Laboratorio 2

Clustering

Autores

Pablo Arango	Juan Diego Trujillo	Santiago Moreno
201630495	201618006	201814353
p.arango	jd.trujillom	s.morenom

Inteligencia de Negocios

Sección 2

Departamento de Ingeniería de Sistemas y Computación

Universidad de Los Andes

Bogotá, Colombia.

1. Perfilamiento de datos

En primer, comenzamos perfilando el archivo que fue entregado para este laboratorio. En este encontramos la siguiente de cantidad de filas y columnas.

Número de filas	Número de columnas
660	11

Al hacer una exploración inicial de los datos, encontramos la siguiente distribución de las columnas:

Tipo de columna	Cantidad
Numérico	9
categóricas	2

Para cada una de las columnas, describimos a que tipo pertenecen:

Nombre	Tipo
Id	Numérica discreta
Limit_bal	Numérica discreta
Sex	Categórica
Education	Numérica discreta
Marriage	Categórica
Age	Numérica discreta
Total Credit Cards	Numérica discreta
Total visits bank	Numérica discreta
Total visits online	Numérica discreta
Total calls made	Numérica discreta

Explorando algunas columnas, nos dimos cuenta de que había una fila con todos sus valores con el carácter "-". La cambiamos y pudimos convertir los valores a numéricos dentro del notebook. A continuación, se presenta una tabla con los principales estadísticos de las columnas numéricas.

Gracias a este análisis, es posible ver que existen anomalías: una de las que se pudo observar, por ejemplo, es que hay algunas edades con valores superiores a 100.

	Limit_bal	Education	Age	Total Credit Cards	Total visits bank	Total visits online	Total calls made
count	652.000000	652.000000	652.000000	6.520000e+02	652.000000	652.000000	652.000000
mean	169907.975460	1.786810	95.831288	1.893556e+05	2.424847	2.601227	3.547546
std	129220.720439	0.779309	1440.329756	4.834941e+06	1.628904	2.949737	2.852634
min	10000.000000	1.000000	21.000000	1.000000e+00	0.000000	0.000000	0.000000
25%	60000.000000	1.000000	28.000000	3.000000e+00	1.000000	1.000000	1.000000
50%	140000.000000	2.000000	33.000000	5.000000e+00	2.000000	2.000000	3.000000
75%	240000.000000	2.000000	41.000000	6.000000e+00	4.000000	4.000000	5.000000
max	630000.000000	6.000000	36745.000000	1.234568e+08	5.000000	15.000000	10.000000

2. Preparación de datos

Antes de empezar a desarrollar la preparación particular que requería cada uno de los algoritmos, se hizo en conjunto una limpieza general de los datos como punto de partida. A continuación, se explican algunos de las cosas que se hicieron:

- Se eliminó una fila en donde todos los valores estaban con el carácter "-"
- Se eliminaron filas que tenían el carácter "?"
- Se eliminaron las filas con edades superiores a lo posible
- Se convirtieron las variables categóricas a numéricas
- Se normalizaron las variables numéricas
- Se eliminaron valores irregulares en *Educacion* y *Total de tarjetas de crédito*
- Se omitieron las columnas *Id* y *Customer* pues no aportaban información relevante

La figura 1 presenta el diagrama de caja con las variables antes de su proceso de normalización. En el diagrama se refleja la necesidad de normalizar las variables para mantener la proporcionalidad en el modelo

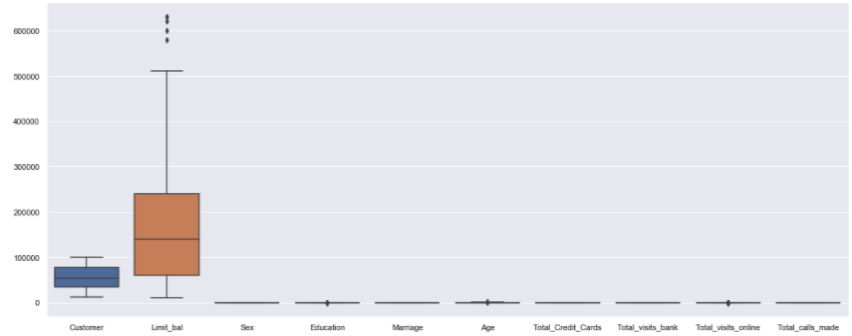


Figura 1: Diagrama de cajas antes de la normalización

Figura 2

La figura 2 presenta un diagrama de caja donde se presentan los valores luego del proceso descrito previamente.

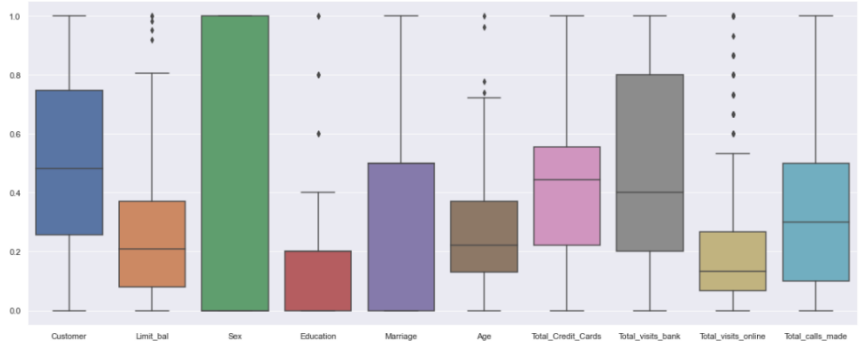


Figura 2: Grafico de cajas de los datos procesados

Dimensiones

- Las dimensiones que se utilizaron para realizar el modelo fueron
    - Total de Tarjeta de Credito
    - Visitas Totales al Banco
    - Visitas Totales Online

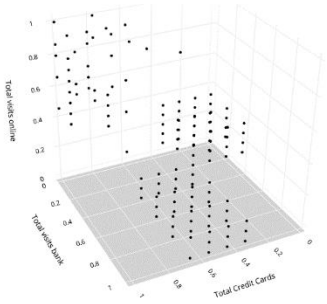


Figura 3: Grafica con los tres ejes de los atributos escogidos

a. K-Means

En el caso de K-Means no se hicieron modificaciones al *dataframe* posteriores a las mencionadas previamente.

b. DBScan

En el caso de DBScan, simplemente se revisó si era necesario hacer un escalado de los datos de tal forma que todos quedaran en una escala similar. Sin embargo, el siguiente grafico muestra que esto no es necesario:

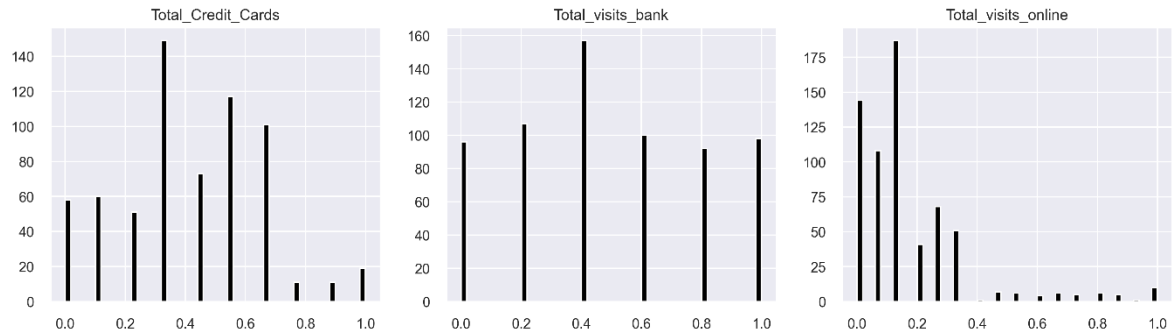


Figura 4: Escalas luego de la normalización

c. Mean shift

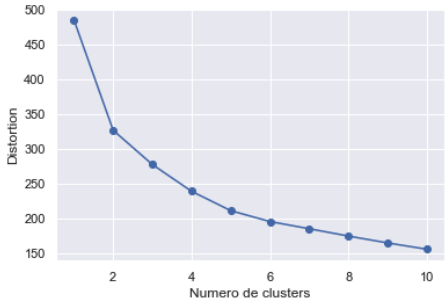
En el caso de Mean-Shift no se hicieron modificaciones al *dataframe* posteriores a las mencionadas previamente.

3. Modelamiento y validación

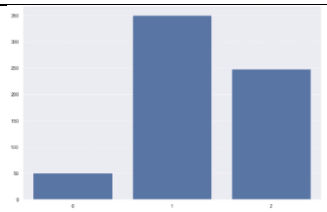
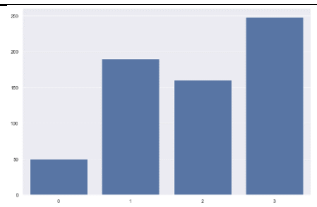
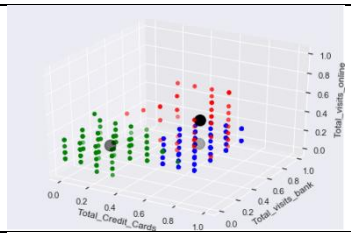
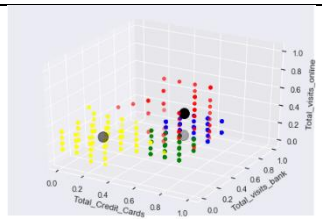
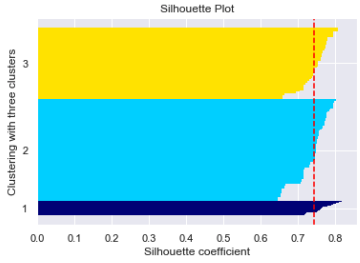
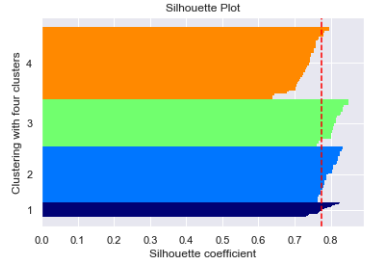
a. K-Means: Pablo Arango

utilizó el método del codo para verificar cual sería el numero óptimo de clústeres:

Método del Codo



Utilizando el resultado grafico arrojado por el método del codo, se tomaron como referencia los valore de K=3 y K=4 para el desarrollo del modelo.

Valor de K	K = 3	K = 4
Tamaño Clústeres		
Representación Grafica		
Coeficiente Silueta		
Score Silueta	0.743	0.773

Con base a los resultados desplegados en la tabla, podemos afirmar que el mejor clustering se dio cuando con 4 clústeres (K =4). Esto debido a que el puntaje de la silueta fue mayor.

b. <u>DBScan: Juan Diego Trujillo</u>

Para el modelo de DBScan es necesario tener en cuenta dos hiperparametros en particular: el primero llamado la distancia épsilon, es la distancia máxima para que dos puntos sean consideras

vecinos el uno del otro; el otro, el número mínimo de vecinos, para que un punto sea considerado como un punto *core.*

Para encontrar un buen par de hiperparametros de tal forma que ambos maximicen el score silueta, se optó por iterar al mismo tiempo que se varían estos dos valores y se encontró lo que se puede apreciar en la siguiente figura.

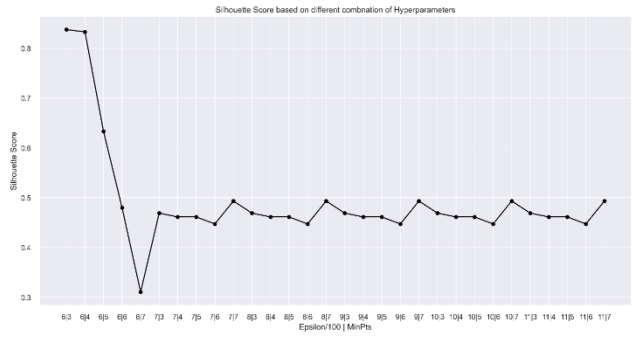


Figura 5: Score silueta según la variación de parámetros

Como se puede observar, el valor del score silueta es bastante elevado, por encima del 90%. Sin embargo, observemos como se comporta el número de clusters según estos valores.

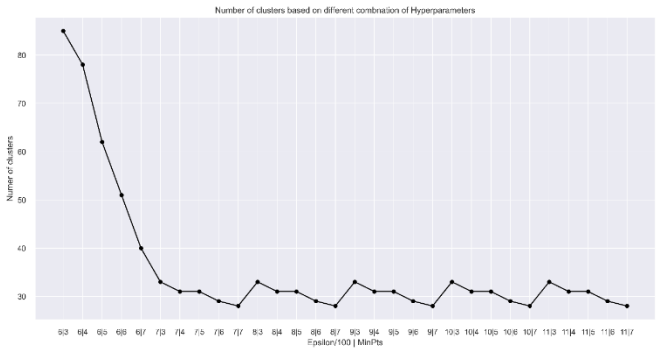


Figura 6: Número de clusters según la combinación de hiperparaemtros

Es posible ver que el número de clusters encontrado es mucho mayor de los encontrados en los otros modelos, llegando a alcanzar casi los 90 para la primera configuración planteada y esto no favorecería a la compañía, pues la idea es tener un número de clusters reducido.

A pesar de haber intentado varias configuraciones cuyos rangos, por ejemplo, iban desde 6 a 12 para la distancia épsilon y de 3 a 8 para el número mínimo de puntos, seguían dando valores muy altos para el número de *clusters*. No obstante, llegaron a encontrase valores cercanos a los 30 clusters pero así mismo el valor silueta era muy bajo (por debajo del 50%)

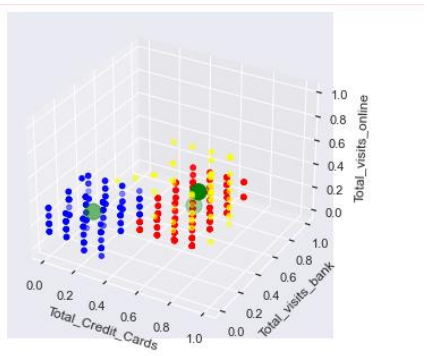
En definitiva, este modelo queda descartado y no lo recomendamos para su uso.

c. Mean shift: Santiago Moreno

Durante el modelamiento de Mean-Shift se tomaron dos posibles grupos de 3 variables para el modelo, un grupo que contenía a el *Total de tarjetas de crédito, Total de visitas al banco y Total de visitas online,* mientras que el en lugar del *Total de visitas online* utilizaba el *Total de llamadas realizadas.*

Tras generar ambos modelos, se dedujo que el modelo que generaba el mejor conjunto de cluster era el primero mencionado, se hizo la prueba comparando el puntaje de ambas siluetas.

A medida que se hicieron avances en el modelo, se hicieron cambios a los datos y se llegó a un grupo de 3 clusters, sin ningún valor atípico evidente. Se generaron múltiples gráficos, entre estos uno que mostraba los clusters en 3D, donde el cluster uno era representado con el color rojo, el 2 con azul, el 3 con amarillo y los centroides con verde:



Para comprobar la validez de este modelo, se utilizó el coeficiente de silueta. Su puntaje dentro de la silueta fue de 0.74%.

4. Comparación y conclusiones

De los resultados arrojados anteriormente se pueden concluir varias cosas. La primera es que el modelo que más se ajusta a las necesidades del proyecto es el de KMeans con K = 4 y los datos normalizados. Esto debido a que, de los tres modelos realizados, este fue el que arrojó el puntaje de silueta más alto de todos. Tal puntaje se utiliza para evaluar la calidad de los clústeres como resultado de un análisis y por tal motivo nos sirve de ayuda para verificar que modelo es el más adecuado.

Adicional a eso, se puede concluir del resultado grafico del KMeans que el numero óptimo de clústeres es 4. Teniendo eso en cuenta también se puede notar que el cluster con el menor número de observaciones es el cluster 0, en el cual sus datos presentan el menor número de visitas online pero el mayor número de visitas al banco con respecto a otros clusteres. El cluster 1 y 2 cuentan con tamaños y características muy similares, sin embargo lo que los diferencia es que los datos del cluster 2 tienen mayores valores para las visitas en línea. Por último, el cluster 3 es el más numeroso de todos y se distingue del resto porque sus datos cuentan con el menor número de tarjetas de crédito totales y visitas físicas al banco.