**Nicolas Segura**

**Jaime Torres**

**David Ruiz**

**Recomendaciones BancAlpes**

**Link del repositorio:** <https://github.com/drvillota/Laboratorio-2>

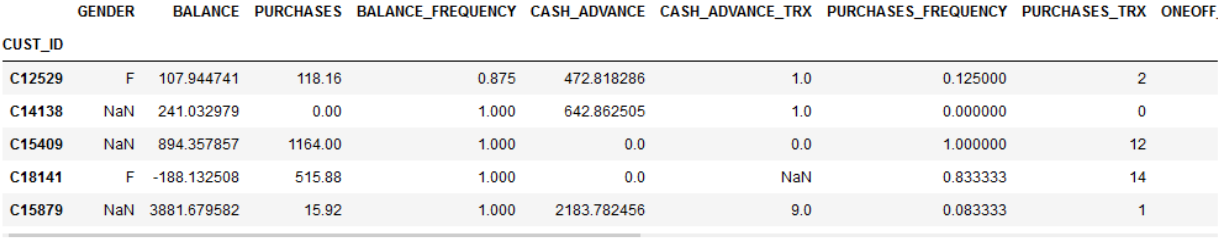
**Link presentación:** [https://www.canva.com/design/](https://www.canva.com/design/DAE6QYrdEwg/0QJInjCSbeaZfFu01aXTtg/view?utm_content=DAE6QYrdEwg&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton)

**El dashboard se encuentra disponible en el repositorio.**

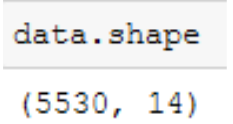
1. **Tratamiento de los datos:**

* **Perfilamiento de datos y Preparación de datos**

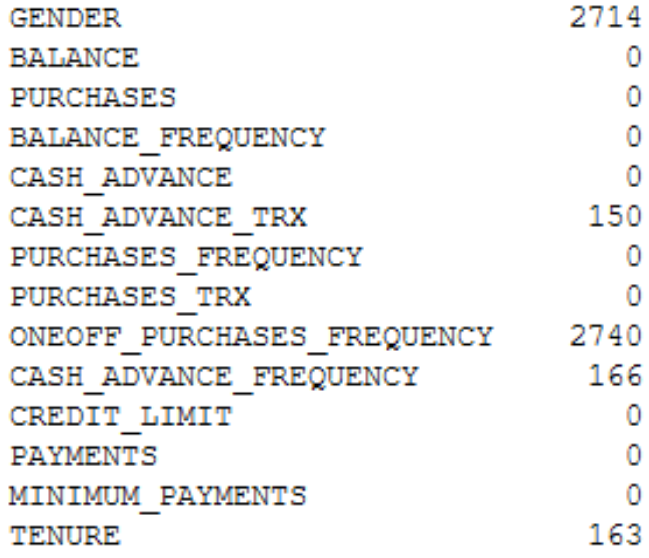
Cargamos y visualizamos los datos, estos tienen la siguiente estructura



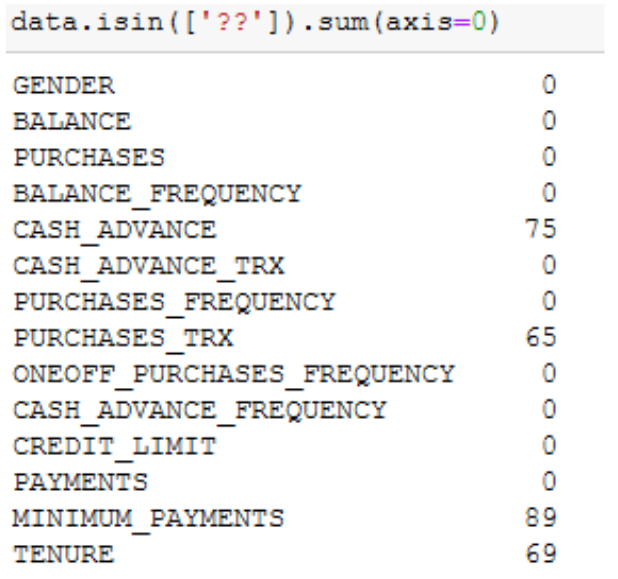
La base de datos cuenta con un total de 5530 filas, cada una con 14 atributos diferentes.



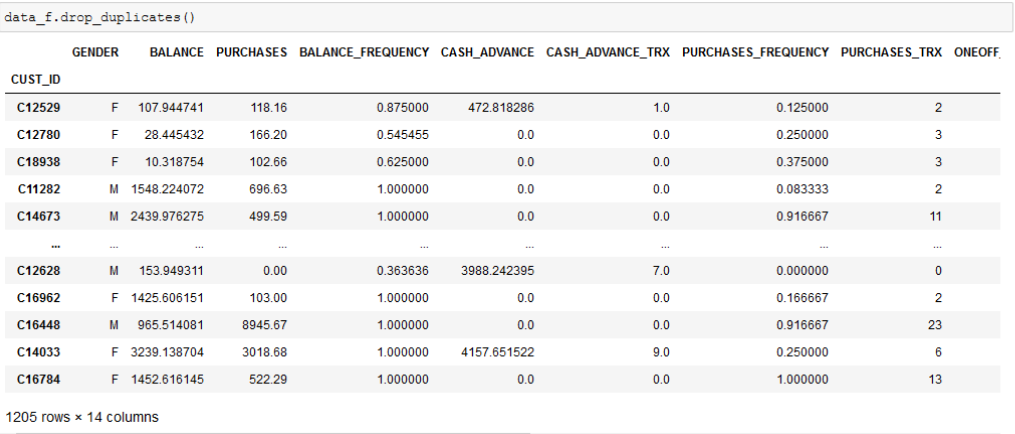
Ahora vamos a visualizar por cada atributo cuantos datos son nulos, aquí podemos ver que la variable “ONEOFF\_PURCHASES\_FRECUENCY” tiene una gran cantidad de datos nulos con 2740, seguido por "GENDER” que tiene 2714 datos nulos.



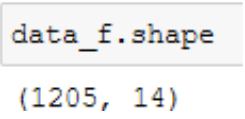
Se revisa cuantos datos tienen caracteres especiales y se borran los que sean registros inválidos.



Se borran los duplicados



Después de realizar el perfilamiento quedamos con 1205 datos.



1. **Modelamiento y Validación:**

**Kmeans (Nicolás Segura Castro):** Antes de desarrollar el modelo nos pudimos percatar de que había una gran correlación entre las columnas: Cash\_Advance, Payments, Balance, Purchases y Credit\_limit. Lo anterior nos permitió entender que estas podían ser los criterios más importantes para el clustering, por lo que los incluimos como variables categóricas y ejecutamos las funciones de codo y silueta obteniendo los siguientes resultados:

Imagen que contiene Código QR

Descripción generada automáticamente

Encontramos que para ambos métodos el número de centroides era 3 y analizando los resultados del cluster encontramos los siguientes tres grupos:

Definimos tres grupos a través del clustering:

* Grupo 1 - 497: Son el grupo más numeroso y son aquellos que tienen balances más bajos, límites menores de crédito y menos pagos.
* Grupo 2 -331: Son un el grupo más equilibrado en cuanto a balance, sin embargo, son los que realizan el mayor número de compras.
* Grupo 3 -377: Son el grupo más pequeño y destaca en general por ser el equilibrio entre los grupos 1 y 2.

Una vez clasificados los grupos, decidimos incluir más variables en el análisis del modelo como Gender y Tenure sin embargo, luego de ejecutar varias veces el modelo no encontramos una correlación clara entre aquellas columnas y las previamente seleccionadas. Fue importante para nosotros el tener en cuenta que para Kmeans se agrupan ciertos grupos de acuerdo con unas características específicas minimizando la suma de distancias entre cada objeto y su cluster. Para este caso utilizamos la distancia cuadrática luego de realizar la normalización y la estandarización de los datos, la cual fue aplicada para el resto de los métodos de agrupamiento.

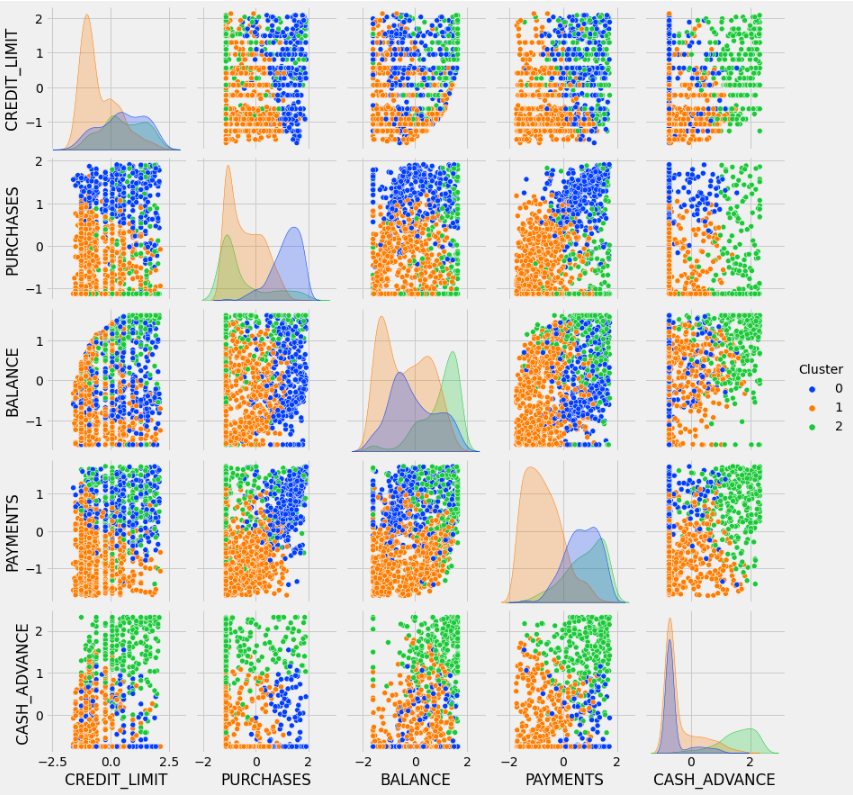
De esta manera, decidimos continuar con los otros métodos de clustering para los resultados obtenidos a través de Kmeans.

Una captura de pantalla de un celular con texto e imágenes

Descripción generada automáticamente con confianza baja

**Spectral Clustering (Jaime Torres):** En Spectral Clustering, los datos son tratados como nodos de un grafo. A partir de aquellos nodos, se trata de desarrollar un problema de partición. Dichos nodos son mapeados en lo que se conoce como un “low-dimensional space” en el que pueden ser segregados por cada uno de los clusters. Es importante recalcar que no se asume de ninguna manera la forma de los clusters, es por eso que Spectral Clustering utiliza un grafo KNN para determinar las distancias que habrá entre cada uno de los datos y sus clusters. Decidimos utilizar este método por su fácil implementación y sus buenos resultados, sin embargo, es importante tener en cuenta que su costo computacional puede ser bastante alto en relación con otros métodos.

A partir de dicho método obtuvimos los siguientes resultados:



Definimos los mismos tres grupos a través del clustering que con Kmeans:

* Grupo 1 - 600: Son el grupo más numeroso y son aquellos que tienen balances más bajos, límites menores de crédito y menos pagos.
* Grupo 2 -270: Son un el grupo más equilibrado en cuanto a balance, sin embargo, son los que realizan el mayor número de compras.
* Grupo 3 -335: Son el grupo más pequeño y destaca en general por ser el equilibrio entre los grupos 1 y 2.

**Agglomerative Clustering (David Ruiz):** Al analizar los tipos de agrupamiento que podíamos utilizar identificamos que agglomerative clustering es implementado normalmente para agrupar objetos en grupos en función de su similitud. En este caso, el algoritmo empieza tratando a cada objeto como un cluster único. Los pares de grupos se fusionan sucesivamente hasta que todos les grupos se fusionan en varios grupos grandes que contengan a los objetos. El resultado es una representación basada en árboles de los objetos y cuenta con una aproximación en donde se considera inicialmente a cada grupo como un solo elemento (hoja) y en cada paso del algoritmo los grupos que son más similares se van combinando en un grupo más grande de nodos. Dicho procedimiento se repite hasta que todos los puntos son miembros de un solo gran clúster. A partir del método obtuvimos los siguientes resultados:

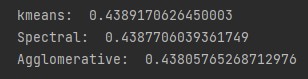
Una captura de pantalla de un celular con texto e imágenes

Descripción generada automáticamente con confianza media

Definimos los mismos tres grupos a través del clustering:

* Grupo 1 - 600: Son el grupo más numeroso y son aquellos que tienen balances más bajos, límites menores de crédito y menos pagos.
* Grupo 2 -270: Son un el grupo más equilibrado en cuanto a balance, sin embargo, son los que realizan el mayor número de compras.
* Grupo 3 -335: Son el grupo más pequeño y destaca en general por ser el equilibrio entre los grupos 1 y 2.

Después de analizar los resultados arrojados por cada uno de los métodos y teniendo en cuenta que se utilizó la función de silueta en todos para escoger el número de clusters (que siempre fue 3), decidimos utilizar el **puntaje de silueta** para escoger el mejor modelo. De esta manera pudimos evidenciar que Kmeans era el que contaba con un puntaje más adecuado, así como se evidencia en la siguiente imagen:



**Conclusiones:**

Teniendo en cuenta los resultados obtenidos a través de los puntajes de silueta concluimos que el modelo más adecuado para este caso es el que fue desarrollado a partir del algoritmo de Kmeans, ya que los datos categóricos pudieron ser fácilmente transformados a numéricos y, además cuenta con una menor complejidad computacional respecto al Spectral Clustering. No obstante, cabe aclarar que en cuanto a complejidad la más baja el algoritmo ganador es Agglomerative Clustering, así que es posible afirmar que para conjuntos de datos extremadamente grandes resultaría más conveniente el uso de este algoritmo en la implementación del modelo.

Basados en lo anterior, le recomendamos entonces a BancAlpes el diseñar y orientar sus campañas de marketing de forma específica a los grupos de *Grupo 1*, *Grupo 2* y *Grupo3* para así mejorar sus ventas, la retención de sus clientes y la calidad de atención. Para finalizar, debe de tenerse en cuenta que, según lo descubierto en este ejercicio, no se deben sesgar ese tipo de campañas frente a atributos como el estado civil, la edad o la educación del cliente.