María Camila Terán — 201822000
Juan Diego Cardona — 201819447
Nicolás Ortega — 201814515

Inteligencia de Negocios
Laboratorio 2

Análisis y perfilamiento de los datos

Para el perfilamiento de datos, el primer paso fue estudiarlos desde el archivo csv. Desde este se pudo evidenciar que hay dos variables categóricas que es necesario modificar a futuro. De igual manera, se observan algunos datos atípicos que hay que modificar como signos de interrogación o que "Male" está representado con una M. Por otro lado, ya en aspectos más técnicos hay 660 datos, pero se evidencian vacíos. De igual manera, para la variable "Education" tiene dos números que representan "desconocido" por lo que es necesario eliminar una en el pre-procesamiento de datos. Además, se presentan datos atípicos en la columna de edad por lo que es necesario revisar y establecer un rango. Para añadir, hay columnas que parecen ser numéricas, pero en realidad no lo son.

Pre-procesamiento y limpieza de los datos

Teniendo en cuenta lo mencionado anteriormente, se inicia la limpieza eliminando los signos de interrogación. Posteriormente, se cambian los datos atípicos de la columna de "Sex" para que valores como "Fmale o Femael" queden como "Female" y valores como "M o Mael" como "Male". Luego, se dejo solo 5 para representar la educación desconocida. Además, en esta columna se eliminó el término "ABC" pues no hacía parte del diccionario de datos. El paso para seguir fue eliminar los valores numéricos de la columna "Marriage" para luego eliminar los nulos. Ahora bien, se verificaron las restricciones y se eliminaron las columnas "Id" y "Costumer" que no son de interés para el análisis a lo que pide BancAlpes. Se estableció el rango de edad menor a 100 y mayor a 21 para finalmente establecer el número de tarjetas de crédito menor o igual a 10. Para el procesamiento de las variables categóricas, se utilizó "One Hot Encoding". Cabe resaltar que con la limpieza de datos, quedaron 639 datos.

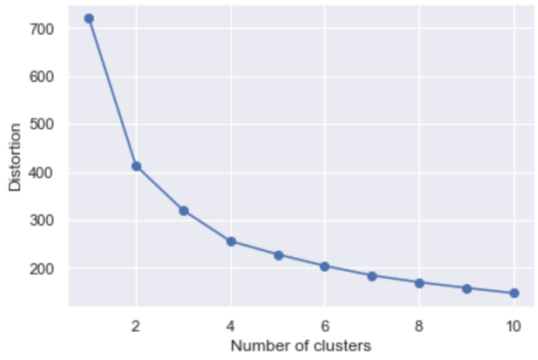
Modelos

K-Means (Juan Diego Cardona)

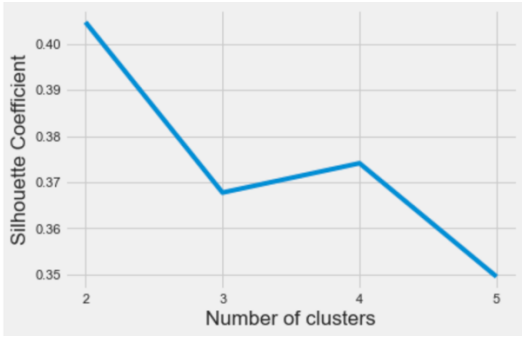
La implementación del algoritmo de clustering K-Means inició con una etapa de procesamiento de datos. Se realizó la limpieza de los datos a nivel de calidad de registros, valores nulos o valores inconsistentes. Posteriormente se realizó una codificación de las variables "*Sex"* y "*Marriage*" utilizando la técnica de *One Hot Encoding* debido a que el algoritmo no maneja variables categóricas y deben ser pasadas a numéricas. Estas nuevas columnas con variables "*dummies"* fueron agregados al conjunto de datos del algoritmo.

También se realizó un proceso de normalización de las variables numéricas con el objetivo de que variables como "*Limit_bal"* (que manejaba valores muy altos) no distorsionara o tuviera más peso que las demás al momento de calcular las distancias.

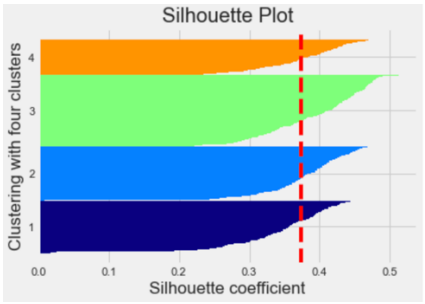
Luego, se realiza un algoritmo de *Plot distortions* con el objetivo de encontrar el número óptimo de clústeres utilizando el método del codo. De este, se obtuvo la siguiente gráfica la cual nos indica que el número óptimo, analizando el nivel de distorsión de los datos a partir de el número de clústeres utilizado es igual a 4.



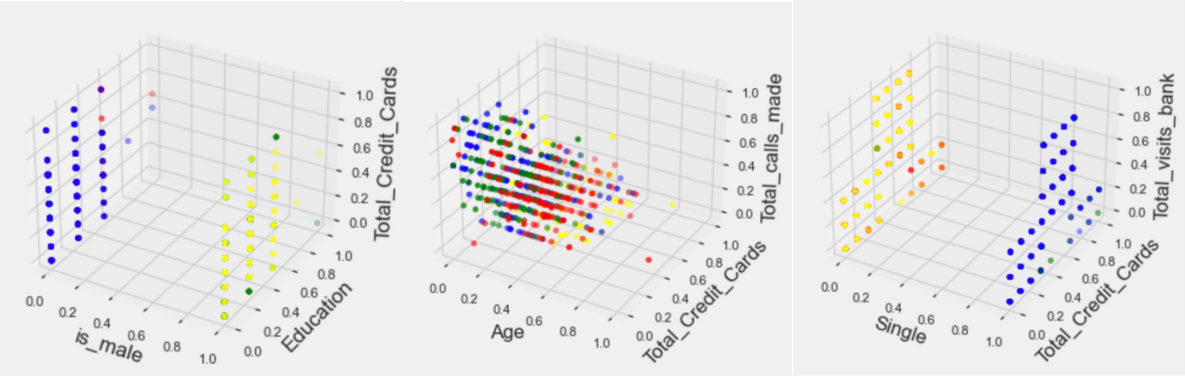
Para corroborar esta información de acuerdo con los valores de la silueta se prosigue a determinar cual es el mejor número de clústeres del cual también se obtiene como resultado el mismo valor.



Ya con el mejor número de clústeres seleccionado, se construyó el modelo de K means y se evaluó su calidad. Dadas las pruebas anteriores, el hiperparámetro seleccionado fue 4.



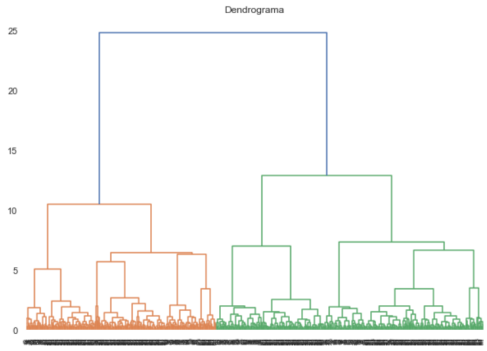
Los resultados finales se despliegan en diferentes vistas las cuales nos permiten deducir ciertas conclusiones. Se puede observar un modelo con 4 clústeres. El modelo posee pocos valores atípicos, no hay valores negativos en los coeficientes de la silueta y hay buena cohesión entre los clústeres. Respecto a las interacciones entre las variables y el agrupamiento entre las columnas, se puede observar que las variables más correlacionadas son: el número de tarjetas de crédito, el número de visitas al banco y el número de llamadas. No existe relación con el género ni el estado civil. Los clientes con más tarjetas realizan más visitas al banco y más llamadas también. Los clientes con menos tarjetas realizan más visitas en línea. Hay una gran relación entre el número de tarjetas de crédito y la edad.



Clustering jerárquico (Maria Camila Terán)

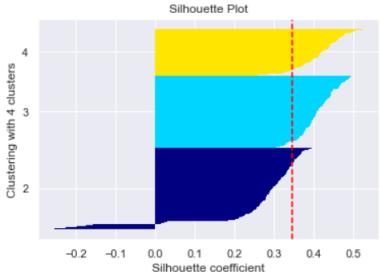
Luego del procesamiento de datos realizado anteriormente, incluyendo la modificación de las variables categóricas, el primer paso fue decidir cuántos clusters realizar. Hay que resaltar que el cluster jerárquico consiste en combinar los datos similares para asignarlos a un mismo cluster. De esta forma, los datos diferentes no quedan cerca. En esta ocasión, se utilizó el método aglomerativo el cual utiliza la distancia euclidiana.

Ahora bien, para seleccionar el número óptimo de clusters se utilizó la herramienta del dendrograma. El resultado fue el siguiente:



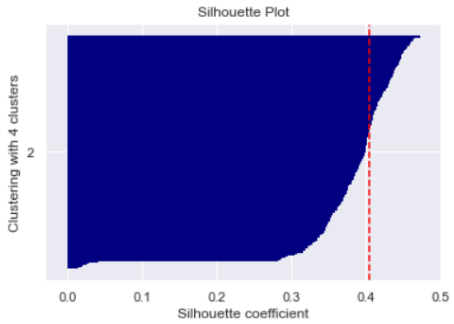
Allí se observan las jerarquías formadas. De esta forma se pueden visualizar las distancias calculadas. Entre más es la distancia de las líneas verticales del dendrograma, mayor será la distancia entre los clusters. En ese sentido, se observan dos opciones factibles: que el número de clusters sea 2 o que el número de clusters sea 4. Esto teniendo en cuenta que se busca trazar una línea en el punto más alto.

Trazando la línea en 8 (4 clusters): se crea el diagrama de siluetas para corroborar si la selección de este número de clusters es adecuada:

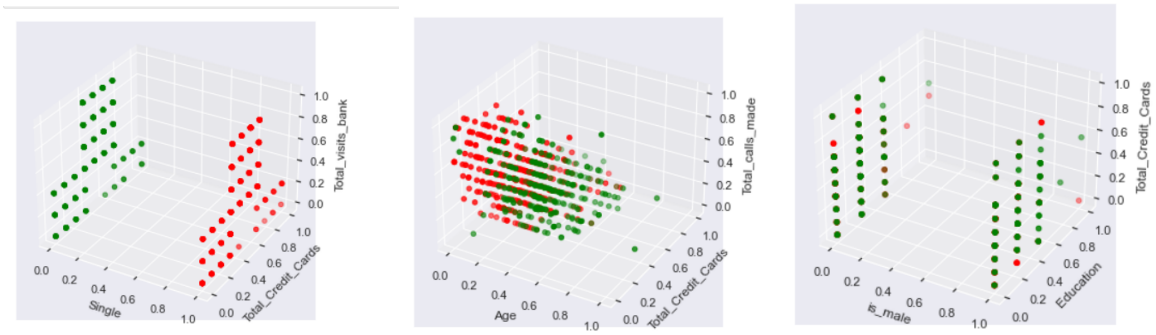


Como se puede evidenciar, los grupos no tienen mucho sentido dado que hay coeficientes negativos. Por ende se traza la línea en 15.

De nuevo se hace el diagrama de la silueta:



Efectivamente ya no hay coeficientes negativos, pero no es muy cercano a uno. Sin embargo, hubo mejorar respecto a si se traza la línea en 8 por lo que se elige este como mejor modelo. Finalmente, se hicieron distintas representaciones visuales con el fin de mirar la relación con las variables. Estas se observan a continuación:

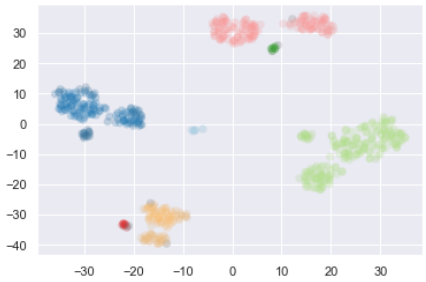


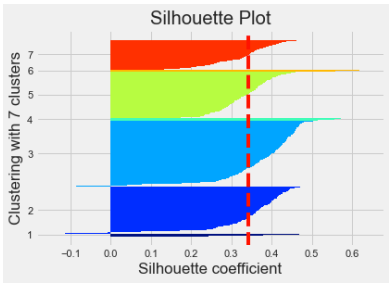
Lo primero que se puede observar es que hay una gran correlación entre la edad, el número de tarjetas de crédito y el número de llamadas. Esto indicia que, entre mayor número de tarjetas de crédito, más llamadas realizará el banco. Por otro lado, se observa que el género influye poco en la cantidad de tarjetas de crédito. Asimismo, se tiene una alta relación entre el número de tarjetas y las visitas al banco. No obstante, si una persona hace más llamadas al banco, menos visitas tendrá. No se ve una gran influencia del estado civil en el número de tarjetas de crédito.

HDBScan (Nicolás Ortega)

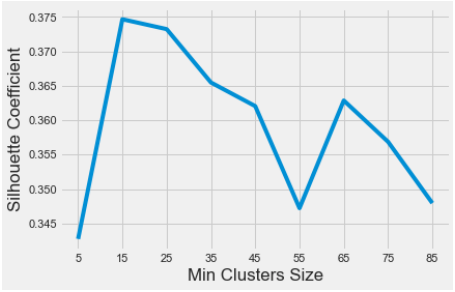
Para la implementación de este algoritmo se tuvo que realizar el procesamiento de los datos mencionado anteriormente, y adicionalmente, se realizaron tareas adicionales para el tratamiento de las variables categóricas. Como el algoritmo no maneja variables categóricas, luego de realizar la limpieza de los datos a nivel de calidad de registros, valores nulos o valores inconsistentes, se realizó una codificación de las variables “Sex” y “Marriage” utilizando la técnica de One Hot Encoding. Esta técnica permitió crear las columnas con las variables “dummies” que se agregaron al conjunto de datos para ser ingresadas al algoritmo. Por otro lado, dado que se está realizando una tarea de clustering y en el proceso del algoritmo se tiene en cuenta la distancia que hay entre los puntos, se realizó un proceso de normalización de las variables numéricas. Esto para que variables como “Limit_bal” (que manejaba valores muy altos) no distorsionara o tuviera más peso que las demás al momento de calcular las distancias. Una vez terminada esta preparación de los datos, se implementó el algoritmo.

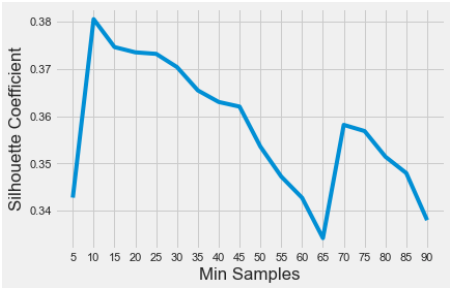
En primera instancia se realizó una ejecución del algoritmo sin fijar los hiperparámetros, y se obtuvieron los siguientes resultados:



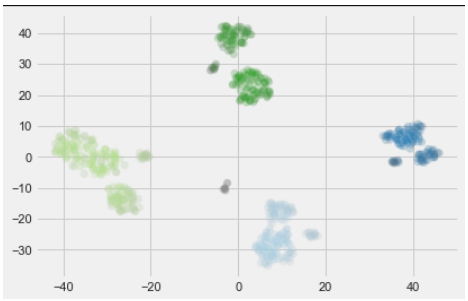


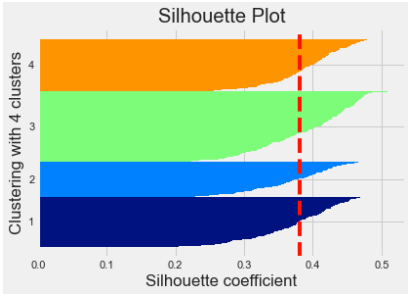
Se puede ver que se forman 7 grupos, pero algunos de estos son muy pequeños e incluso no tienen buena cohesión, puesto que reportan un coeficiente de silueta negativo. Con base en esto, se decidió realizar una búsqueda de hiperparámetros que pudieran mejorar el modelo. Este proceso se realizó de forma iterativa, es decir probando con una serie de valores y evaluando modelos resultantes para observar el que diera un mejor resultado. Los hiperparámetros a buscar fueron "min_clusters_size" y "min_samples", a continuación se muestran los resultados de la búsqueda:





Los valores dieron mejores resultados para los hiperparámteros fueron 15 y 10, respectivamente. Con estos valores se estimó un nuevo modelo:

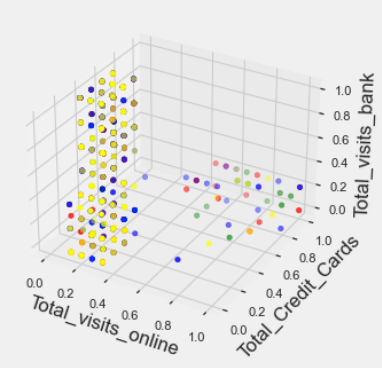


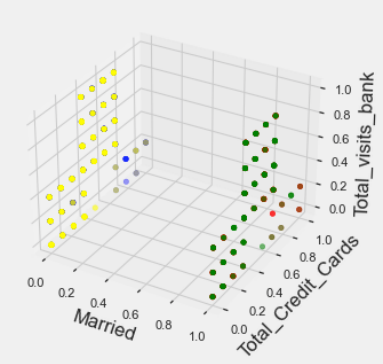


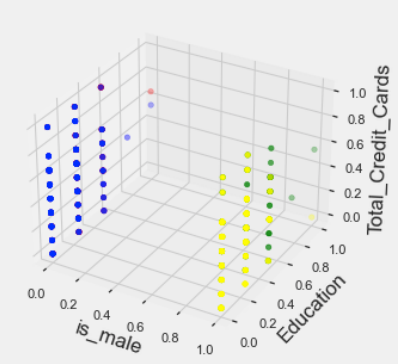
En este nuevo modelo hay menos clusters y menos valores atípicos. Los clusters tienen una mejor cohesión entre sí, no hay valores negativos para el coeficiente de la silueta, que si bien no es demasiado cercano a 1, logró mejorar con respecto a la primera ejecución del algoritmo.

Finalmente, se realizaron algunas observaciones de la interacción y agrupamiento entre las columnas y se obtuvieron algunas conclusiones. En general, el comportamiento de los clientes no variaba con el género, ni con estado de casado, soltero u otro. De forma similar, existía un comportamiento homogéneo entre las personas de todas las edades y nivel de educación. Las variables más correlacionadas eran el número de

tarjetas de crédito, el número de visitas al banco y el número de llamadas. Los clientes con más tarjetas realizan más visitas al banco y más llamadas también. Los clientes con menos tarjetas realizan más visitas en línea.







Análisis de resultados, comparación de los modelos y recomendación

De los grupos observados en los anteriores resultados podemos ver que los que los algoritmos identifican clusters muy similares de acuerdo con los datos suministrados. Después de realizar los análisis necesarios en cada caso y obtener un respectivo “mejor modelo” por algoritmo, podemos ver que se forman 4 grupos generales, y que todos los modelos presentan valores muy parecidos de coeficiente de silueta, esto quiere decir que los clusters que presentan tienen una cohesión similar. Si bien en el caso del algorimto de clustering jerárquico solo se forman dos grupos, se puede ver que estos mantienen una cohesión aceptable en cuánto a la métrica del coeficiente de la silueta.

Por otro lado, es importante tener en cuenta que los modelos se construyeron haciendo uso de una buena cantidad de variables, lo que hace un poco más compleja la segmentación. Esto con el objetivo de capturar la mayor cantidad de información posible dada del negocio y así poder identificar las variables más relevantes para la campaña propuesta. A la luz de los resultados de los modelos pudimos observar que las variables que presentaban un comportamiento más correlacionado eran el número de tarjetas de crédito, el número de visitas al banco (tanto en línea como físicas) y el número de llamadas. Los grupos más identificables se formaban de la interacción de estas variables con la variable del número de tarjetas de crédito de los clientes. Es posible observar que cuantas más tarjetas tiene un cliente, hace más visitas físicas al banco, y de manera similar, cuantas menos tiene hace más visitas en línea. En el lado opuesto, fue posible concluir que las variables de estado civil, educación y edad de los clientes no eran muy significativas ni se relacionaban tan claramente con las otras, pues en cada caso los comportamientos de estas variables eran homogéneos y no cambiaban significativamente para cada categoría.