Tarea 5: Aprendizaje no supervisado

Índice

1. Introducción

El objetivo de este artículo es realizar una exploración detallada del método de clustering K-means, uno de los enfoques clásicos y ampliamente utilizados en la literatura contemporánea. Aunque existen otros métodos, tanto tradicionales como avanzados, en esta ocasión nos centraremos exclusivamente en el análisis y las aplicaciones del K-means.

1.1. Clustering y sus Aplicaciones

Los métodos de clustering, también conocidos como técnicas de agrupamiento, tienen como objetivo principal dividir un conjunto de datos en grupos o çlusters.[en] función de similitudes entre los datos. Estos grupos deben ser tal que los datos dentro de un grupo sean más similares entre sí que con los datos de otros grupos. Los métodos de clustering son ampliamente utilizados en diversas disciplinas y aplicaciones. Aquí te presento algunas de sus principales utilidades:

- Segmentación de Mercado: Las empresas utilizan el clustering para segmentar a sus clientes en diferentes grupos según sus comportamientos de compra, preferencias, características demográficas, entre otros. Esto permite implementar estrategias de marketing dirigidas y más efectivas.
- Reducción de Dimensionalidad: En conjuntos de datos de alta dimensionalidad, el clustering puede ser utilizado para crear representaciones reducidas de los datos, agrupando características similares.
- Detección de Anomalías: Los datos que no se agrupan claramente en ningún cluster pueden ser considerados como anomalías o outliers. Esto es útil en aplicaciones como la detección de fraudes.

2. Descripción y Contexto de los Datos

El conjunto de datos con el que trabajamos proviene de un subconjunto refinado de la base de créditos de la plataforma "Lending Clubçorrespondiente al período 2007-2018. Hemos retenido únicamente los créditos que se encuentran completamente saldados y aquellos que han incurrido en incumplimiento. Esto nos proporciona un total de 261,185 registros, que hemos dividido en conjuntos de entrenamiento y prueba. A través de un proceso de selección de variables, decidimos conservar 22 características. las cuales se encuentran estandarizadas para asegurar su compatibilidad.

3. Método K-means

3.1. Descripción

El algoritmo K-means es una técnica de clustering que busca particionar un conjunto de N puntos en K clusters en el que cada punto pertenece al cluster cuyo centroide es más cercano. Matemáticamente, el objetivo es minimizar la suma de las distancias al cuadrado entre cada punto y el centroide de su cluster.

Dado un conjunto de observaciones $\{x_1, x_2, \ldots, x_N\}$, donde cada observación es un vector d-dimensional, K-means busca particionar las N observaciones en K ($K \leq N$) conjuntos $S = \{S_1, S_2, \ldots, S_K\}$ con el objetivo de minimizar la suma de las distancias al cuadrado entre los puntos y los centroides de los clusters:

$$\min \sum_{i=1}^{K} \sum_{x \in S_i} ||x - \mu_i||^2$$

donde μ_i es el centroide del cluster S_i y puede ser calculado como:

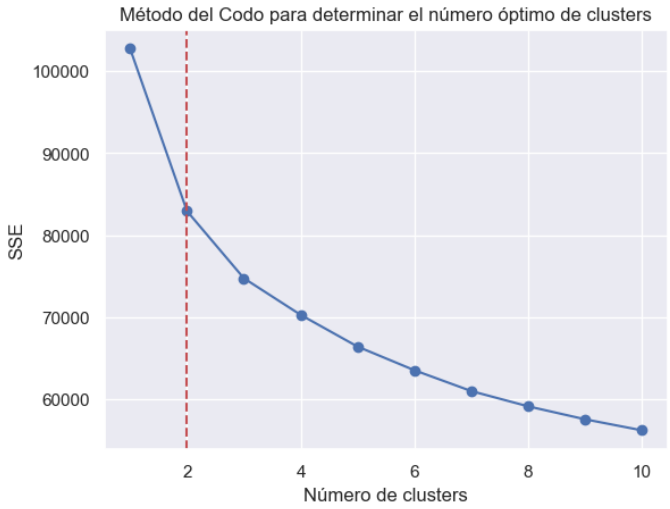
$$\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x$$

3.2. Elección numero de clusters

Seleccionar el número óptimo de clusters (k) en K-means es fundamental para obtener resultados significativos.

3.2.1. Inercia - Criterio del codo

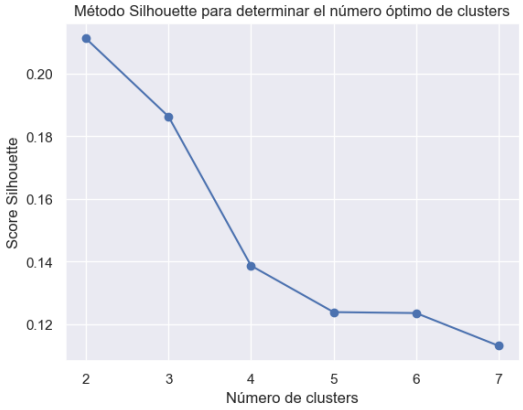
Este método implica correr el algoritmo de K-means para un rango de valores de k (por ejemplo, k de 1 a 10), y luego para cada valor de k se computa la suma de errores cuadrados (SSE). Cuando se grafica la línea de SSE para cada valor de k, el punto donde la SSE comienza a estabilizarse y a producirse un çodo.[es] generalmente considerado un buen indicador del número óptimo de clusters



Nuestro conjunto de datos no muestra un punto de inflexión marcado, comúnmente conocido como çodo", en la evaluación de la cantidad óptima de clusters. Sin embargo, al observar la variación en la suma de las distancias al cuadrado respecto al número de clusters, notamos que la tasa de cambio empieza a desacelerarse y los valores parecen comenzar a estabilizarse en $K = 2$

3.2.2. Score Silhouette

El coeficiente de Silhouette mide cuán cerca cada punto en un cluster está de los puntos en los clusters vecinos. Los valores varían entre -1 y 1, donde un valor alto indica que el objeto está bien emparejado con su propio cluster y mal emparejado con los clusters vecinos. Si la mayoría de los objetos tienen un valor alto, entonces la configuración del clustering es apropiada. Si muchos puntos tienen un valor bajo o negativo, entonces la configuración del clustering puede tener demasiados o muy pocos clusters.



En nuestro caso, el valor más alto del coeficiente Silhouette se logra para $K = 2$, lo que sugiere que esta es la cantidad más adecuada de clusters para nuestro conjunto de datos. Es importante señalar que un coeficiente Silhouette alto es indicativo de una estructura de clusters bien definida.

3.2.3. Índice Davies-Bouldin

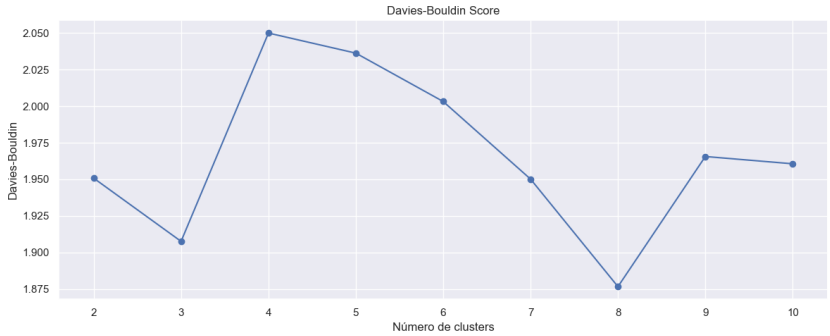
Es una métrica utilizada para evaluar la calidad de las particiones obtenidas a través de algoritmos de clustering. Se basa en el cociente entre la suma de la dispersión interna de dos clusters y la distancia entre los centros de estos clusters. La idea es que buenos clusters tendrán baja dispersión interna y estarán alejados entre sí.

Formalmente, el índice Davies-Bouldin DB para k clusters se define como:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} \left(\frac{S_i + S_j}{D_{ij}} \right)$$

Donde:

- S_i es la dispersión media de cada punto desde el centroide en el cluster i.
- D_{ij} es la distancia entre los centroides de los clusters i y j.
- k es el número total de clusters.



La métrica proporciona una relación entre la distancia media dentro de un cluster y la distancia media entre clusters. Para el índice Davies-Bouldin, valores más bajos ndican una mejor partición. Bajo este criterio la mejor particion serian 8 clusters.

3.2.4. Índice Calinski-Harabasz

tambien conocido como score V de Calinski-Harabasz. Este índice ofrece una relación entre la dispersión entre clusters y la dispersión dentro de los clusters. Un valor más alto indica mejores resultados de clustering, ya que significa que los clusters están bien separados entre sí y densamente empaquetados.

Formalmente, el índice Calinski-Harabasz CH se define como:

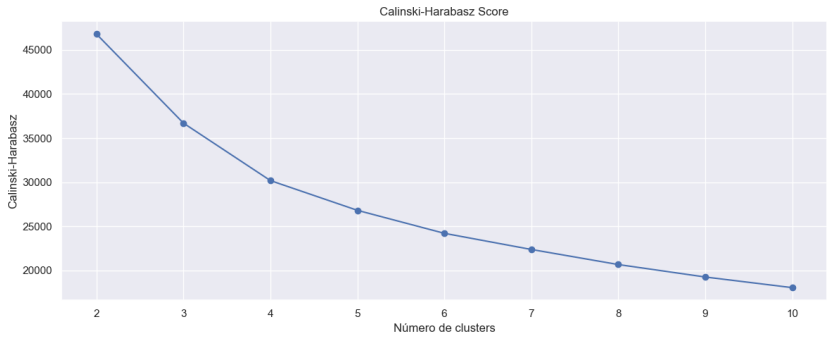
$$CH = \frac{B/(k-1)}{W/(n-k)}$$

Donde:

- B Dispersión entre clusters, que es la suma de las distancias cuadradas entre los centroides de los clusters y el centroide global, ponderado por el número de puntos en cada cluster.
- W Dispersión dentro de los clusters, que es la suma de las distancias cuadradas entre cada punto y su centroide de cluster.
- k Número de clusters.

- n Número total de puntos.

Un valor más alto indica que la separación entre clusters es significativamente mayor que la dispersión dentro de los clusters, lo que es deseable en un buen clustering.

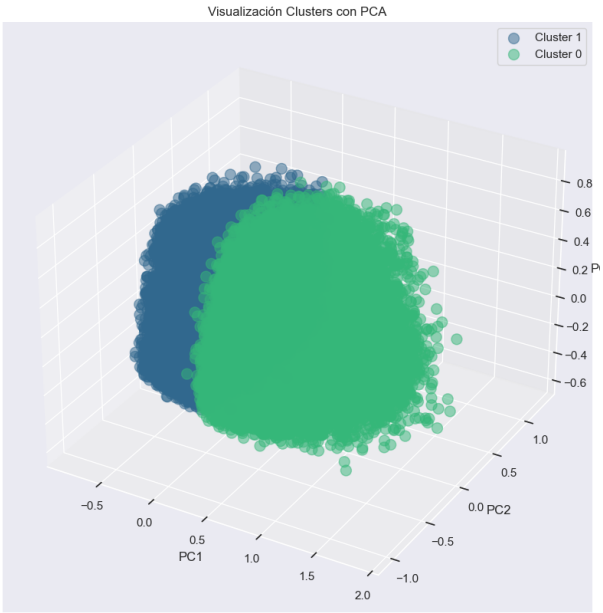


3.3. Interpretación Resultados

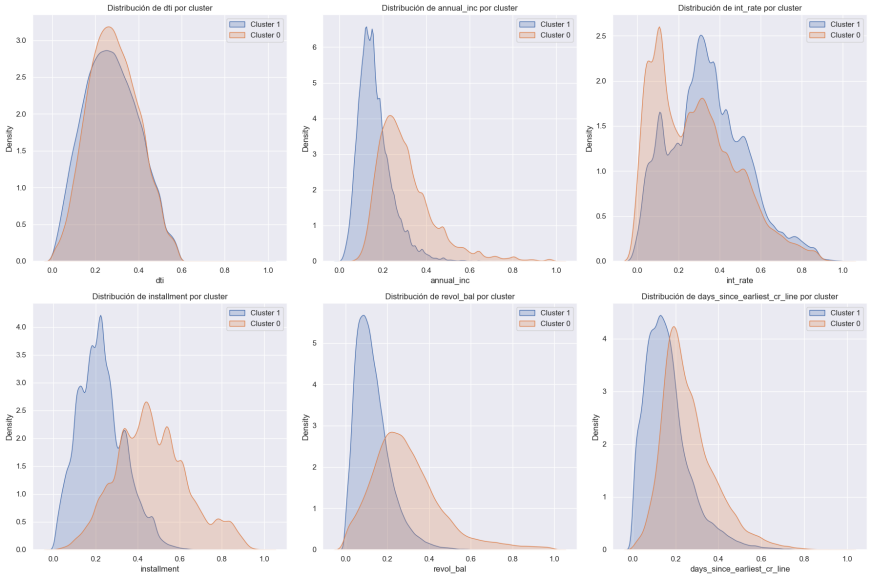
Después de optar por dividir los datos en dos clusters, eleccion respaldada por la concordancia entre 3 de los 4 criterios empleados, que señalaban $K = 2$ como el número óptimo de clusters, esta decisión se alinea coherentemente con la naturaleza de nuestros datos. Principalmente, estamos tratando con dos categorías distintas: préstamos "buenosz préstamos "malos".

A continuación, se presenta el análisis de los resultados. Para una interpretación y visualización más sencilla, destacaremos únicamente seis variables.

Debido a la alta dimensionalidad de nuestras variables, decidimos aplicar una reducción de dimensionalidad para visualizar nuestros clusters. Optamos por el método PCA, conservando los tres componentes principales, los cuales explican el 55,5 % de la varianza total.



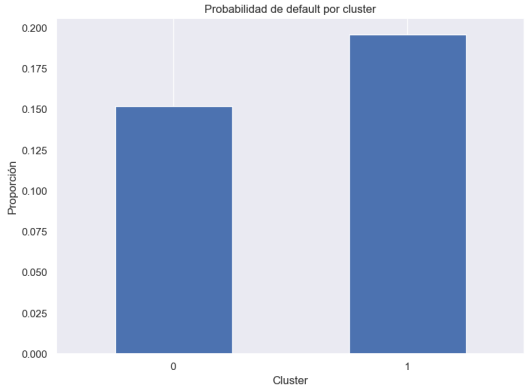
cluster	dti	annual_inc	int_rate	installment	revol_bal	days_since_earliest_cr_line
0	19.39	96,900.58	11.99	667.94	28,325.94	6,673.12
1	18.70	53,625.71	13.46	332.00	12,604.83	4,535.52



Con base en las gráficas de distribución y en los centroides , podemos catalogar los clusters de la siguiente manera

- Cluster 0: Clientes con una mayor antigüedad, cuyos ingresos anuales superan el promedio. Gracias a su sólido historial crediticio, cuentan con una línea de crédito revolvente más amplia. Esta trayectoria les permite acceder a préstamos con tasas de interés más bajas, y suelen tener pagos mensuales altos y un índice de ingresos/deuda ligeramente superior.
- Cluster 1: Se trata de clientes más recientes con ingresos que oscilan entre medios y bajos. Generalmente, debido al mayor riesgo asociado, se les otorgan créditos con tasas de interés más elevadas, tienden a presentar pagos mensuales más bajos.

Generalmente, podríamos interpretar a los clientes del grupo 1 como de mayor riesgo en comparación con los del grupo 0. Esta percepción se puede corroborar observando la proporción de créditos en default en cada categoría. Sin embargo, es importante señalar que el aprendizaje no supervisado, por sí mismo, no es el método más adecuado para aplicar en metodologías de credit scoring.



4. Referencias

- sciencedirect.com/science/article/abs/pii/S0038012119305440
- es.wikipedia.org/wiki/K-medias
- sciencedirect.com/science/article/abs/pii/S0038012119305440