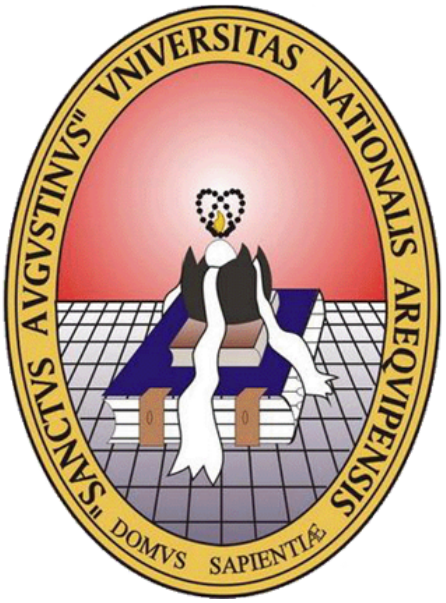
Universidad Nacional De San Agustín

Facultad de Producción y Servicios

Maestría en Ciencias de la Computación



Tema: Proyecto final del curso

Curso: Recuperación de Información

Presentado por:

Weimar Ccapatinta Huamani

Docente: ANA MARIA CUADROS VALDIVIA

2023

Problema:
El problema hace referencia a la necesidad de clasificar a los clientes en diferentes grupos o segmentos con características similares que permita:

- Personalización de productos y servicios, La entidad financiera puede querer ofrecer productos y servicios personalizados que se adapten a las necesidades y preferencias de diferentes segmentos de clientes.

Objetivos:

Encontrar la segmentación de clientes con características y comportamientos similares
Desarrollar perfiles de cliente detallados para cada segmento.

Preparar los datos de clientes: El objetivo es recopilar y preparar los datos relevantes de los clientes.
Aplicar técnicas de clustering: Se aplicará técnicas de clustering como k-means, a los datos de clientes para agrupar a los clientes en segmentos con características y comportamientos similares.

Introducción.

La segmentación de clientes en el ámbito bancario es una estrategia fundamental para comprender mejor las características y comportamientos financieros de los clientes. Al utilizar técnicas de segmentación, como el algoritmo de clustering K-means, podemos agrupar a los clientes en diferentes segmentos con características similares.

En este estudio, nos enfocamos en la segmentación de clientes de un banco utilizando el algoritmo K-means. El objetivo es identificar patrones y tendencias en los datos financieros de los clientes para mejorar la personalización de los productos y servicios bancarios, así como para diseñar estrategias de marketing más efectivas.

Además, al utilizar el algoritmo K-means, podemos identificar segmentos de alto valor, segmentos con mayor riesgo crediticio o segmentos con potencial de crecimiento. Esto nos brinda información valiosa para la toma de decisiones estratégicas en términos de retención de clientes, adquisición de nuevos clientes y diseño de campañas de marketing personalizadas.

Trabajos relacionados.

- Credit Card Holders Segmentation Using K-mean Clustering with Autoencoder
- Designing Progressive and Interactive Analytics Processes for High-Dimensional Data Analysis
- Using Unsupervised Machine Learning Techniques for Behavioral-based Credit Card Users Segmentation in Africa
- Credit Users Segmentation for improved Customer Relationship Management in Banking

Resumen

La segmentación de clientes es un enfoque comúnmente utilizado en el campo del análisis de datos y el marketing para agrupar a los clientes en diferentes segmentos basados en características similares. El algoritmo K-means es una técnica de clustering popular que se puede aplicar para realizar esta segmentación.

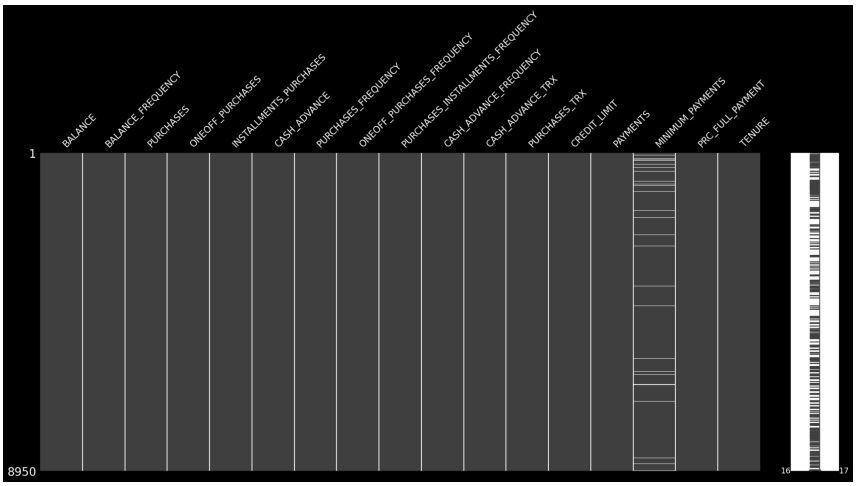
Descripción y preprocesamiento de datos

El conjunto de datos “Credit Card Dataset for Clustering” resume el comportamiento de uso de aproximadamente 9000 titulares activos de tarjetas de crédito durante los últimos 6 meses. El archivo está a nivel de cliente con 18 variables de comportamiento.

Para el siguiente paso de análisis es necesario tener los datos limpios y listos para ser usados por ellos es importante:

- Eliminar variables que no sean relevantes en este caso la variable “CUST_ID” el cual no es relevante para la segmentación ya que es un dato que va incrementando conforme se registren nuevos clientes.
- Eliminar los valores atípicos, en este caso consideramos los outliers como característica importante de los clientes que serán útiles para la segmentación.

- Imputar los valores faltantes en este caso para las variables “CREDIT_LIMIT” “MINIMUM_PAYMENTS” el cual presenta valores nulos según la imagen se encuentran a lo largo de los registros de los clientes por ello la imputación se realizará utilizando la mediana.



Análisis de datos.

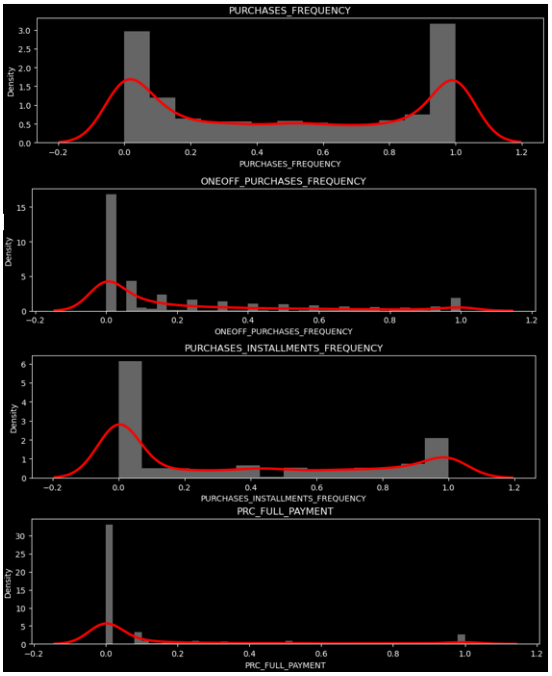
Esta etapa es considerada crucial ya que en ella aprenderemos el comportamiento de nuestros datos, con la ayuda de los siguientes métodos: análisis exploratorio de datos que incluye análisis univariado y multivariado, y las medidas de tendencia central.

Análisis univariado: a continuación un breve resumen estadístico de las variables.

- *El balance medio es $1564*
- *La frecuencia del balance se actualiza muy a menudo*
- *El promedio de las compras es $1000*
- *El importe máximo de compra no recurrente es en promedio $600*
- *El promedio de la frecuencia de las compras está cerca de 0.5*
- *El promedio del límite de crédito es $4500*
- *El porcentaje de pago completo es 15%*
- *Los clientes llevan de promedio en el servicio 11 años*

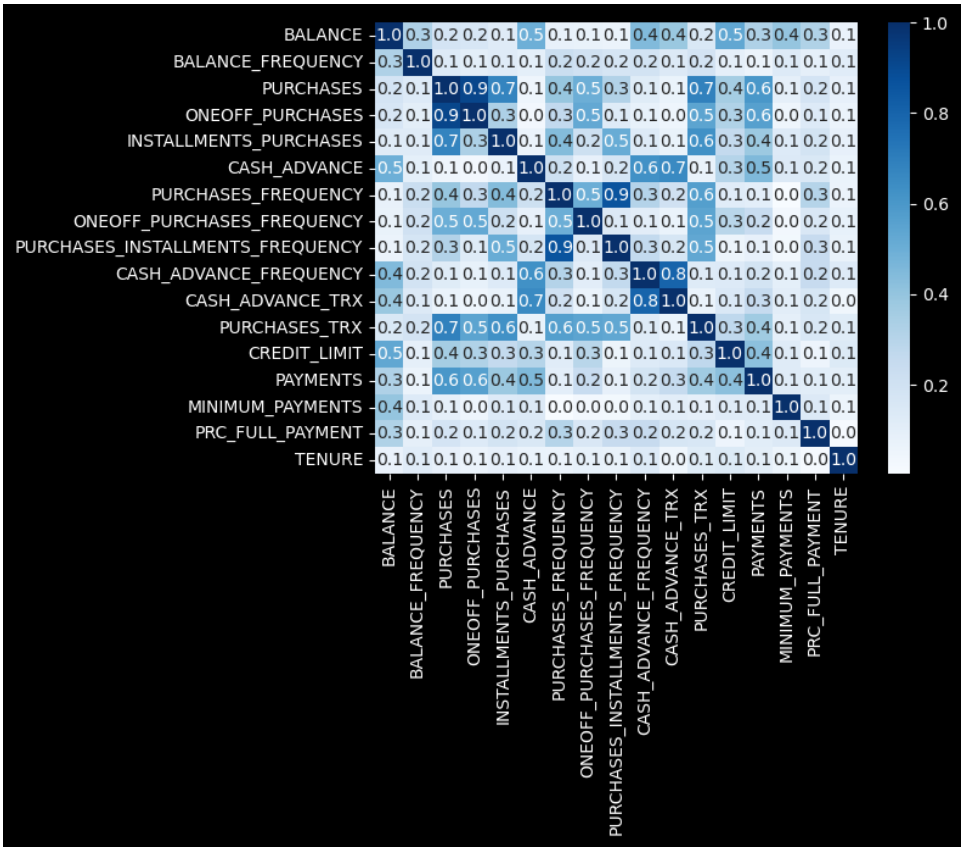
Adicionalmente la función distplot combina la función matplotlib.hist con la de seaborn kdeplot(), nos muestra la densidad de una probabilidad el cual será muy útil para seguir obteniendo mas información de nuestros datos por ejemplo:

- Para el campo 'PURCHASES_FREQUENCY', se puede apreciar dos grupos diferentes de clientes
- Para los campos 'ONEOFF_PURCHASES_FREQUENCY' y 'PURCHASES_INSTALLMENT_FREQUENCY' la gran mayoría de usuarios no pagan todo de golpe ni a plazos
- Muy pocos clientes pagan su deuda por completo 'PRC_FULL_PAYMENT'



Análisis multivariable: herramientas como la matriz de correlación son importantes para ver el comportamiento entre variables. del cual podemos deducir lo siguiente:

- Todas las correlaciones son positivas
- Existe una alta correlación entre PURCHASES, ONEOFF_PURCHASES, INSTALLMENTS_PURCHASES, PURCHASES_TRX, es común que a más compras se haga estas pueden ser a contado o a crédito y a su vez implica mayor número de transacciones.

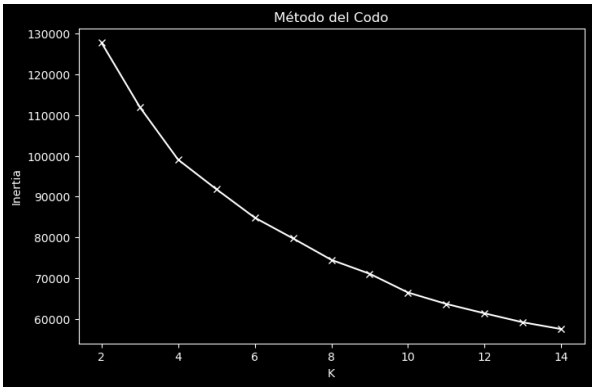


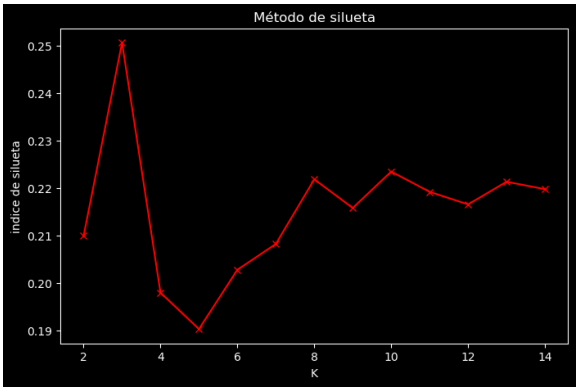
Ingeniería de funciones

Antes de aplicar el método de Kmean lo primero es normalizar los datos para asegurar que todas las características tengan un rango similar.

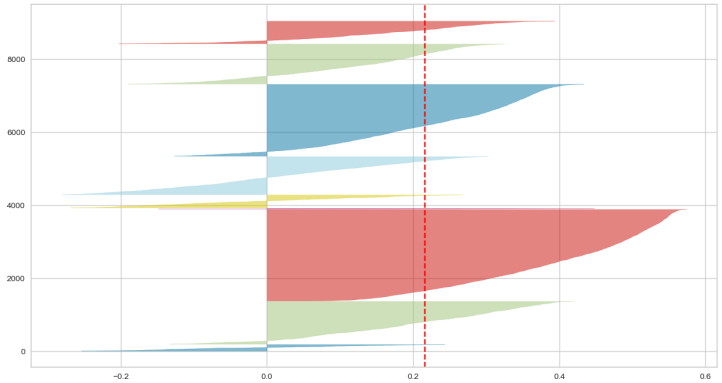
Elegimos el algoritmo k-Means, ya que se conoce entre todos los demás métodos de agrupación en clústeres por ser el más rápido y más adecuado a la dispersión de los datos, y el más utilizado en la segmentación de clientes.

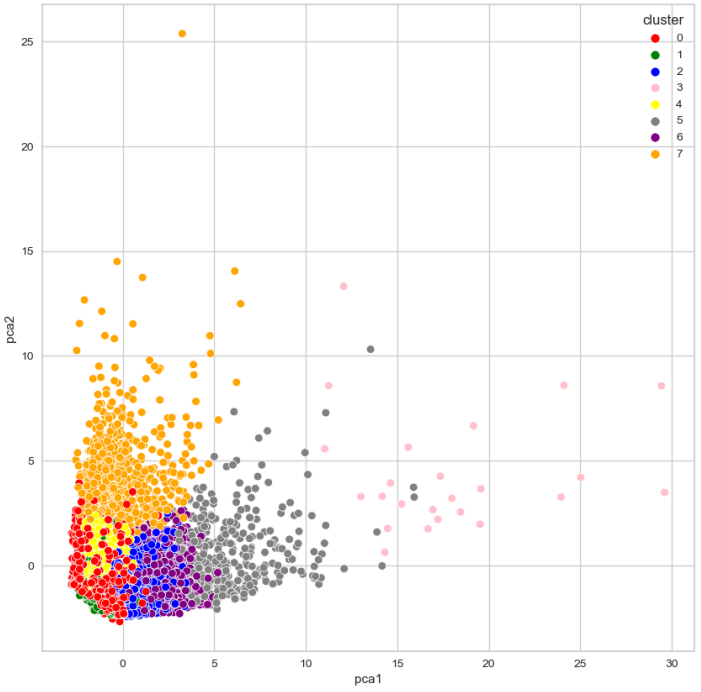
Para iniciar el método K Means es necesario asignarle de forma manual el número de clusters para ello utilizaremos el método del codo que mide la distancia promedia del centroide a todos los puntos del clúster y el coeficiente de silueta.
En el método del codo se puede visualizar el quiebre en el cluster = 9, a partir del cual la inercia baja de forma plana. pero en la gráfica del índice de silueta que mide la calidad del agrupamiento con 3 clusters llega el 25 % el cual es mayor al resto, pero mantendrá una inercia alta en el método del codo. Por esta razón se elige el segundo mejor del método de la silueta que sería 8.
Tengamos en cuenta que el número de clústeres proporcionado por estos métodos confirma la regla general que sugiere que el número de clústeres no debe ser ni pocos ni muchos para adaptar una estrategia de marketing y optimizar los recursos.



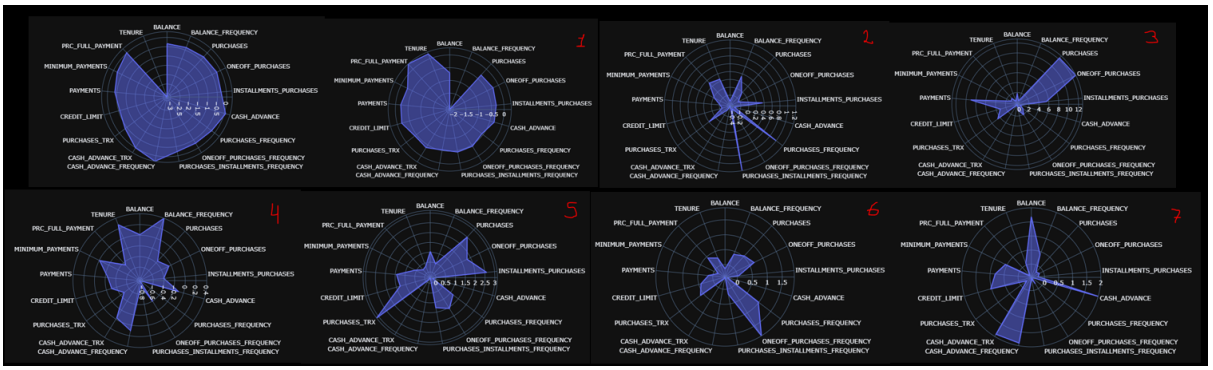


Aplicando el método K-means, a continuación podemos observar en el diagrama de siluetas, cuyo coeficiente de silueta promedio es 0.22 quedando por debajo del óptimo esperado (1), también podemos observar la presencia de coeficientes menores a cero lo cual indica que los clientes no pertenecen al cluster en el que se encuentran.
Actualmente tenemos 17 características las cuales son casi imposibles poder visualizarlas todas en un solo gráfico, por ello aplicaremos PCA para reducir a 2 componentes principales de esta manera podremos visualizar los clusters de una mejor manera.
En la gráfica podemos observar que los clusters están claramente separados, todos los punto que pertenecen al mismo color son un cluster, en el podemos observar que el cluster nro 3 tiene los puntos más dispersos, asimismo que se puede observar que algunos puntos se superponen.





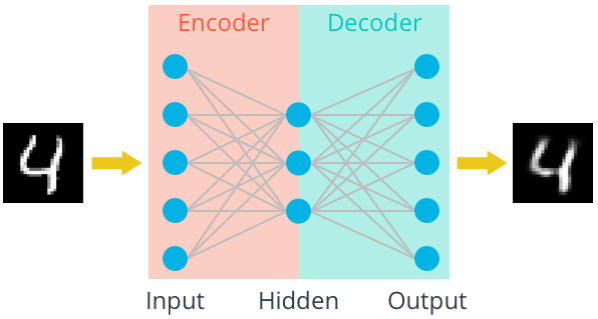
Encontrar las características de estos 8 clusters/segmentos es complicado



Autoencoder

Un autoencoder es una arquitectura de red neuronal que se utiliza para el aprendizaje no supervisado y la compresión de datos. Su objetivo principal es aprender una representación de baja dimensión de los datos de entrada y luego reconstruir los datos de manera precisa. Esto se logra mediante la codificación (toma los datos de entrada y los transforma en una

representación de baja dimensión llamada "espacio latente") y decodificación (Recibe el código latente y lo reconstruye en una representación similar a los datos de entrada originales) de los datos en una estructura en forma de "codificador-decodificador".

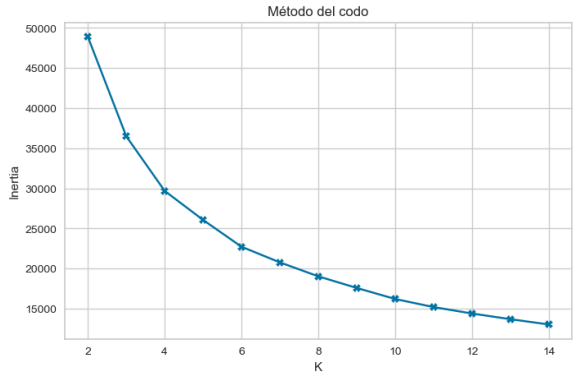


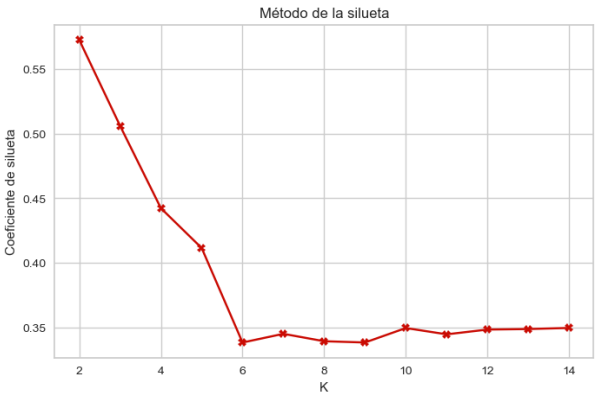
Tenemos una entrada de 17 características las cuales se reducirá a 10 características de la capa central los cuales serán suministrados al K-means.

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 17)]	0
dense_8 (Dense)	(None, 7)	126
dense_9 (Dense)	(None, 500)	4000
dense_10 (Dense)	(None, 500)	250500
dense_11 (Dense)	(None, 2000)	1002000
dense_12 (Dense)	(None, 10)	20010
dense_13 (Dense)	(None, 2000)	22000
dense_14 (Dense)	(None, 500)	1000500
dense_15 (Dense)	(None, 17)	8517

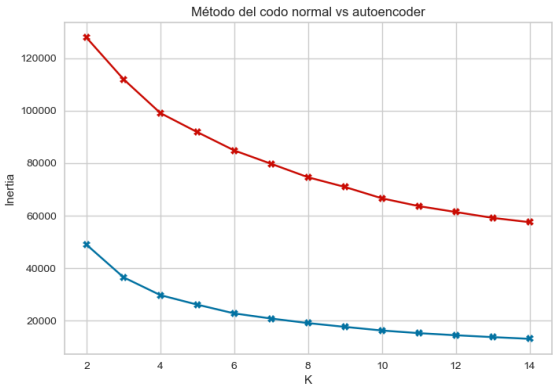
Total params: 2,307,653
Trainable params: 2,307,653
Non-trainable params: 0

A continuación procedemos a encontrar el número de clusters. En el método del codo se puede visualizar el quiebre en el cluster = 4, pero en la gráfica del índice de silueta que mide la calidad del agrupamiento con 2 clusters llega el 57 % el cual es mayor al resto, pero mantendrá una inercia alta en el método del codo. Por ello seleccionaremos 4 clusters.

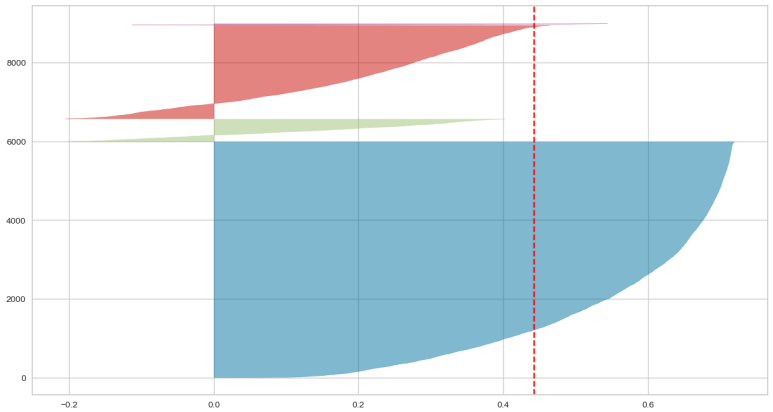


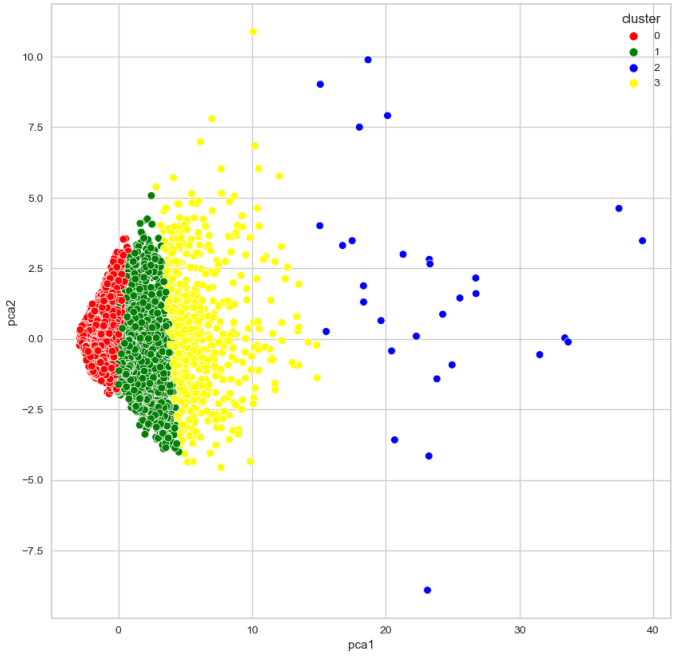


Adicionalmente podemos observar que con el autoencoder (línea azul) obtuvimos menor inercia.

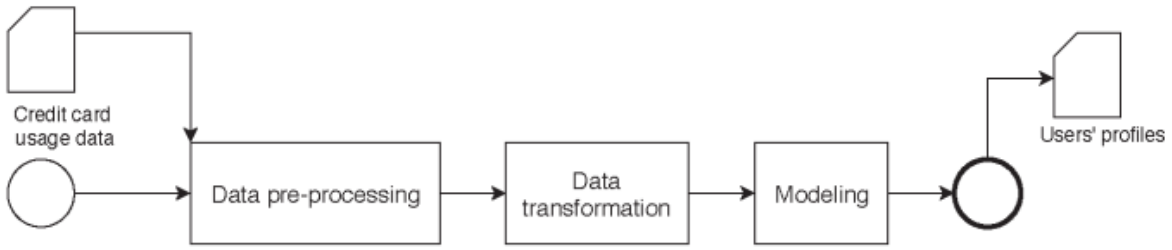


Nuevamente aplicando el método K-means ahora nuestro set de datos reducida a 10 características, en el diagrama de siluetas su coeficiente de silueta promedio es 0.44 mayor al obtenido anteriormente, también podemos observar la presencia de coeficientes menores a cero pero en menor cantidad incluso se ve bien defino el cluster 0. Aplicando el PCA para reducir a 2 componentes principales en dicha gráfica podemos observar que los clusters están claramente separados y que es el pca1 es el que define el cluster al cual pertenecen los datos, siendo el cluster 2 el más disperso y en poca cantidad, a simple vista ya no vemos que los puntos se superponen como sucedía en el en diagrama anterior.





Metodología (Propuesta del modelo de análisis visual - Pipeline)



Primeramente se ha entendido el problema del negocio que estaba dirigido claramente al departamento de marketing de este banco.

A partir de entender cuáles eran los puntos clave, hemos visualizado el dataset para hacernos una idea de cómo estaba formado, cuáles eran las variables, su rango de valores e incluso los valores faltantes que hemos rellenado con la técnica de reemplazo por la mediana después de alguna visualización, hemos hecho una mezcla de los histogramas con la Kde (estimación de densidad), también hemos realizado una matriz de correlaciones para entender las variables y cómo se relacionan las unas con las otras, claramente se vió que había mucha correlación entre las variables,
Con el algoritmo de clustering de K-means utilizamos el método del codo y el coeficiente de la silueta que nos ayuda a decidir el número óptimo de cluster a la hora de agrupar la información.
A partir de aquí hemos empezado a hacer un primer clustering con ocho grupos que ya nos ha dado una pequeña idea de estos ocho segmentos a los que podríamos lanzar una campaña de marketing.
Hicimos un primer PCA para poder visualizar todas las características proyectadas en las dos componentes más importantes, las dos que se llevan la mayor parte de la información. Si bien se puede distinguir los 8 clusters se puede observar puntos superpuestos con diferentes clusters esto se confirma con el diagrama de silueta que muestra clientes en un cluster al cual no corresponden (coeficiente menor a 0).

Sin embargo, es un poco complicado darles un nombre a cada grupo. lo que nos ha llevado a la segunda parte, intentar eliminar esas correlaciones entre las variables antes de hacer la segmentación.

Primero hacemos una reducción de la dimensión, pasar de 17 a solamente 10 y volver a aplicar primero el método del codo y el coeficiente de silueta para detectar el óptimo K, que en este caso ha sido 4 volver a aplicar el método de K-means, hacer un nuevo clustering y finalmente volver a visualizarlo en 2D a través del nuevo PCA que nos muestra una segmentación más ordenada.

A partir de ahora cualquier nuevo cliente se le aplicará primero la reducción a diez variables a través del encoder, luego se le aplica el clustering para ver dónde caen y finalmente se le aplica

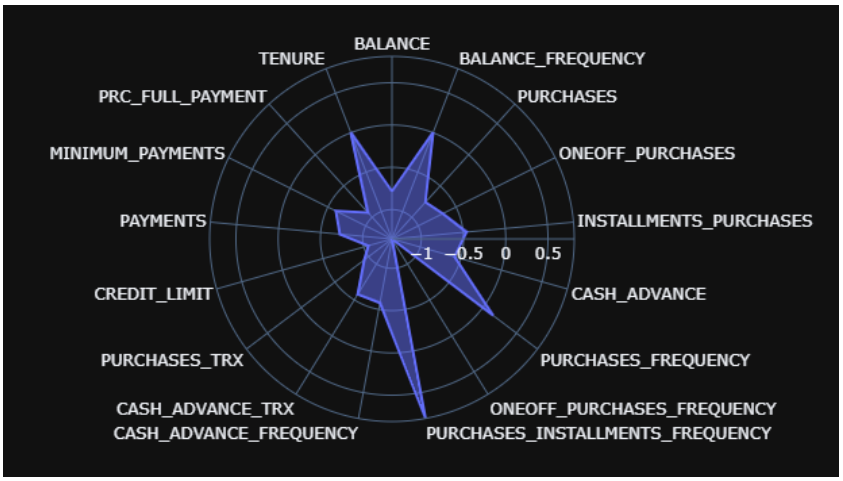
el PCA para poder pintarlo en esta región. Y podemos decidir claramente si está cerca de los rojos, de los azules, los verdes o los amarillos.
Por tanto, nuevo cliente que nos ingrese al banco, nuevo cliente que sabremos exactamente qué ofrecerle, cuáles son sus necesidades y también ayudará a que nuestro departamento de marketing pueda ofrecer productos, servicios y promociones personalizadas.

Discusión

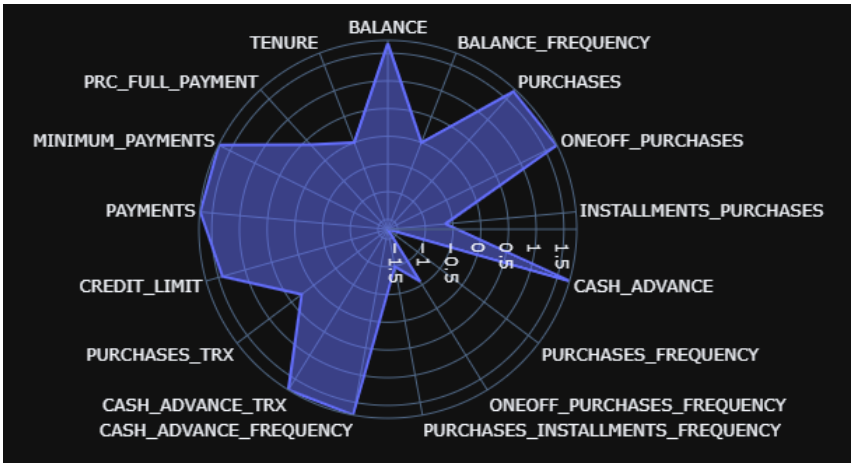
En este estudio, utilizamos el algoritmo de clustering K-means para segmentar a los clientes de nuestro banco con el objetivo de mejorar nuestras estrategias de marketing y personalización. A partir de los resultados obtenidos, se pueden extraer varias conclusiones importantes.

En primer lugar, los resultados de la segmentación revelaron la existencia de cuatro segmentos distintos de clientes. Estos segmentos se diferencian en función a la conducta en el uso de la tarjeta del cliente. Cada segmento representa una oportunidad única para adaptar nuestras estrategias de marketing y abordar las necesidades específicas de cada grupo.

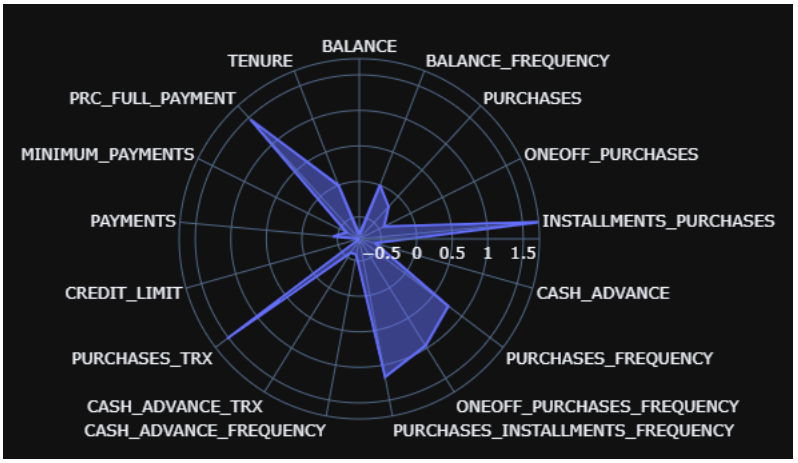
Grupo 1:
realiza compras con mucha frecuencia, frecuencia nula en las compras de golpe o directas, límite de crédito bajo, el monto por compra realizada rodea los 210 $, la cuenta del cliente tiene el saldo más bajo ronda los 300$
“intenta pagar en menor cantidad con intereses y cuenta con un saldo muy bajo”



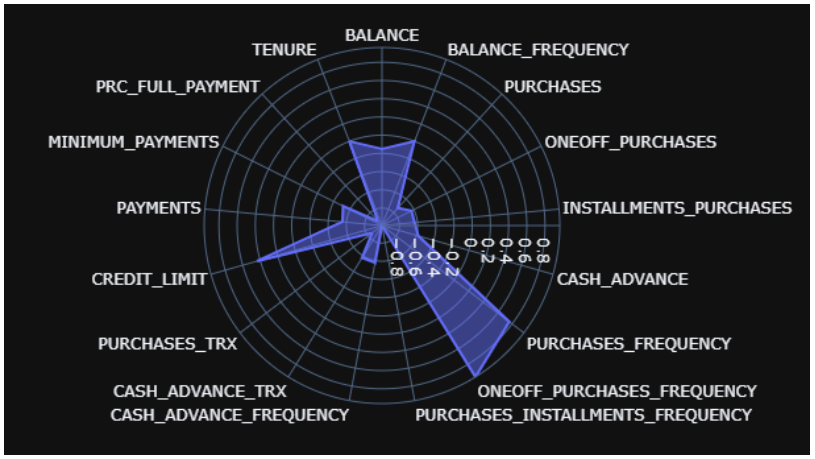
Grupo 2:
tiene un saldo elevado entorno a 6000$, piden anticipo en efectivo bien elevado con alta frecuencia, baja frecuencia de compra, el número de compras rodea los 22000$, limite de tarjeta alrededor de 7500 $
“gente que pide anticipos constantemente, usan tarjeta de crédito como préstamo”



grupo 3:
Límite de crédito bajo, porcentaje de pago completo alto 44%
“personas capaz de devolver completamente su préstamo candidato a incremento del límite de crédito”



grupo 4:
El número de compras rodea los 422$, pero tiene alta frecuencia de compras y alta frecuencia de compras de golpe pero también de montos pequeños, no pide anticipos, límite de crédito alto
“realiza muchas compras pero en montos pequeños”



Es importante destacar que la segmentación de clientes utilizando K-means nos permite optimizar nuestras estrategias de marketing al dirigir nuestros recursos de manera más eficiente y personalizada. Además, el uso de este enfoque de clustering nos brinda una base sólida para futuras investigaciones y mejoras en nuestras estrategias de segmentación.

Sin embargo, es importante mencionar algunas limitaciones de este estudio. Aunque el algoritmo K-means es ampliamente utilizado y eficaz, la elección del número óptimo de clusters puede ser subjetiva y requerir validación adicional. Además, la segmentación se basa en los datos disponibles y puede no capturar todas las dimensiones relevantes para la personalización efectiva.

Conclusiones.

En este estudio, utilizamos el algoritmo de clustering K-means para segmentar a los clientes del banco con el objetivo de comprender mejor sus características y necesidades. A partir de los resultados obtenidos, se ha identificado diferentes segmentos de clientes que nos proporcionan información valiosa para la toma de decisiones estratégicas en marketing, los cuatro grupos obtenidos tienen las siguientes características: “intenta pagar en menor cantidad con intereses y cuenta con un saldo muy bajo”, “gente que pide anticipos constantemente, usan tarjeta de crédito como préstamo”, “personas capaz de devolver completamente su préstamo candidato a incremento del límite de crédito”, “Realiza muchas compras pero en montos pequeños”

Antes de realizar el método de clusterización con K-means es recomendable hacer un correcto preprocesamiento de datos, posteriormente elegir adecuadamente el número de clusters y también es recomendable reducir el número de dimensiones (características) sin perder mucha información en la reducción, para ello es recomendable el uso de autoencoders ya que están entrenadas para una reconstrucción precisa.

Repositorio.

El código desarrollado en el presente estudio se encuentra en el siguiente link: https://github.com/weicap/MCC_Information_Recovery/tree/main/Final_Project