Introducción.

El objetivo de este reporte es desarrollar una clasificación de grupos de observaciones o clústeres desde una base de datos de un banco alemán. Este ejercicio no tiene como objetivo predecir correctamente un resultado ya conocido, sino que se debe poder generar una clasificación sólo con la información disponible, encontrando grupos cuyos miembros sean lo más uniformes posibles, a la vez que se diferencian lo máximo de los elementos de los otros grupos. Este enfoque de trabajo es conocido como no supervisado, siendo el modelo de los K medios uno de los más usados.

Selección del número de clústeres y las variables.

El modelo de los K medios tiene ese nombre porque su objetivo es clasificar un conjunto de datos en **K** clústeres mediante el cálculo del promedio de cada clúster para cada una de las variables que los datos tengan. El vector resultante con el promedio de cada una de las variables para cada clúster se conoce como el centro del clúster o *centroide*. La posición de estos centros dependerá de las variables que se usen para la clasificación y del número de clústeres en que se quiere dividir la muestra; de esto se sigue que el éxito o fracaso de la clasificación dependerá, en última instancia, de una correcta selección de estos 2 factores.

Si bien la selección de variables y valores para **K** se acerca más a un proceso dialéctico en el que el resultado de uno influye en la selección del otro, el primer paso se dio al seleccionar las variables con las que se iniciaría esta danza.

1. Las variables.

   El conjunto de datos consta de 2 variables numéricas contínuas (***amount*** y ***age***), 7 variables numéricas ordinales (***savings***, ***employed***, ***rate***, ***residence***, ***property***, ***credits*** y ***job***), 4 variables categóricas binarias (***persons***, ***telephone***, ***foreign*** y ***credit***) y 4 variables categóricas de más de 2 opciones (***status***, ***history***, ***personal*** y ***housing***).

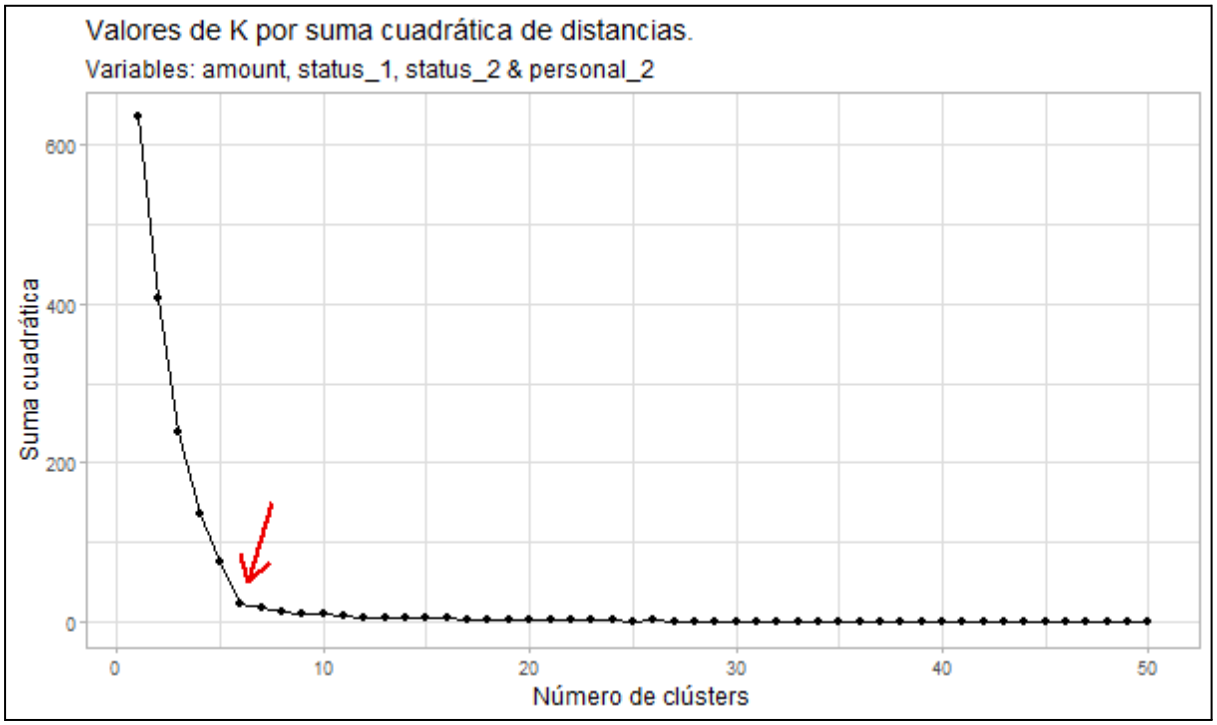
   El primer filtro aplicado fue descartar aquellas variables que tuviesen escasa varianza, pues se estimó que poco podrían aportar en la clasificación: ***foreign***, ***persons***, ***job*** y ***credits*** quedaron fuera. El siguiente paso fue transformar en variables binarias las variables categóricas no binarias para luego normalizar todas las variables y evitar el sesgo de escala. Se volvió a aplicar una segunda selección por escasa varianza y ***housing*** fue descartada.

   De estas 18 variables seleccionadas (9 variables numéricas y binarias originales más 9 variables binarias que se construyeron desde 3 variables categóricas no binarias originales) se probaron varios enfoques para iniciar la selección, siendo el más fructífero aquel que iniciaba con las 2 variables numéricas contínuas (***amount*** y ***age***), a las cuales se les agregó o quitó alguna de las otras 16 variables. La selección final se discutirá en el análisis de resultados.

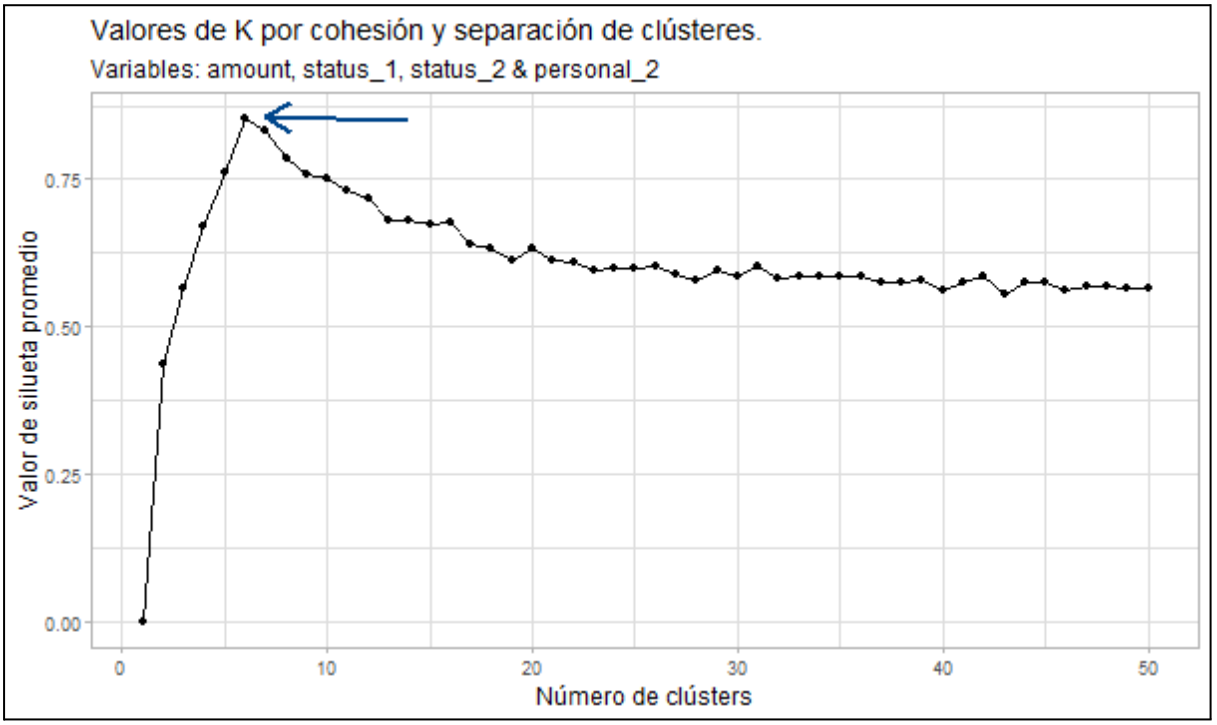
2. El valor de K.

Como se explicó anteriormente, la elección del valor de **K** se hizo en paralelo a la selección de las variables del modelo. Durante el proceso fue curioso notar que, para algunas selecciones de variables, el mejor valor de **K** era más o menos evidente dentro de las primeras 20 opciones de clústeres, pero para otras selecciones, especialmente al evaluar el valor de silueta promedio, el mejor **K** parecía no converger en ningún valor, incluso simulando 200 clústeres.

Una de las maneras para encontrar el mejor valor de **K** es mediante la determinación de *el codo*: se calcula la resta entre los valores de un registro para cada una de las variables del modelo de clasificación, y los valores que para esas variables tiene el *centroide* del clúster al cual pertenece dicho registro (recordando que el *centroide* corresponde al vector que tiene el valor promedio de cada una de las variables de entre todos los registros de ese clúster, no teniendo que corresponder necesariamente a una determinada observación del conjunto de datos); el resultado obtenido se eleva al cuadrado y se suma con todos los resultados obtenidos para todas las observaciones de todos los clústeres. Ese valor nos da la suma cuadrática total de las distancias a los *centroides* que se obtiene para un determinado valor de **K**. A mayor cantidad de clústeres que se pueden generar menor es este resultado, pero a partir de cierto valor la diferencia entre sumas cuadráticas no es tan importante. Al graficar el número de clústeres versus la suma cuadrática total se puede observar en algún punto un quiebre en la tendencia, que asemeja un codo humano, lo cual le da el sobrenombre a esta técnica.Como se ve en el gráfico de abajo, se encontró un codo claro con un valor de **K** de 6 con una suma cuadrática total de distancias a los *centroides* de ***23,47***.



Otra método que se puede usar para determinar el mejor valor de **K** es *la silueta*, una técnica que mide cuán *parecidas* son las observaciones dentro de cada clúster, a la vez que estima cuán *separadas* están de los otros clústeres. La forma de definir la *cohesión* (lo *parecido* que son las observaciones dentro de un clúster) es mediante el cálculo de la distancia promedio (en este caso, la distancia euclidiana) entre todos los registros que pertenecen a un mismo clúster; a su vez, la *separación* se calcula mediante el promedio de las distancias de cada observación del clúster a la observación más cercana de cualquier otro clúster que no sea el propio. Al valor obtenido de la *separación* se le resta el valor obtenido de *cohesión*, diferencia que es dividida por el valor mayor entre estos dos, *cohesión* y *separación*, obteniéndose finalmente el valor de silueta que es una cifra entre -1 y 1: mientras más cercano a 1 mejor definidos los clústeres y más diferenciados de otros estarán. Como se observa en el gráfico inferior, al igual que en el gráfico de el codo el valor de **K** óptimo es de 6 clústeres, para un valor de silueta promedio de ***0,854***.



Finalmente, podemos evaluar los clústeres según su valor de interpretabilidad, es decir ¿del número de clústeres que genere se puede sacar alguna información relevante? ¿Los valores de las variables de los *centroides* dicen algo? En este sentido, los clústeres generados con un valor de **K** de 6 también pueden decir algo, pero esta idea se desarrollará mejor cuando se realice el análisis de los resultados obtenidos.

Análisis de resultados.

Luego de todo el estudio de variables y valores de **K**, el modelo de K medias quedó de la siguiente forma:

1. Variables usadas:

   amount *= monto del crédito.*
   status_1 *= persona sin cuenta corriente.*
   status_2 *= persona con cuenta corriente con deuda.*
   personal_2 *= hombre soltero o mujer no soltera.*

2. Cantidad de clústeres generados (valor de K): ***6***.

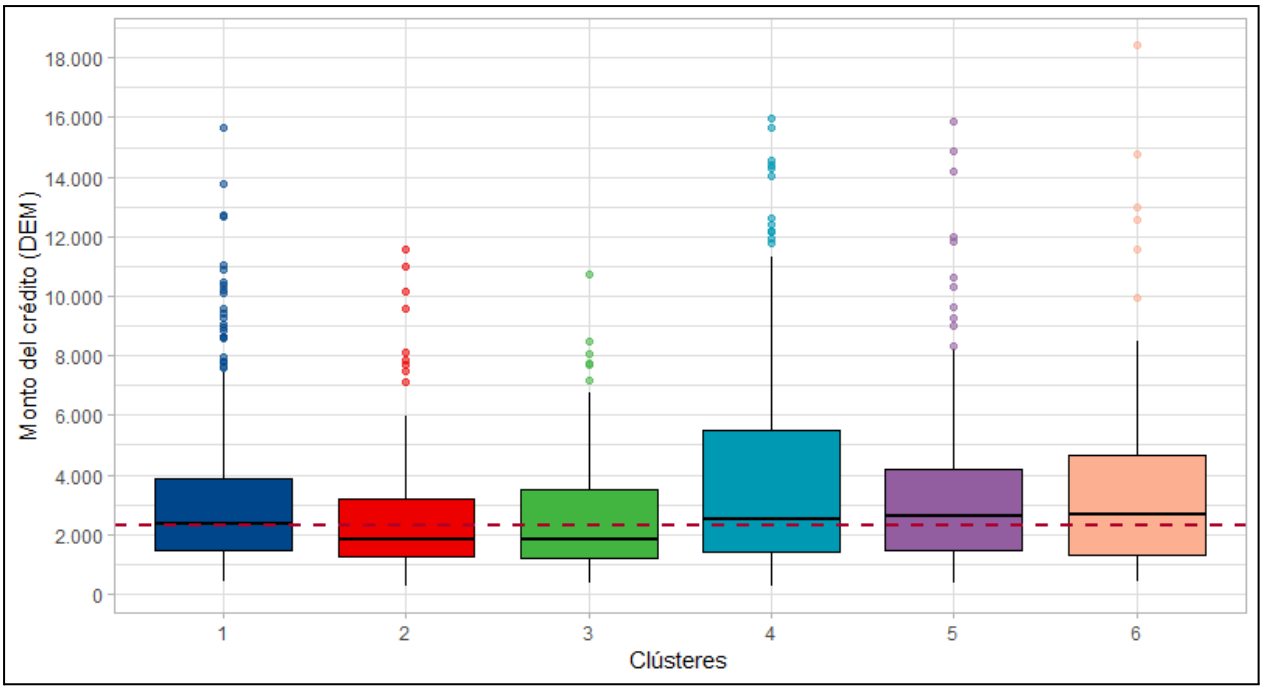
Los centroides de cada clúster quedaron definidos como se ve en la siguiente tabla (si bien para el cálculo de los clústeres se usó el promedio, en el caso de ***amount*** se muestra la mediana, que será de mayor utilidad para la interpretación del resultado):

Clúster	amount	status_1	status_2	personal_2
1	2333.0	0	0	0
2	1840.5	0	0	1
3	1838.5	1	0	1
4	2520.0	0	1	0
5	2589.5	1	0	0
6	2651.0	0	1	1

El tamaño de cada uno de los clústeres es:

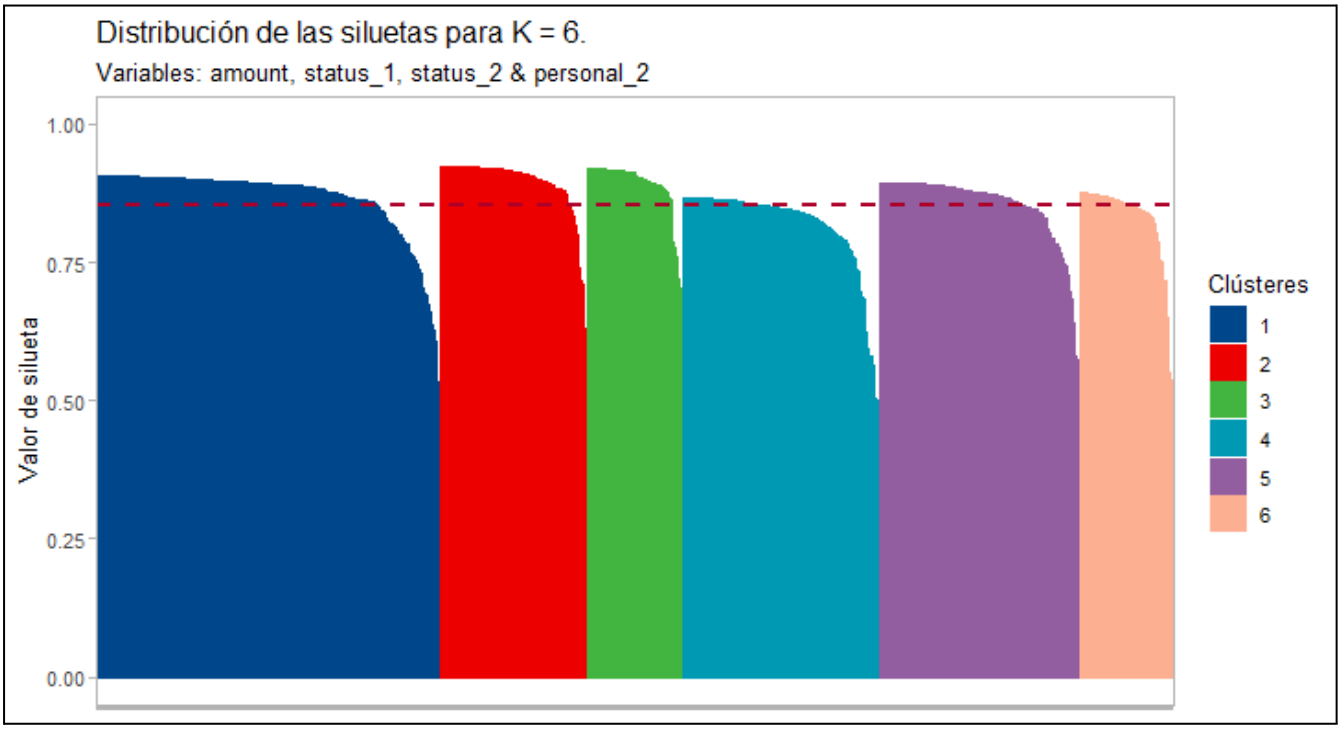
- ***Clúster 1***: 321 observaciones.
- ***Clúster 2***: 136 observaciones.
- ***Clúster 3***: 88 observaciones.
- ***Clúster 4***: 183 observaciones.
- ***Clúster 5***: 186 observaciones.
- ***Clúster 6***: 86 observaciones.

Al mirar los valores de los *centroides* se puede observar que las variables binarias quedan comprendidas dentro de algún clúster de forma completa. Sin embargo, la variable ***amount*** es más difícil de interpretar, pues es numérica contínua y las cifras obtenidas no parecen estar muy diferenciadas. Para lograr una mejor interpretación se comparó el valor de la mediana para ***amount*** de cada uno de los clústeres con el valor de la mediana para todo el conjunto de datos de la misma variable (cuyo valor es de ***2.320***), resultando en 3 grupos de clústeres claramente identificables: 3 clústeres con un *elevado monto de crédito solicitado* (el 4, el 5 y el 6), 1 clúster con un *mediano monto de crédito solicitado* (el 1, el más cercano a la mediana total) y 2 clústeres con un *bajo monto de crédito solicitado* (el 2 y el 3). Al mirar el tamaño de las cajas del gráfico de *boxplot* resalta que el grupo del bajo monto presenta los datos con menor dispersión, mientras que en el grupo con mayor monto están los clústeres con mayor dispersión.



Para ver qué tan pertinentes son estos 6 grupos de clasificación se puede observar *la silueta promedio* de cada uno, evaluando si la *cohesión* dentro del clúster y la *separación* con otros clústeres es adecuada. Recordando que el valor de *cohesión*/*separación* promedio para todos los datos es de 0,854, los valores de silueta promedio obtenidos para cada clúster son: ***0,86*** para el clúster 1; ***0,89*** para el clúster 2; ***0,89*** para el clúster 3; ***0,81*** para el clúster 4; ***0,85*** para el clúster 5; ***0,82*** para el clúster 6. Al estar todos estos valores cercanos a 1, es posible decir que la clasificación obtenida es exitosa.

Las siluetas de todos los clústeres pueden verse en el siguiente gráfico.



Con la información ya expuesta fue posible otorgar una etiqueta a cada clúster, fijándose en las diferencias relevantes que presentan:

- ***Clúster 1***: Clientes con cuenta corriente y alguna cantidad ahorrada, mujeres solteras u hombres no solteros, que solicitaron un monto de crédito medio.
- ***Clúster 2***: Clientes con cuenta corriente y alguna cantidad ahorrada, mujeres no solteras u hombres solteros, que solicitaron un monto de crédito bajo.
- ***Clúster 3***: Clientes sin cuenta corriente, mujeres no solteras u hombres solteros, que solicitaron un monto de crédito bajo.
- ***Clúster 4***: Clientes con cuenta corriente con deudas, mujeres solteras u hombres no solteros, que solicitaron un monto de crédito alto.
- ***Clúster 5***: Clientes sin cuenta corriente, mujeres solteras u hombres no solteros, que solicitaron un monto de crédito alto.
- ***Clúster 6***: Clientes con cuenta corriente con deudas, mujeres no solteras u hombres solteros, que solicitaron un monto de crédito alto.

Retomando un punto que quedó esbozado en la sección anterior, estas etiquetas no sólo permiten clasificar a los clientes, sino que también parecen tener sentido de interpretabilidad, lo cual le da más fuerza a la segmentación de las observaciones en 6 clústeres usando las variables ya mencionadas.

Con los clústeres ya caracterizados, podemos hipotetizar sobre la utilidad práctica de esta segmentación. Por ejemplo, pensando en apuntar a un cierto “tipo de cliente” para ofrecerle una cuenta corriente en el banco en el cual está solicitando el crédito, nos fijaríamos en aquellas personas que sean clasificadas en los clústeres 3 y 5, adoptando distintas estrategias de márketing para cada uno de estos segmentos. Las posibilidades son infinitas, pues al ser un modelo de análisis no supervisado, la interpretación de la clasificación dependerá del objetivo que queramos alcanzar, teniendo muchos caminos a seguir y no dependiendo de la adecuación a un resultado ya conocido.

Evaluación 3a
Curso Minería de datos
Wladimir Richard Parada Rebolledo
Néstor Patricio Rojas Ríos