (2018) [11] en el cual se busca la segmentación de 100 prestatarios de un banco por medio de K-Means, para la implementación de un programa de lealtad que consiste en la emisión de 3 tipos de tarjetas “platinum”, “gold” y “plateada”. Para lo anterior, consideraron variables como

el monto del préstamo, antigüedad del cliente. El algoritmo se implementó en dos ocasiones, la primera con las variables mencionadas y para la segunda se estandarizaron las variables. Posteriormente definieron 6 estrategias comerciales en las que se le da cierta importancia a cada variable. Por último, el trabajo concluye que la decisión de qué implementación escoger (con variables normales o estandarizadas) depende de a qué le quiere dar más relevancia la institución financiera.

De acuerdo con el trabajo realizado por Smeureanu, Ruxanda y Badea (2013) [18] donde se busca segmentar en dos grupos a los clientes de un banco de Rumania. Primero se identificaron a los clientes que dejan mayores ganancias al banco a través de cierto comportamiento, para esto calcularon el promedio de la suma mensual de sus transacciones, los clientes que superaron los 4,500 RON fueron clasificados como *"affluent"* y el resto como *"mass"*. Una vez que fueron asignadas las categorías se realizó la selección de variables sobre una base estadística y también sobre el juicio de expertos. El trabajo enlista 31 variables predictoras que contemplan el monto y número de transacciones además de algunos lugares donde fueron hechas como tiendas de joyerías, de libros, hoteles y restaurantes por mencionar algunas. Posterior a esto se aplican 2 técnicas de machine learning, redes neuronales con una capa oculta consiguen una exactitud del 93.8% pero solo un 41% en la identificación de la clase de interés (affluent); al usar SVM obtienen una exactitud del 97%, acertando a un 83% la clase de interés.

En algunas ocasiones resulta necesario reducir el número de variables en el conjunto de datos con la finalidad de disminuir el costo computacional al momento de procesarlos, y al mismo tiempo se busca conservar las características principales que describen a la base, un ejemplo de esto se puede encontrar en la investigación de Alkhayrat, Aljnidi y Aljouma (2020) [1], en donde se realizó una comparación del resultado de aplicar las técnicas: Análisis de componentes principales (PCA) y Red neuronal autocodificadora (Autoencoder Neural Network), para reducir la dimensionalidad del conjunto de datos y posteriormente definir una segmentación de usuarios de telecomunicaciones, para ello contaron con una base de datos con 220 variables. Para ello, generaron tres segmentaciones diferentes: con los datos originales, apli-

cando PCA y aplicando la Red neuronal autocodificadora. Gracias a la reducción de dimensionalidad lograron reducir el número de variables a 20, consiguiendo mejores resultados en las métricas de Silueta y Davies-Bouldin en comparación a las obtenidas al momento de segmentar con las 220 variables iniciales; siendo la segmentación utilizando la base con la reducción de dimensionalidad generada por la Red neuronal autocodificadora la que mejor se desempeñó, pero al mismo tiempo hacer la acotación de que si un mapeo lineal puede explicar la varianza de los datos, es mejor usar el PCA.

Ciertamente contar con un volumen significativo de información, puede hacernos pensar en que se está capturando en gran parte todas las características del objeto de estudio, en este caso, los clientes, pero en algunos momentos no es posible contar con esa cantidad de datos, pero a pesar de eso, es posible conseguir resultados relevantes, como se presenta en el trabajo de Kilari (2022) [10], cuya finalidad fue segmentar a los clientes de una tienda de un centro comercial mediante el algoritmo K-Means, utilizando una base de datos con 200 registros y 5 variables, de las cuales solo conservaron 2, el ingreso anual de los clientes y la calificación de gasto que les otorgó la tienda (entre más alta, significaba que compraban más), apoyándose de la métrica de Silueta, definen el número de cúmulos a 5, los cuales al ser graficados se distinguen claramente. Culminaron el trabajo realizando recomendaciones para cada grupo de clientes, resaltando aquellos con un ingreso alto - calificación de gasto alta e ingreso alto - calificación de gasto baja, con el objetivo de fidelizar a esos los clientes.

Una alternativa a K-Means, es el algoritmo de Partición Alrededor de Medoids (PAM) que utiliza K-Medoids para realizar el agrupamiento, tal y como se puede ver en el trabajo de Elguera (2018) [8], apoyándose de dicha herramienta, se realizó una segmentación de clientes de un casino, donde después de realizar el preprocesamiento de la información, conservaron 8 de las 20 variables iniciales, tales como el monto acumulado, de recarga, número de visitas al mes, edad, sexo, entre otras. Logrando encontrar 3 clusters, para posteriormente utilizar esa segmentación para entrenar un árbol de clasificación que les permitió definir de una mejor forma el perfil de los clientes, apoyándose de las reglas generadas por el clasificador.

Capítulo 2

Metodología

2 Metodología

Dos de los términos más comunes al hablar de las tareas que se pueden llevar a cabo usando aprendizaje de máquina o automático son *aprendizaje supervisado* y *no supervisado*, los cuales nos indican el tipo de tarea que se llevará a cabo y cuál será el resultado esperado. Cada tipo de aprendizaje se conforma por un conjunto de algoritmos que tienen distintas finalidades.

2.1. Aprendizaje supervisado

La finalidad en este tipo de tareas es aprender la relación existente entre un conjunto de datos de entrada y la variable objetivo correspondiente a cada instancia del conjunto de datos. Desde el punto de vista de Provost (2013) [13], hablando metafóricamente, un maestro “supervisa” al aprendiz proveyéndole información con un conjunto de datos.

Dentro del aprendizaje supervisado se encuentran dos principales subclases, que son las tareas de clasificación y regresión, éstas se distinguen por el tipo de variable objetivo que se quiere predecir, para la primera el objetivo es predecir el valor de una variable categórica, mientras que para la segunda se busca predecir el valor para una variable numérica continua.

Algunos de los algoritmos más comunes para tareas de aprendizaje supervisado son [12]:

- K-NN: K Vecinos más cercanos (K-Nearest Neighbors)
- Regresión lineal
- Regresión logística
- SVM: Máquinas de Vectores de Soporte (Support Vector Machine)
- Clasificador ingenuo de Bayes (Naive Bayes Classifier)

2.2. Aprendizaje no supervisado

En las tareas de este tipo, no se cuenta con una variable objetivo la cual se quiera predecir, en este caso solo se tiene el conjunto de datos de entrada y se busca identificar ciertas regularidades dentro de la información, es decir, se le pide al algoritmo que a partir de la información de entrada se extraiga conocimiento. Como opina Provost (2013) [13] que al tener un conjunto de ejemplos que no cuentan con una variable objetivo, el aprendiz no recibirá información acerca del propósito del aprendizaje, sino que será dejado que construya sus conclusiones acerca de qué tienen en común los ejemplos recibidos. En el aprendizaje no supervisado se encuentras dos tareas principales:

Transformación de los datos, este campo contiene los algoritmos que crean una nueva representación de los datos, ya sea para una mejor interpretación o para ser usados en otros algoritmos de machine learning. Una de las aplicaciones más comunes de este tipo de trasformaciones es la reducción de dimensionalidad que toma un conjunto de datos de alta dimensionalidad y encuentra una nueva representación de los datos que resume las principales características del conjunto original.

Algoritmos de clustering, buscan dividir el conjunto de datos en diferentes grupos de tal forma que en cada grupo se encuentren observaciones muy similares y que al mismo tiempo sea muy distintas a observaciones pertenecientes a otros grupos.

Uno de los restos que hay que enfrentar a la hora de utilizar un algoritmo de aprendizaje no supervisado es al momento de evaluar si el algoritmo aprendió algo útil, ya que en contraste con los algoritmos supervisados, ahora no contamos con una etiqueta para evaluar el desempeño.

Algunos de los algoritmos usados para tareas de aprendizaje no supervisado son [12]:

- K-Means
- K-Medoids

- GMM (Modelo de Mezclas Gaussianas)
- Agrupamiento jerárquico
- DBSCAN

2.3. Estudio del modelo K-Means

Es uno de los algoritmos más utilizados para elaborar análisis de clusters, a grandes rasgos, el algoritmo intenta identificar los centroides de los clusters que sean los más representativos en ciertas regiones de los datos. El algoritmo se puede resumir en dos pasos: asignación de cada observación del conjunto de datos al centroide más cercano y definición del nuevo centroide del cluster como la media de los datos que fueron asignados a éste. El algoritmo termina cuando las observaciones ya no cambian del centroide asignado [12].

Matemáticamente se describe de la siguiente forma:

Dado un conjunto de datos

$$X = \{x^t\}_{t=1}^{N}$$

Con k vectores de referencia,

$$m_j \mid j = 1, \ldots, k$$

Una función de error

$$E(\{m_i\}_{i=1}^{k} \mid X) = \sum_t \sum_i b_i^t \parallel x^t - m_i \parallel^2$$

Donde

$$b_i^t = \begin{cases} 1 \text{ si } \| x^t - m_i \| = \min_j \| x^t - m_j \| \\ \\ 0 \text{ en otro caso} \end{cases}$$

Los mejores vectores de referencia son aquellos que minimizan la función de error; dado que b_i^t también depende de m_j, no se puede resolver este problema de minimización de forma analítica. Por lo que se tiene un proceso iterativo llamado K-Means para esto. Primero se inicia con valores aleatorios para m_i y en cada iteración se calcula b_i^t para todos los x^t, que son las *etiquetas estimadas*. Si b_i^t es 1 se dice que x^t pertenece al grupo m_i, y una vez teniendo dichas etiquetas se minimiza la ecuación de error, derivando respecto a m_i e igualando a 0 se obtiene

$$m_i = \frac{\sum_t b_i^t x^t}{\sum_t b_i^t}$$

es decir, el vector de referencia se define como la media de todas las observaciones que representa, este es un proceso iterativo ya que una vez que se calculan los nuevos m_i, b_i^t cambian, por lo que se tiene que recalcular, lo que a su vez afecta m_i, estos dos pasos se repiten hasta que m_i se estabiliza [2].

Para describir el proceso anterior desde una perspectiva más gráfica, se muestra la siguiente figura 2.1, donde se implementa el algoritmo sobre una base de datos sintética y se buscan 3 clusters. En esta figura se puede observar la inicialización, donde se seleccionan los centros de forma aleatoria, a continuación se asignan los puntos más cercanos a cada centro y en las iteraciones siguientes se recalculan los centros con el promedio de los puntos asignados a cada cluster. Posterior a la segunda iteración se puede observar que la asignación de los puntos no cambia, por lo cual se finaliza el algoritmo [12].

Dada nueva información, cada observación será asignada al cluster más cercano, en la figura 2.2, se observan el área correspondiente a cada cluster [12].

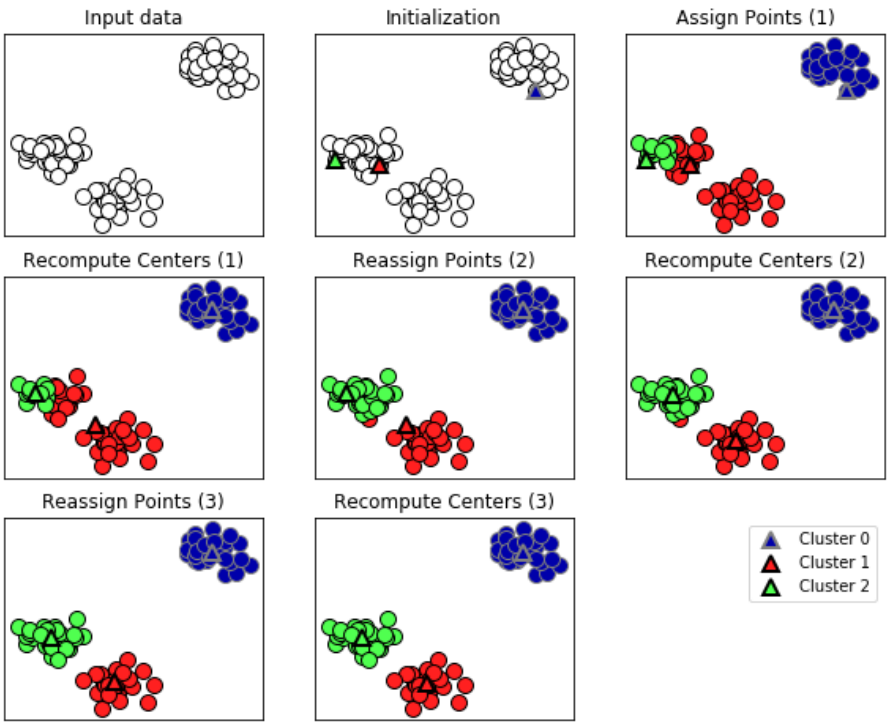

Figura 2.1: Etapas del algoritmo K-Means
Fuente: Müller, 2016, p.169

2.3.1. Limitaciones

Dado que cada cluster se define por su centro, esto quiere decir que cada cluster es de forma conexa, como resultado de esto *K-Means* solo captura formas “simples”, es decir, el algoritmo asume que todos los grupos tienen el mismo diámetro, por lo que al encontrarse con patrones más complejos no se obtendrá un buen resultado, como el que se observa en la figura 2.3.

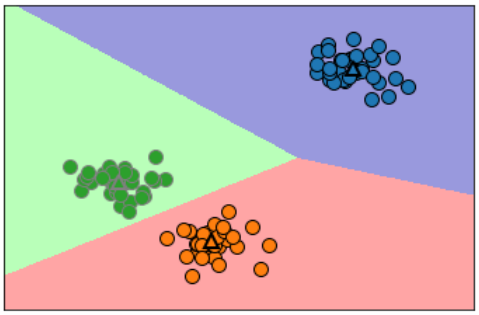

Figura 2.2: Centros y fronteras de los clusters definidos por K-Means
Fuente: Müller, 2016, p.170

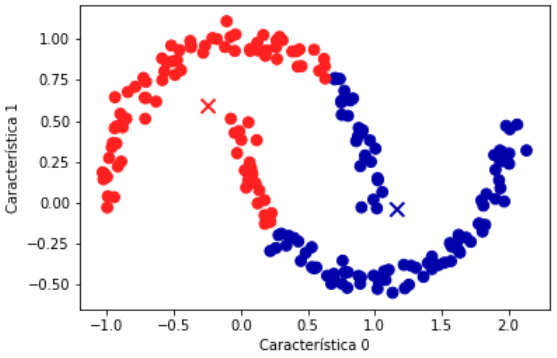

Figura 2.3: Desempeño de K-Means en patrones particulares
Fuente: Müller, 2016, p.170

2.3.2. Elección y validación del modelo

Una de las principales situaciones que se deben enfrentar a la hora de utilizar este algoritmo es elegir el valor de *K*, es decir, en cuántos grupos se deben dividir los datos, de tal forma que los miembros de cada cluster sean similares, pero muy diferentes a los miembros de otros.

Dentro de las formas más comunes para elegir el valor de *K* esta el *método del codo*, partiendo del hecho de que *K-Means* busca elegir los centroides que minimicen la inercia [15]:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (\| x_i - \mu_j \|)^2$$

Donde C es el conjunto de clusters, μ_j el centroide del cluster j y x_i las observaciones de un conjunto de datos X. Por lo que al dividir el conjunto de datos en K clusters con $k \in 1,2,3,\ldots,n$ se puede obtener un valor de inercia para cada valor de K.

Conociendo el número de clusters y los valores de la inercia para cada uno, se pueden elaborar gráficas como la que se observa en la figura 2.4, en la cual se observa que a partir de 5 clusters en adelante el valor de la inercia no cambia drásticamente comparada con los valores obtenidos utilizando 4 clusters o menos. Esto es debido a que al incrementar el número de cúmulos es equivalente a tener más centroides con menores datos asociados a éstos disminuyendo la distancia entre ellos. Una forma para entender dicho comportamiento es pensar en los casos extremos, el primero es cuando $K = 1$, es decir, el conjunto de datos en su estado natural, se va a elegir un solo centroide y se sumarán las distancias de todos los puntos a éste, este caso es en el que se alcanza el valor máximo de la inercia. Por otro lado está el caso donde $K = N$, donde N es la cardinalidad (número de observaciones) del conjunto de datos, por lo que cada observación será su propio centroide haciendo que no exista distancia que medir, por lo que el valor de la inercia será 0.

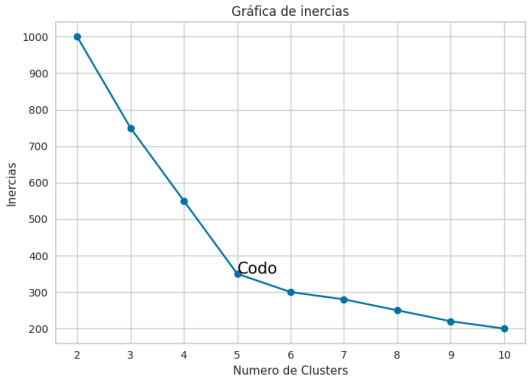

Figura 2.4: Ejemplo de gráfica de inercias para la elección del valor de K
Fuente: Elaboración propia

Regresando al *método del codo* para la elección del valor de K, utilizando la

gráfica del as inercias, se utiliza aquel donde se nota una disminución drástica del decrecimiento en el valor de las inercias (donde se dibuja un codo), para el ejemplo mostrado en la figura 2.4 se debería tomar $K = 5$.

Adicional al método anterior se puede utilizar el *Coeficiente Silueta* cuyos valores ofrecen una métrica para calificar la definición los clusters obtenidos. Para cálcular este coeficiente se necesitan hacer los siguientes cálculos [14]:

- a: La distancia media entre un punto y todos los demás puntos en el mismo cúmulo.
- b: La distancia media entre un punto y todos los demás puntos del cluster más cercano.

Una vez que se obtienen los valores de a y b, se realiza el cálculo del *coeficiente Silueta*:

$$S = \frac{b - a}{\text{máx}(a, b)}$$

Ya que no es posible obtener distancias negativas y suponiendo que $a = 0$, S toma el valor de 1, por otro lado, cuando $b = 0$ se obtiene $S = -1$, o lo que es lo mismo $-1 \leq S \leq 1$. Además se cuenta con descripciones particulares para los siguientes valores de S:

- $S = -1$: Los puntos están asignados erróneamente en un cluster
- $S = 0$: Los clusters se traslapan
- $S = 1$ Los puntos se encuentran perfectamente asignados

Capítulo 3

Análisis exploratorio y procesamiento de los datos

3 Análisis exploratorio y procesamiento de los datos

Posiblemente una de las etapas que requiere de un mayor enfoque de trabajo, es la del análisis exploratorio de los datos, por ser aquella donde se hace una revisión de todas la variables que conforman la base de datos, a pesar de aplicar técnicas que ayudan a resumir el comportamiento de las mismas, como las medidas de tendencia central y dispersión, algunas variables requerirán atención especial, en particular cuando existan casos con datos faltantes. Posteriormente y como se podrá observar más adelante, se analiza la correlación y distribución de las variables para poder seleccionar las que brinden mayor información para el problema que se desea abordar. Por último, se aplican técnicas de escalamiento de los datos y al mismo tiempo se proponen tres propuestas del tratamiento de los datos con la finalidad de elegir aquel que contribuya a una mejor segmentación de los datos.

3.1. Análisis exploratorio

Partimos de una base inicial que cuenta con 8,399 observaciones y 12 variables que se enlistan y describen en la tabla 3.1, esta base publicada en Kaggle[1] (sitio en el que se publican bases de datos para que la comunidad realice proyectos de ciencia de datos, además de publicar competiciones para implementar modelos de aprendizaje automático, deep learning, etc. para resolver problemas en específico a cambio de premios económicos) originalmente se conformó con la finalidad de modelar la morosidad de los clientes de un banco peruano para revelar los hallazgos que permitan a la institución de mejores herramientas al momento de otorgar créditos.

Una vez que se tienen las variables identificadas, se inicia el análisis de exploración de datos, comenzando con la identificación de valores faltantes. En este caso

[1] Bank Defaul Analysis: https://www.kaggle.com/datasets/luishcaldernb/morosidad

Variable	Descripción
vivienda	Tipo de vivienda
edad	Edad del cliente
dias_lab	Días trabajados en el empleo actual
exp_sf	Número de meses del cliente desde que adquirió un producto financiero
nivel_ahorro	Clasificación del ahorro del cliente, 0 para los que no tiene ahorros y 12 para los que tienen un nivel alto de ahorro
ingreso	Monto de ingreso del cliente
linea_sf	Línea de crédito del cliente
deuda_sf	Monto de deuda del cliente
score	Calificación crediticia del cliente, a mayor número, mejor perfil crediticio
zona	Lugar de residencia del cliente
clasif_sbs	Calificación otorgada por el regulador del país (SBS), 0: normal, 1: con problemas potenciales, 2: deficiente, 3: dudoso y 4: pérdida
nivel_educ	Máximo nivel educativo del cliente

Cuadro 3.1: Variables y su definición
Fuente: Elaboración propia

solo 3 variables presentan dicha situación como lo muestra la tabla 3.2.

Variable	Número de observaciones faltantes
exp_sf	1,830
linea_sf	1,127
deuda_sf	461

Cuadro 3.2: Variables con datos faltantes
Fuente: Elaboración propia

Un punto adicional que se debe contemplar es la correlación que existe entre las variables, el análisis se desarrolló solo con aquellas variables que tengan una correlación < 0.40 entre ellas. Esto para eliminar aquellas variables que proporcionan la misma información al modelo; la matriz de correlación se puede observar en la figura 3.1.

Por lo que en este punto se contempla el uso de las siguientes variables (aquellas que no se encuentran en la matriz de correlación son variables categóricas):

	edad	dias_lab	exp_sf	ingreso	linea_sf	deuda_sf	score
edad	1.000000	0.480000	0.360000	0.300000	0.280000	0.110000	0.430000
dias_lab	0.480000	1.000000	0.350000	0.230000	0.190000	0.050000	0.270000
exp_sf	0.360000	0.350000	1.000000	0.420000	0.500000	0.210000	0.380000
ingreso	0.300000	0.230000	0.420000	1.000000	0.530000	0.290000	0.380000
linea_sf	0.280000	0.190000	0.500000	0.530000	1.000000	0.360000	0.320000
deuda_sf	0.110000	0.050000	0.210000	0.290000	0.360000	1.000000	0.200000
score	0.430000	0.270000	0.380000	0.380000	0.320000	0.200000	1.000000

Figura 3.1: Coeficiente de correlación
Fuente: Elaboración propia

- vivienda
- dias_lab
- nivel_ahorro
- linea_sf
- deuda_sf
- score
- zona
- clasif_sbs
- nivel_educ

Otro aspecto a revisar en esta fase son las medidas de tendencia central y de dispersión de los datos, para poder ver sus características más generales, las cuales se pueden observar en el cuadro 3.3 (se omiten las medidas de las variables vivienda, nivel_ahorro, zona, clasif_sbs y nivel_educ por ser variables categóricas, pero se abordan posteriormente).

Por otro lado en cuanto a las variables categóricas se tiene la siguiente información:

De acuerdo a lo que se observa en el cuadro 3.3, las variables **linea_sf** y **deuda_sf** tienen datos faltantes, ya que su conteo no llega a las 8,399. Para corregir esta situación se hizo la imputación de la mediana.

Medida	dias_lab	linea_sf	deuda_sf	score
Conteo	8,399	7,272	7,938	8,399
Media	5,556	11,987	6,111	197
std	2,153	21,323	11,178	20
min	2,956	0	0	134
25%	4,174	1,169	478	182
50%	4,904	4,030	2,259	197
75%	6,182	12,087	5,755	212
max	20,700	121,543	57,094	266

Cuadro 3.3: Medidas de tendencia central y dispersión de las variables numéricas
Fuente: Elaboración propia

Medida	vivienda	nivel_ahorro	zona	clasif_sbs	nivel_edu
Conteo	8,399	8,399	8,399	8,399	8,399
Únicos	3	13	25	5	4
Top	FAMILIAR	12	Lima	0	UNIVERSITARIA
Frec.	5,853	6,619	4,980	5,282	4,802

Cuadro 3.4: Resumen de las variables categóricas
Fuente: Elaboración propia

En cuanto a las variables categóricas, se presentan las siguientes gráficas para obtener más información de cómo se distribuyen en cada una de sus categorías.

Dado lo que se observa en la figura 3.2, se destaca que la mayoría de los clientes de este banco cuentan con una vivienda familiar.

En cuanto a los niveles de ahorro, en la figura 3.3 podemos ver que casi la totalidad de los clientes caen en un nivel de ahorro 12.

En la figura 3.4 se puede observar que la mayoría de los clientes se concentran en la zona Lima.

Para le caso de la clasificación que otorgó el regulador, de acuerdo a lo que se observa en la figura 3.5, se tienen menos categorías y la mayoría tienen una calificación de 0: "normal".

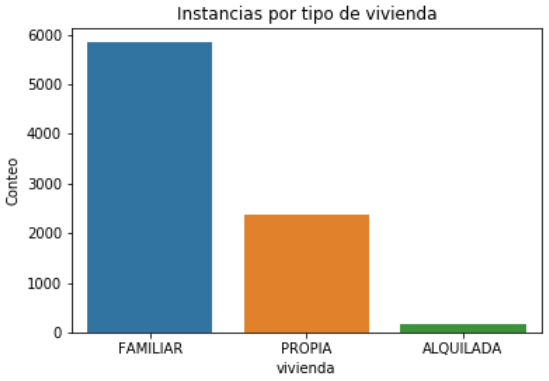

Figura 3.2: Distribución de los datos por tipo de vivienda

Fuente: Elaboración propia

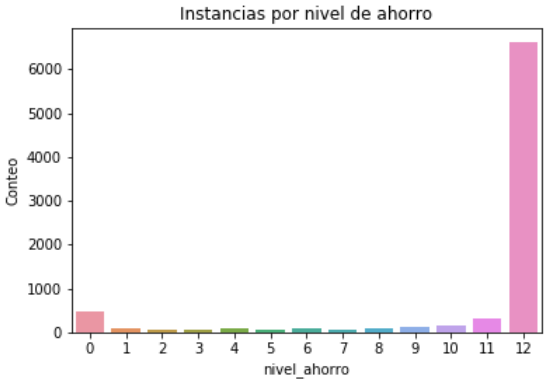

Figura 3.3: Distribución de los datos por nivel de ahorro

Fuente: Elaboración propia

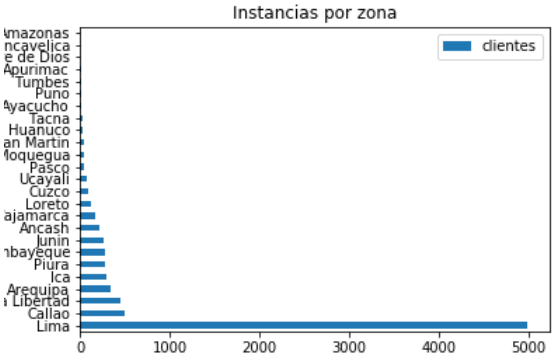

Figura 3.4: Distribución de los datos por zona

Fuente: Elaboración propia

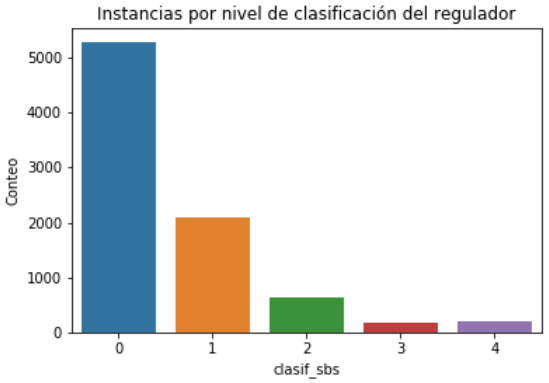

Figura 3.5: Distribución de los datos por clasificación del regulador

Fuente: Elaboración propia

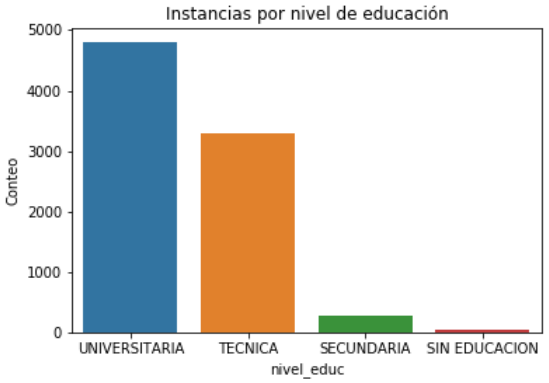

Figura 3.6: Distribución de los datos por nivel educativo

Fuente: Elaboración propia

De las gráficas anteriores principalmente se deben resaltar las variables **zona** y **nivel_ahorro** que cuentan con demasiadas categorías y al mismo tiempo la mayoría de los datos caen dentro de pocas de éstas, lo cual indica que no van a contribuir información relevante al momento de aplicar el algoritmo, por lo que para los próximos pasos descritos en este análisis serán omitidas.

A continuación se presentarán los histogramas de las variables numéricas para ver la forma en la que se distribuyen:

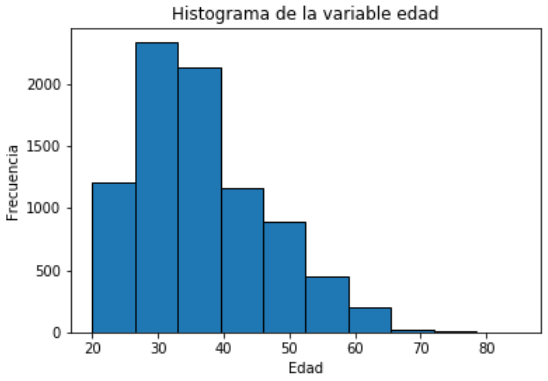

Figura 3.7: Distribución de los datos por edad

Fuente: Elaboración propia

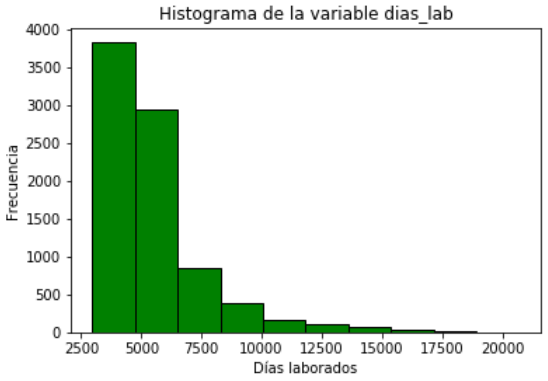

Figura 3.8: Distribución de los datos por días laborados

Fuente: Elaboración propia

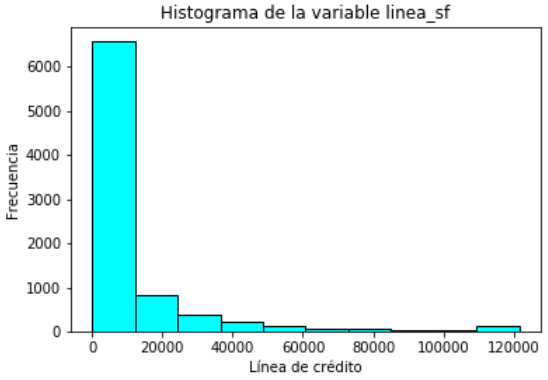

Figura 3.9: Distribución de los datos por línea de crédito

Fuente: Elaboración propia

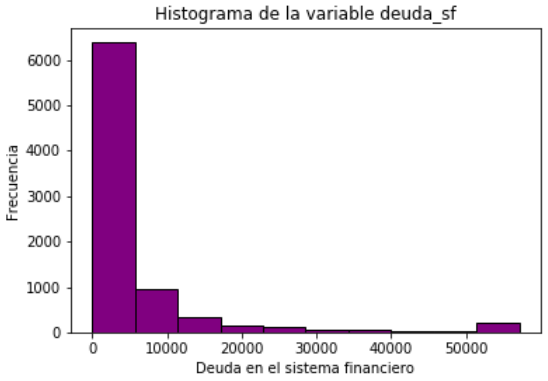

Figura 3.10: Distribución de los datos por deuda del cliente en el sistema financiero

Fuente: Elaboración propia

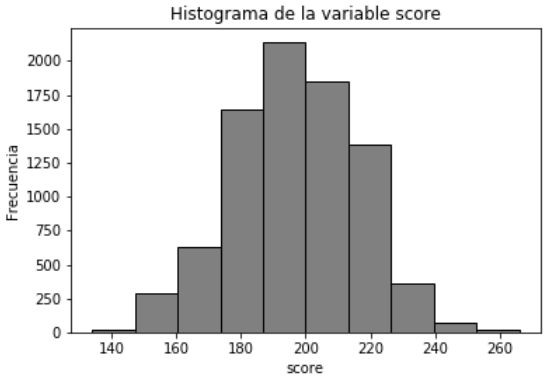

Figura 3.11: Distribución de los datos por score

Fuente: Elaboración propia

Como se puede observar en los histogramas anteriores, tres de las variables numéricas (figuras 3.8, 3.9 y 3.10) están sesgadas a la derecha, es decir, existen datos atípicos; dada esta situación y el número de variables con las que se cuentan se decidió abordar el análisis realizando los siguientes experimentos:

1. Segmentación con las variables seleccionadas imputando datos atípicos
2. Segmentación eliminando los registros con datos atípicos
3. Segmentación utilizando solo variables numéricas

La intención de realizar ese ejercicio fue para ver el comportamiento del algoritmo *K-Means* bajo esas condiciones y seleccionar el mejor escenario utilizando la

métrica **Silueta**.

3.2. Preprocesamiento

Dados los experimentos que se planteó realizar, se requirió de la preparación de las siguientes bases de datos:

- *base_sin_outliers*: Base de datos con todas las variables y se eliminaron los registros donde alguna de las variables tuviera algún valor atípico
- *base_imp*: Base de datos con todas las variables y se imputó el valor de la mediana correspondiente para aquellos valores considerados como atípicos
- *base*: Base de datos con un subconjunto de las variables tomadas de la base a la cual se le eliminaron variables correlacionadas y se le imputó la mediana correspondiente a los valores faltantes de cada variable

Para la identificación de datos atípicos se utilizaron las fórmulas *límite inferior* = $Q_1 - 1.5 * IQR$ y *límite superior* $= Q_3 + 1.5 * IQR$ donde Q_1 y Q_3 son el primer y tercer cuartil respectivamente e $IQR = Q_3 - Q_1$ el rango intercuartil. El el cuadro 3.5 se presentan los valores que se obtuvieron para cada variable.

Variable	**Límites**
dias_lab	1,162 - 9,194
linea_sf	0 - 22,402
deuda_sf	0 - 13,066
score	137 - 257

Cuadro 3.5: Límites para la identificación de valores atípicos
Fuente: Elaboración propia

Una vez que se conocen los límites para cada variable se generaron las bases mencionadas previamente, eliminando registros e imputando valores. Las medidas de tendencia central de *base*, *base_sin_outliers* y *base_imp* se encuentran en los cuadros 3.6, 3.7 y 3.8 respectivamente.

Medida	dias	linea_sf	deuda_sf	score
Conteo	8,399	8,399	8,399	8,399
Media	5,556	11,987	6,111	197
std	2,153	21,323	11,178	18
min	2,956	0	0	134
25%	4,174	1,169	478	182
50%	4,904	4,030	2,259	197
75%	6,182	12,087	5,755	212
max	20,700	12,1543	57,094	266

Cuadro 3.6: Medidas de tendencia central y dispersión de las variables numéricas
Fuente: Elaboración propia

Medida	dias	linea_sf	deuda_sf	score
Conteo	6,444	6,444	6,444	6,444
Media	5,061	4,563	2,484	193
std	1,234	4,612	2,812	19
min	2,956	0	0	146
25%	4,113	1,122	139	178
50%	4,721	3,960	1,631	194
75%	5,725	5,822	3,797	205
max	9,165	22,341	13,056	254

Cuadro 3.7: Medidas de tendencia central y dispersión de las variables numéricas eliminando datos atípicos

Fuente: Elaboración propia

3.2.1. Codificación de variables categóricas

Para esta situación se utilizó la técnica **Label encoder** sobre las variables **vivienda** y **nivel_educ**, la cual asigna un número entero a cada una de las categorías quedando

del siguiente modo:

- **vivienda**
  - 0: SECUNDARIA, 1: SIN EDUCACION, 2: TECNICA, 3: UNIVERSITARIA
- **nivel_educ**
  - 0: FAMILIAR, 1: PROPIA, 2: ALQUILADA

Medida	dias	linea_sf	deuda_sf	score
Conteo	8,399	8,399	8,399	8,399
Media	5,106	4,653	2,442	197
std	1,247	4,474	2,745	19
min	2,956	0	0	146
25%	4,174	1,500	263	182
50%	4,782	4,030	1,632	197
75%	5,725	5,249	3,507	212
max	9,165	22,341	13,056	257

Cuadro 3.8: Medidas de tendencia central y dispersión de las variables numéricas con datos atípicos imputados

Fuente: Elaboración propia

3.2.2. Escaladores

Como parte del procesamiento, para las variables numéricas particularmente, se aplicaron los escaladores **StandardScaler** y **RobustScaler** con la finalidad de reducir la escala de las variables e identificar mejoras en el desempeño del algoritmo. En cuanto al **StandardScaler** sabemos que se asegura de que la información escalada tenga media 0 y varianza 1 [17]. Del mismo modo **RobustScaler** se asegura de que los datos estén en la misma escala, pero a diferencia de la técnica anterior, éste usa la

mediana y los cuartiles en lugar de la media y la varianza [16]. En la figura 3.12 se puede observar como cambian los datos aplicando los escaladores mencionados.

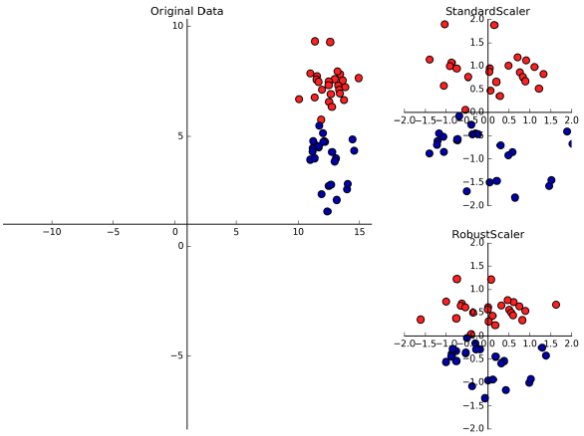

Figura 3.12: Transformación usando StandardScaler y RobustScaler

Fuente: Müller, 2016, p.133

En este punto se cuentan con todos los elementos para la elaboración de los tres escenarios mencionados previamente. Dicho ejercicio se presenta en el siguiente capítulo.

Capítulo 4

Segmentación de clientes con el algoritmo K-Means

4 Segmentación de clientes con el algoritmo K-Means

En este apartado se presentan los resultados obtenidos de la implementación del algoritmo utilizando las tres bases de datos descritas previamente en el capítulo 3 (*base_sin_outliers*, *base_imp* y *base*) aplicandoles los escaladores **StandardScaler** y **RobustScaler**, además de una reducción de dimensionalidad a dos componentes utilizando la técnica de análisis de componentes principales *PCA*, esto último con la finalidad de poder visualizar el resultado del algoritmo seleccionado en una gráfica de dos dimensiones identificando la asignación de los datos a cada uno de los clusters y finalizar con la descripción de éstos.

4.1. Resultados de la generación de los escenarios

Para la elección del número de cúmulos así como el escalador a aplicar, se calculó el coeficiente *Silueta* para cada uno de los escaladores y conjuntos de datos para una segmentación de 2 a 5 grupos, dicha delimitación obedece a la naturaleza del trabajo, es decir, al ser una solución estratégica empresarial resulta complicado, desde una perspectiva comercial, dar seguimiento a numerosos segmentos, es por ello que se busca tener el número mínimo de grupos diferenciados a atender.

Los valores del coeficiente *Silueta* obtenidos se pueden observar en el cuadro 4.1. Para este ejercicio, las variables seleccionadas en el conjunto de datos **base** fueron **dias_lab**, **linea_sf** y **deuda_sf**.

De acuerdo con lo visto en el capítulo 2, en cuanto más se acerque el coeficiente *Silueta* al valor de 1 significa que los cúmulos se encuentran diferenciados de mejor forma, es por ello que considerando los datos presentados en el cuadro 4.1 se puede observar que el mejor resultado se obtiene utilizando el conjunto de datos **base** (que cuenta con las variables **dias_lab**, **linea_sf** y **deuda_sf**) escalando con

Núm. Clusters	base_sin_outliers		base_imp		base	
	Standard	Robust	Standard	Robust	Standard	Robust
2	0.356	0.456	0.360	0.528	0.694	0.779
3	0.412	0.496	0.394	0.437	0.622	0.782
4	0.426	0.369	0.394	0.359	0.468	0.677
5	0.384	0.389	0.386	0.350	0.485	0.644

Cuadro 4.1: Valores del Coeficiente Silueta para cada base con diferentes números de clusters

Fuente: Elaboración propia

RobustScaler y segmentando en 3 clusters. Por otro lado, observando la gráfica 4.1 que muestra las inercias para este escenario, no se logra identificar un *codo*, es por ello que la segmentación se realizó conforme a la información que proporcionó el coeficiente *Silueta*.

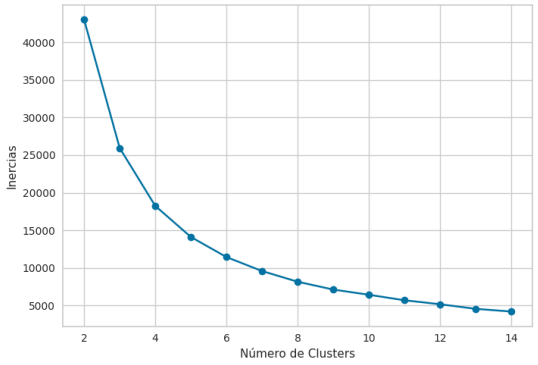

Figura 4.1: Inercias del modelo con las variables **dias_lab**, **linea_sf** y **deuda_sf**

Fuente: Elaboración propia

4.1.1. Modelo seleccionado

El modelo definitivo para segmentar a los clientes de la base de datos en cuestión, utilizará las variables **dias_lab**, **linea_sf** y **deuda_sf**, previamente procesadas y escaladas con **RobustScaler**. Aprovechando que los datos se redujeron a dos dimensiones, en la gráfica 4.2 se puede observar la asignación de los clientes a cada uno de los 3 clusters.

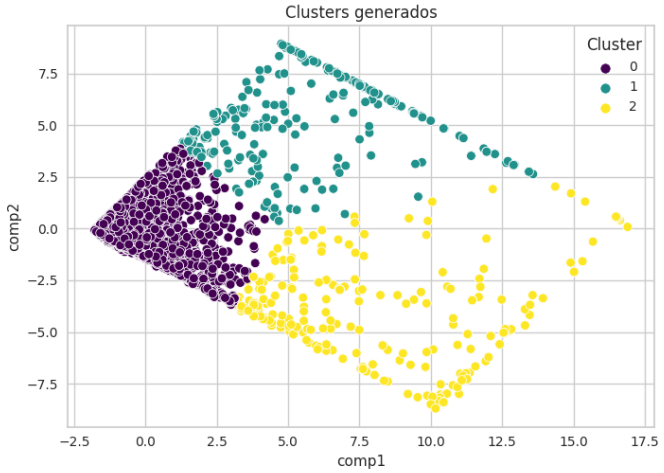

Figura 4.2: Datos segmentados por Cluster, utilizando las primeras dos componentes principales (comp1 y comp2) obtenidas por el análisis de componentes principales (PCA)

Fuente: Elaboración propia

4.1.2. Descripción de los clusters

Una de las tareas más importantes que se deben realizar una vez obtenidos los grupos deseados, es realizar la descripción de éstos, es decir, encontrar las principales características que los definen y con ellas poder dirigir acciones personalizadas a cada uno de los grupos.

Al revisar el tamaño de los cúmulos, en el cuadro 4.2 se resalta que el cluster 0 concentra el $90\,\%$ de las observaciones y solo un $5\,\%$ se encuentra asignado en cada uno de los otros dos grupos, a la par de eso, en le cuadro 4.3 se puede observar que el cluster 0 cuenta con los promedios más bajos comparado con los otros dos, además en la experiencia de este autor, este comportamiento es común en el ámbito financiero, ya que por lo general pocos clientes cuentan con un monto significativo de ingresos, ahorros o línea de crédito, mientras la mayor parte de los clientes tienden a tener montos menores en los rubros mencionados.

Con apoyo, principalmente del cuadro 4.3, se encontraron las siguientes características en cada uno de los clusters, además se les proporcionó un nombre para

Cluster	Núm. Observaciones	Proporción
0	7,596	90%
1	421	5%
2	382	5%

Cuadro 4.2: Distribución de observaciones por cada cluster
Fuente: Elaboración propia

Cluster	dias_lab	linea_sf	deuda_sf
0	5,456	6,425	3,212
1	5,752	24,135	44,486
2	7,320	85,728	14,089

Cuadro 4.3: Promedio en cada cluster de las variables utilizadas
Fuente: Elaboración propia

futuras referencias a éstos:

- **General (Cluster 0)**: Contiene a los clientes con el menor promedio de días laborados, línea de crédito y monto de deuda, este cluster concentra el 90% de los clientes analizados
- **Medio (Cluster 1)**: Clientes con un promedio de días laborados ligeramente mayor a los clientes del cluster **General** pero con una línea de crédito *6 veces más grande* y con el promedio de deuda más grande
- **Premium (Cluster 2)**: Clientes con el promedio más alto en la línea de crédito y con un nivel de deuda poco más de *4 veces* la del cluster **General**

Capítulo 5

Solución estratégica

5 Solución estratégica

En esta última sección se presentan las acciones comerciales que se recomiendan para cada grupo definido, además de la implementación de un árbol de decisiones, algoritmo de aprendizaje supervisado, para clasificar a futuros clientes e identificar en automático la oferta de productos o servicios que se les deberá hacer.

5.1. Oferta comercial

Ahora que ya se conocen las características de los clientes que sirvieron para agruparlos, se definieron las siguientes acciones para cada grupo, con la intención de desarrollarlos y retenerlos (teniendo en cuenta que 2 de las 3 variables que se utilizan están relacionadas al crédito):

- **General**: Al ser el grupo más grande y con menor línea de crédito y de deuda, la recomendación es mantener las condiciones actuales y ofrecer productos de ahorro.
- **Medio**: Inicialmente ofrecer productos de ahorro. Al ser el grupo con el promedio de deuda más grande, la institución deberá analizar la morosidad de los clientes que sean asignados a este grupo y posteriormente aquellos que cuenten con las condiciones pertinentes, ofrecerles un aumento de línea de crédito.
- **Premium**: El objetivo principal de la institución con respecto a los clientes asignados a este grupo será retenerlos con programas de fidelización y promociones como descuentos y compras a meses sin intereses, además de ofrecer productos de ahorro e inversión, ya que son los clientes con la línea de crédito más grande, pero lo más importante, con una deuda moderada, lo que podría deberse a una morosidad baja.

5.2. Clasificador

En esta etapa se utilizaron las etiquetas asignadas por el algoritmo *K-Means* como la variable objetivo para entrenar un *Árbol de decisiones*, con una profundidad máxima de 3 niveles y un $20\,\%$ de los datos (1,680) en el conjunto de prueba. En la figura 5.1 se puede observar el árbol generado para esta clasificación.

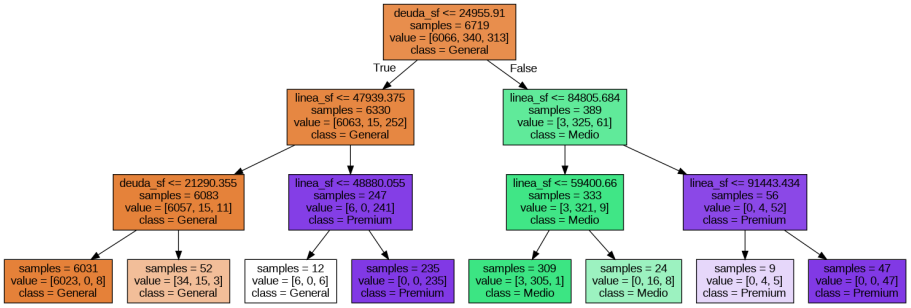

Figura 5.1: Visualización del árbol generado

Fuente: Elaboración propia

De acuerdo a los resultados obtenidos en la métrica *Accuracy* que se presentan en el cuadro 5.1, se puede apreciar que los datos se están clasificando de forma correcta, pero se debe considerar el hecho de que se está trabajando con clases desbalanceadas, recordando que el grupo **General** concentra el $90\,\%$ de los clientes, por lo que si los clientes de los otros dos grupos son clasificados incorrectamente, no se lograría ver un impacto significativo en el *Accuracy*. Dicha situación se puede descartar revisando los valores de las métricas *Precision* y *Recall*, las cuales se pueden observar en el cuadro 5.2.

Datos	***Accuracy***
Entrenamiento	$99.29\,\%$
Prueba	$99.05\,\%$

Cuadro 5.1: *Accuracy* de los conjuntos de datos de entrenamiento y prueba
Fuente: Elaboración propia

Tomando en cuenta que la métrica *Precision* indica la proporción de aciertos al clasificar un elemento de forma correcta y *Recall* mide las observaciones de

Grupo	*Precision*	*Recall*
General	100 %	100 %
Medio	92 %	94 %
Premium	97 %	91 %

Cuadro 5.2: Métricas *Precision* y *Recall* del conjunto de datos de prueba
Fuente: Elaboración propia

una clase asignadas correctamente a una clase [12], es posible afirmar que incluso los clientes pertenecientes a los grupos **Medio** y **Premium** en su mayoría se logran clasificar correctamente.

Dados los resultados obtenidos, el clasificador implementado formará parte de la solución estratégica para la clasificación de nuevos clientes y asociación de productos y servicios.

5.2.1. Ejemplo práctico de clasificación

Para mostrar la clasificación de los clientes a alguno de los tres grupos definidos, del conjunto de prueba se tomaron al azar tres clientes, cuyos valores para cada variable se pueden observar en el cuadro 5.3.

Cliente	*dias_lab*	*linea_sf*	*deuda_sf*
1	5,725	5,828	0
2	4,874	22,580	1,948
3	8,374	40,481	48,629

Cuadro 5.3: Valores redondeados de las variables de los clientes seleccionados para el ejemplo de clasificación
Fuente: Elaboración propia

También se creó una clase en Python llamada `clasificacion(modelo)`, que recibe como argumento el modelo previamente entrenado y cuenta con el método `asignacion(array)` cuyo argumento recibe el arreglo numérico que contiene los valores de las variables de los clientes que se quieren clasificar, dicho método devuelve una tabla con los datos ingresados, el grupo al que fue asignado, así como una descripción corta de la acción recomendada. Lo anterior se puede observar gráficamente

en la figura 5.2.

```
ej = clasificacion(model)
```

```
ej.asignacion(ejemplo)
```

	dias_lab	linea_sf	deuda_sf	Grupo	Acción
0	5725	5827.55	0.00	General	Ahorro
1	4874	22580.00	1947.81	General	Ahorro
2	8374	40481.20	48628.55	Medio	Ahorro y evaluación de línea de crédito

Figura 5.2: Visualización del grupo asignado además de la acción recomendada

Fuente: Elaboración propia

De este modo, al contar con la información que se presenta en la figura 5.2, resulta sencillo realizar las recomendaciones definidas, optimizando la asignación de recursos humanos que atenderán a los clientes, así como la de las distintas estrategias que se pudieran asignar en el futuro para cada uno de los grupos definidos.

Conclusiones

Conclusiones

Con lo expuesto a lo largo de este trabajo, hemos establecido un proceso básico para la segmentación de información proveniente de bases de datos de clientes bancarios que nos permite asociar acciones para una atención mejor dirigida, cumpliendo con lo establecido en los objetivos del proyecto. Se destaca la importancia del análisis y procesamiento de los datos, pasando por una etapa muy relevante, la selección de variables, en donde generalmente se buscaba construir un modelo que utilice la mayoría de las variables disponibles, sin embargo esto no siempre es posible o lo más óptimo, ya que de acuerdo a las métricas de evaluación de segmentación para este caso, aquellos con más variables no presentaron lo resultados más consistentes. Sin embargo fue posible ajustarlo utilizando un grupo reducido de características a los cuales se les pudo asignar una oferta de productos y servicios acorde a las características utilizadas.

Tal y como se mencionó al inició de este trabajo, resulta fundamental que las empresas cuenten con el conocimiento del perfil de clientes que conforman su cartera, ya sea por razones regulatorias o comerciales, derivado del análisis realizado, la institución bancaria puede tomar mejores decisiones y optimizar el uso de recursos económicos como los presupuestos para las campañas y los recursos humanos como los ejecutivos que tienen el contacto directo con el cliente. Gracias a las métricas de evaluación, se puede destacar la calidad del modelo definido, el cual además puede adaptarse para incorporar una o más variables que se quisieran considerar en el análisis; en contraste a lo realizado en este trabajo, al momento de descartar variables, se debe tener en mente que derivado de la naturaleza de la información, al ser una base de datos de acceso público y resaltando la sensibilidad inherente a la misma, resulta complicado conseguir más datos que ayuden a robustecer el modelo. Por último se debe tomar en cuenta que las variables utilizadas para esta segmentación, están relacionadas al crédito y aunque el principal negocio de una institución bancaria se deriva del otorgamiento de crédito, no se debe perder de vista la captación de los recursos de los clientes, por lo que desatender dicho rubro impactaría directamente

en la capacidad de la institución para otorgar créditos así como el incumplimiento de algunas normas regulatorias.

Bibliografía

[1] ALKHAYRAT, M., ALJNIDI, M., AND ALJOUMAA, K. A comparative dimensionality reduction study in telecom customer segmentation using deep learning and pca. *Journal of Big Data 7* (2020), 1–23.

[2] ALPAYDIN, E. *Introduction to machine learning*. MIT press, 2020.

[3] AN, J., KWAK, H., JUNG, S.-G., SALMINEN, J., AND JANSEN, B. J. Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data. *Social Network Analysis and Mining 8*, 1 (2018), 54.

[4] ARYUNI, M., MADYATMADJA, E. D., AND MIRANDA, E. Customer segmentation in xyz bank using k-means and k-medoids clustering. In *2018 International conference on information management and technology (ICIMTech)* (2018), IEEE, pp. 412–416.

[5] BALDEÓN, L. H. C. Bankdefaultanalysis, 2021. Accedido en junio de 2023.

[6] BAUTISTA RUEDA, F. R., ET AL. Análisis de la aplicación de la segmentación en el sector financiero. Master's thesis, Maestría en Administración de Empresas-MBA–Virtual, 2021.

[7] DONG, D., ZHANG, J., AND YE, J. Research on customer segmentation method of commercial bank based on data mining. In *3rd International Conference on Innovation Development of E-commerce and Logistics (ICIDEL 2017)* (2017), pp. 62–65.

[8] ELGUERA VEGA, R. M. Segmentación de clientes de un casino utilizando el algoritmo partición alrededor de medoides (pam) con datos mixtos.

[9] KAGGLE. Bankdefaultanalysis. `https://www.kaggle.com/datasets/luishcaldernb/morosidad`. Accedido en junio de 2023.

[10] KILARI, H., EDARA, S., YARRA, G. R. S., AND GADHIRAJU, D. V. Customer segmentation using k-means clustering. *International Journal of Engineering Research & Technology (IJERT) 11*, 03 (2022), 303–208.

[11] MIHOVA, V., AND PAVLOV, V. A customer segmentation approach in commercial banks. In *AIP conference proceedings* (2018), vol. 2025, AIP Publishing LLC, p. 030003.

[12] MÜLLER, A. C., AND GUIDO, S. *Introduction to machine learning with Python: a guide for data scientists*. "O'Reilly Media, Inc.", 2016.

[13] PROVOST, F., AND FAWCETT, T. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. "O'Reilly Media, Inc.", 2013.

[14] ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20* (1987), 53–65.

[15] SCIKIT-LEARN. 2.3 clustering. `https://scikit-learn.org/stable/modules/clustering.html`. Accedido en junio de 2023.

[16] SCIKIT-LEARN. sklearn.preprocessing.robustscaler. `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html#sklearn.preprocessing.RobustScaler`. Accedido en junio de 2023.

[17] SCIKIT-LEARN. sklearn.preprocessing.standardscaler. `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`. Accedido en junio de 2023.

[18] SMEUREANU, I., RUXANDA, G., AND BADEA, L. M. Customer segmentation in private banking sector using machine learning techniques. *Journal of Business Economics and Management 14*, 5 (2013), 923–939.

[19] TIMARÁN-PEREIRA, S., HERNÁNDEZ-ARTEAGA, I., CAICEDO-ZAMBRANO, S., HIDALGO-TROYA, A., AND ALVARADO-PÉREZ, J. El proceso de descubrimiento de conocimiento en bases de datos. *Descubrimiento de patrones de desem-*

peño académico con árboles de decisión en las competencias genéricas de la formación profesional (2016), 63–86.

[20] VOICAN, O. Using data mining methods to solve classification problems in financial-banking institutions. *Economic Computation & Economic Cybernetics Studies & Research 54*, 1 (2020).