“Inteligencia de Negocios: Laboratorio 1”

Felix S. Rojas Casadiego, Juan S. Alegría Z., Daniel Reales
Universidad de los Andes, Bogotá, Colombia
{fs.rojas, j.alegria, da.reales}@uniandes.edu.co
Fecha de presentación: febrero 19 de 2023

Tabla de contenido

1. Entendimiento de los datos

Se quiere obtener información relevante para BiciAlpes sobre la seguridad en las vías. Por esto, se quiere conocer cuáles son los factores que más impactan en los accidentes viales que involucran ciclistas, lo que también podría ayudar a las autoridades y planificadores urbanos en la implementación de mecanismos que reduzcan la ocurrencia de accidentes, así como otros planes de movilidad sostenible.

En los datos recopilados de fuentes abiertas de la alcaldía se pudo observar que existe una columna que indica la severidad de un accidente de acuerdo con si este es fatal, serio o leve. Por tanto, se buscará aplicar tres algoritmos de clustering para poder entender los factores que indicen en la severidad de los accidentes.

El entendimiento de los datos es una etapa fundamental en todo proyecto de análisis de datos. En esta fase, se busca comprender la calidad y características de los datos que se tienen a disposición, de tal manera que se puedan identificar los procesos de limpieza y preparación necesarios para lograr los objetivos del proyecto. Uno de los primeros pasos es determinar si los datos son suficientes para el alcance del proyecto. En caso de serlo, se procede a realizar un perfilamiento completo que incluya estadística descriptiva y gráficos sobre los datos. En este sentido, se debe tener en cuenta el número de datos disponibles, es decir, las filas y columnas que conforman el conjunto de datos.

En el caso particular de los datos de BiciAlpes, se cuenta con un total de 5338 filas y 15 columnas. En cuanto a las variables, se tienen diferentes tipos de variables, como categóricas y numéricas. En el caso de las variables numéricas, se pueden determinar diferentes estadísticos descriptivos como la media, la varianza, la desviación estándar, entre otros. En el caso particular de las variables numéricas proporcionadas, se tiene la variable "Number_of_Vehicles" que es numérica y representa el número de vehículos involucrados en el accidente. La media para esta variable es de 1.00, lo que significa que en promedio hay un solo vehículo involucrado en cada accidente. Por otro lado, la variable "Speed_limit" también es numérica y representa el límite de velocidad. La media para esta variable es de 33.52, lo que indica que la mayoría de los accidentes ocurrieron en carreteras donde el límite de velocidad es relativamente alto.

Por otro lado, en el caso de las variables categóricas, es importante conocer las categorías y en qué proporción se presentan. La variable "Accident_severity" es categórica y representa la severidad del accidente (Fatal = 1, Serio = 2, Leve = 3). En este caso, se puede observar que la mayoría de los accidentes son de gravedad leve (con un valor de 3), seguidos por los de gravedad seria (con un valor de 2), y finalmente los de gravedad fatal (con un valor de 1).

Es importante mencionar que una parte fundamental de esta etapa está relacionada con el análisis a nivel de calidad de datos y, en particular, a nivel de las dimensiones de calidad, como la completitud, unicidad, consistencia y validez. Este análisis permite identificar posibles errores o inconsistencias en los datos, lo que lleva a la necesidad de realizar actividades de preparación adicionales. En el caso de los datos proporcionados, se puede observar que la columna "Unnamed: 14" no presenta valores y, por lo tanto, no proporciona información útil para el análisis.

2. Preparación de datos

La preparación de datos es un proceso fundamental en el análisis de datos que permite transformar los valores actuales de acuerdo con los algoritmos que se van a utilizar y el objetivo del negocio a resolver. El proceso de preparación de datos incluye la limpieza de los datos, que implica el manejo de los datos nulos (missing values) y los valores atípicos (outliers).

Para esto, se creó un pipeline que tiene como propósito ser utilizado para limpiar los datos de forma general antes de pasarlos a cualquier limpieza necesaria para cualquier algoritmo en particular. Se realizó one-hot encoding, imputación y normalización, todo consolidado en un pipeline para su reutilización.

En la sección de carga y exploración de datos, se utilizaron los datos "Datos_BiciAlpes.csv", que se cargaron

en un DataFrame y se observaron los primeros 20 resultados. Luego se observaron las distribuciones de las variables generando diagramas de barras. Esto se hizo para ver si había outliers y tratar de aplicar transformaciones para corregir esto.

Después, se eliminó una variable que resultó ser irrelevante (Vehicle_Type) y se eliminó otra columna (Unnamed: 14) debido a la gran cantidad de valores nulos.

Finalmente, se realizó un análisis de las variables categóricas de la base de datos y se convirtieron a string para su posterior análisis. También se reemplazaron algunas variables por etiquetas del diccionario para las que faltaban y se mostró su distribución.

3. Modelamiento

3.1 Algoritmo de K-Means

El algoritmo k-means es un método utilizado en el análisis de datos y el aprendizaje automático para agrupar datos en grupos similares. El objetivo es encontrar k grupos (donde k es un número predefinido) que minimicen la distancia media entre los puntos de cada grupo.

El algoritmo comienza seleccionando k puntos aleatorios como centros iniciales de los clusters. A continuación, cada punto de datos se asigna al grupo cuyo centro esté más próximo en términos de distancia euclídea.

Una vez asignados todos los puntos a un grupo, se calcula el centroide de cada grupo, que es el punto medio de todos los puntos del grupo. Los centroides se convierten en los nuevos centros de los grupos.

Este proceso se repite de forma iterativa: en cada iteración, los puntos se reasignan a los grupos más cercanos y se vuelven a calcular los centroides. El algoritmo se detiene cuando los centroides de los grupos ya no cambian o cuando se alcanza un número máximo de iteraciones.

Una vez finalizado el proceso de agrupación, los grupos pueden utilizarse para analizar los datos y tomar decisiones fundamentadas basadas en las similitudes y diferencias entre los grupos.

El algoritmo kmeans tiene algunos problemas. Por ejemplo, no siempre encuentra la solución óptima para agrupar los datos. Esto se debe a que el algoritmo puede quedar atrapado en un mínimo local y no ser capaz de encontrar la mejor solución global. En otras palabras, el resultado final puede depender de los centros iniciales de los grupos seleccionados aleatoriamente. El número de grupos a encontrar también debe seleccionarse antes de ejecutar el algoritmo. Si el número de grupos elegido no es el adecuado, el resultado de la agrupación puede ser inútil. Si se selecciona un número de grupos demasiado grande, los grupos pueden ser demasiado pequeños y no tener sentido. Por otro lado, si se selecciona un número de grupos demasiado pequeño, los grupos pueden ser demasiado grandes y no distinguir las diferencias entre los datos. Además, el algoritmo k-means se basa en la distancia euclidiana para asignar los puntos de datos a los grupos. Esto significa que el algoritmo sólo es eficaz para agrupar datos que tienen límites de grupos lineales. Si los datos tienen límites de grupo complejos o no lineales, el algoritmo k-means puede producir una agrupación imprecisa o ineficaz.

En el algoritmo de K-Means, se hizo la elección de los hiperparámetros utilizando el método del codo y el coeficiente de silueta para así encontrar el número óptimo de grupos.

3.2 Algoritmo de Gaussian Mixture

El algoritmo Gaussian Mixture es un método utilizado en el análisis de datos y el aprendizaje automático para modelar la distribución de probabilidad de los datos y agruparlos en grupos similares. El objetivo es encontrar k grupos (donde k es un número predefinido) que maximicen la verosimilitud de los datos observados.

El algoritmo comienza seleccionando k distribuciones normales aleatorias como centros iniciales de los clusters. A continuación, cada punto de datos se asigna al grupo cuya distribución esté más próxima en términos de probabilidad.

Una vez asignados todos los puntos a un grupo, se actualizan los parámetros de las distribuciones normales para reflejar los datos del grupo. Los parámetros incluyen la media, la varianza y la ponderación de la distribución.

Este proceso se repite de forma iterativa: en cada iteración, los puntos se reasignan a los grupos más probables y se actualizan los parámetros de las distribuciones. El algoritmo se detiene cuando la diferencia entre la verosimilitud de dos iteraciones consecutivas es menor que un umbral predefinido o cuando se alcanza un número máximo de iteraciones.

Una vez finalizado el proceso de agrupación, los grupos pueden utilizarse para analizar los datos y tomar decisiones fundamentadas basadas en las similitudes y diferencias entre los grupos.

A diferencia del algoritmo k-means, el algoritmo Gaussian Mixture es capaz de modelar distribuciones complejas y no lineales. Además, el número de grupos a encontrar no necesita ser seleccionado antes de ejecutar el algoritmo, ya que el algoritmo puede determinar automáticamente el número óptimo de grupos

utilizando criterios de información, como el criterio de información de Akaike o el criterio de información bayesiano. Sin embargo, el algoritmo es más computacionalmente costoso que k-means y puede tener dificultades para converger en conjuntos de datos grandes o complejos.

3.3 Algoritmo de K-Modes

El algoritmo k-modes es una técnica de agrupamiento utilizada en análisis de datos y aprendizaje automático para agrupar variables categóricas. El objetivo es encontrar k grupos (donde k es un número predefinido) que minimicen la distancia entre los modos de cada grupo.

El algoritmo comienza seleccionando k modos aleatorios como centros iniciales de los clusters. A continuación, cada punto de datos se asigna al grupo cuyo modo esté más próximo en términos de distancia Hamming.

Una vez asignados todos los puntos a un grupo, se calcula el modo de cada grupo, que es el valor más común en cada variable categórica. Los modos se convierten en los nuevos centros de los grupos.

Este proceso se repite de forma iterativa: en cada iteración, los puntos se reasignan a los grupos más cercanos y se vuelven a calcular los modos. El algoritmo se detiene cuando los modos de los grupos ya no cambian o cuando se alcanza un número máximo de iteraciones.

Una vez finalizado el proceso de agrupación, los grupos pueden utilizarse para analizar los datos y tomar decisiones fundamentadas basadas en las similitudes y diferencias entre los grupos.

El algoritmo de k-modes también tiene algunas limitaciones. Al igual que en k-means, la elección del número de grupos adecuado es crucial para obtener una buena solución. Además, la elección de la distancia adecuada para medir la similitud entre los puntos de datos y los modos también es importante. Otras decisiones importantes incluyen la inicialización de los modos y el criterio de parada adecuado para evitar iterar indefinidamente.

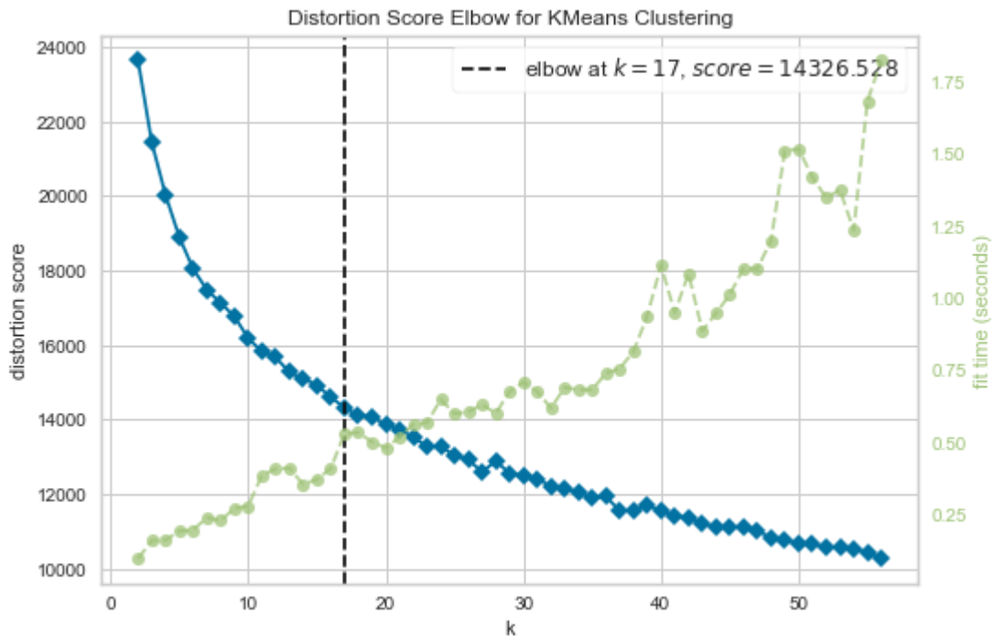
En resumen, el algoritmo de k-modes es una técnica de agrupamiento útil para datos categóricos, pero al igual que cualquier algoritmo de agrupamiento, requiere decisiones cuidadosas en la configuración de sus hiperparámetros para obtener resultados precisos y útiles.

4. Validación

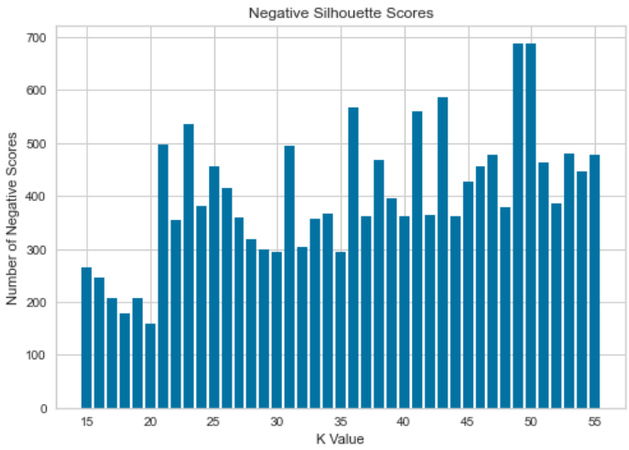
4.1 Validación cuantitativa

4.1.1 K-means

Se logró determinar inicialmente que, con el método del codo, el número óptimo de clusters está en el rango de 17 clusters.



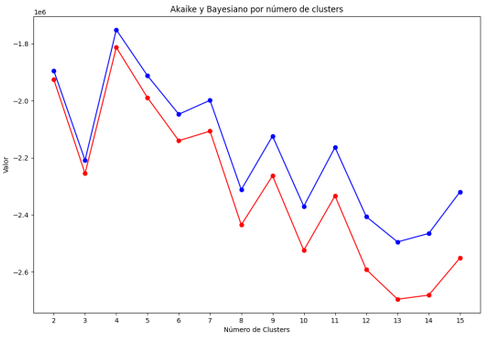
El coeficiente de silueta utiliza un parámetro de distancia para medir lo lejos que está un punto de su cluster en comparación con el centroide de otro cluster diferente. Si este valor es negativo, este punto de datos está más cerca de otro clúster que el cluster asignado. Por ello, optimizando el número de valores negativos del coeficiente de silueta para un número k, se llegó a que con 20 clusters se obtiene el menor valor de puntajes negativos.



Por otro lado, se utilizaron hiperparámetros como “random_state=32” para poder replicar los resultados al iniciar siempre en el mismo estado y “init=’k-means++’”: para seleccionar los centros de los clusters iniciales de forma inteligente para acelerar la convergencia.

4.1.2 Gaussian Mixture

Vamos a utilizar el algoritmo de Gaussian Mixture para realizar el primer procesamiento de clustering. El hiperparametro sobre el cual debemos optimizar las dos medidas de costo (El AIC y el BIC) es el número de clusters.
Para poder saber el número de clusters optimo vamos a entrenar el algoritmo para un rango de 2 a 15 clusters. Si hay alguno que minimiza el AIC y el BIC tomaremos ese valor para el número de clusters.



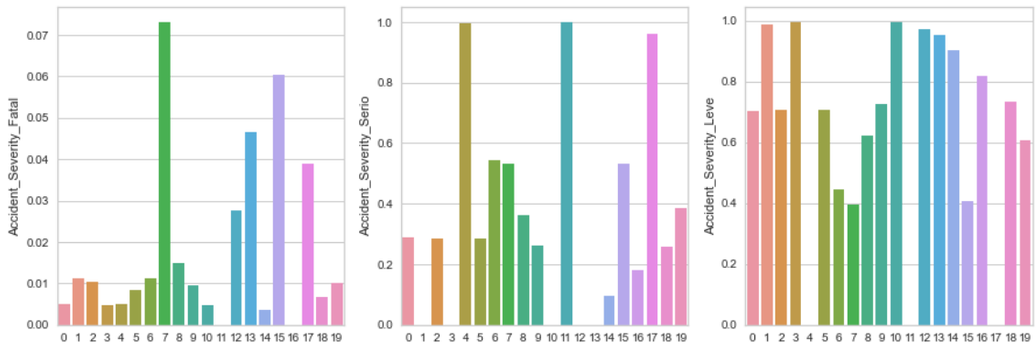
4.1.3 K-modes

Inicialización: como en k-means, la inicialización de los modos puede afectar significativamente el resultado del algoritmo. Una opción común es inicializar los modos de forma aleatoria varias veces y elegir la mejor solución.
Distancia: para medir la distancia entre un punto de datos y un modo, es común utilizar la distancia Hamming, que mide la cantidad de atributos diferentes entre los dos puntos. Otras medidas de distancia también se pueden utilizar, como la distancia de Jaccard o la distancia euclidiana modificada.
Número de clusters (k): elegir el número adecuado de clusters es importante para obtener una buena solución. Una forma común de hacerlo es probar diferentes valores de k y elegir el que maximice una medida de calidad, como la silueta o la suma de las distancias cuadradas dentro del cluster.
Criterio de parada: como en k-means, es importante elegir un criterio de parada adecuado para evitar iterar indefinidamente. Los criterios comunes incluyen un número máximo de iteraciones, convergencia de los modos o una disminución mínima en la función objetivo.

4.2 Validación cualitativa

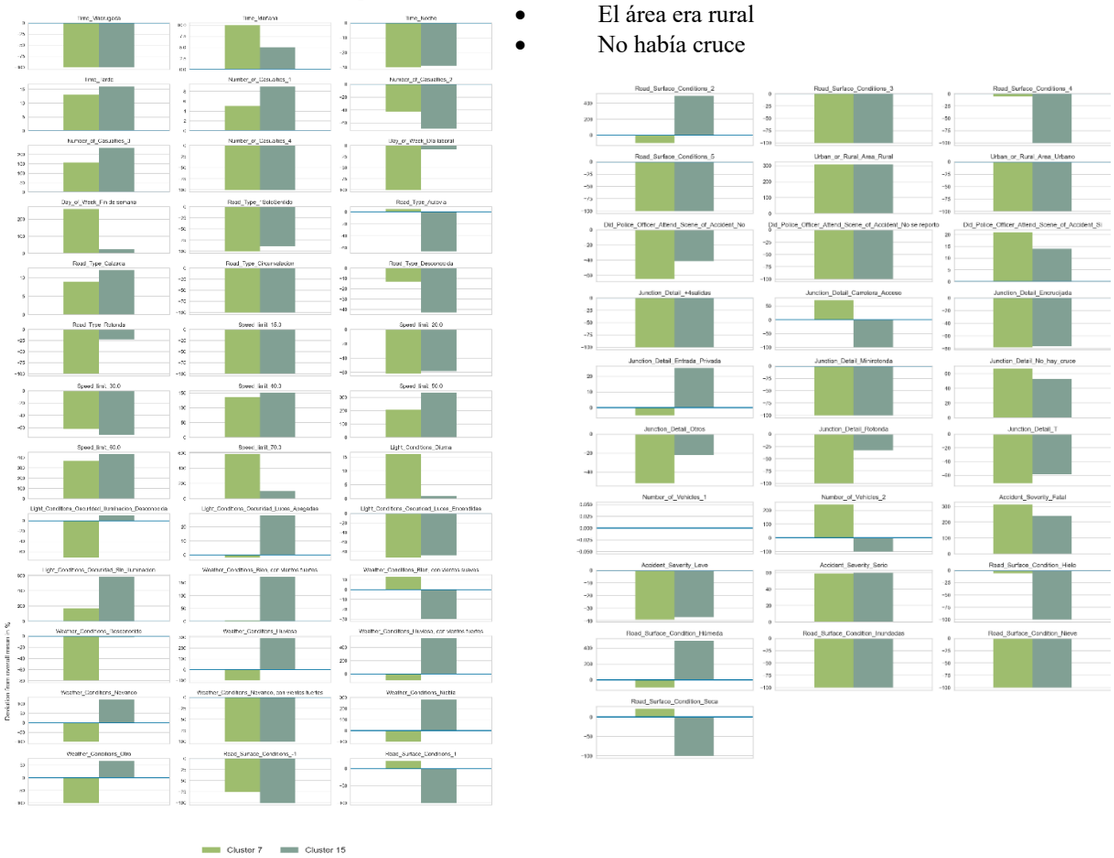
4.2.1 K-means

Los resultados cualitativos muestran que los clústeres 7, 13, 15 y 17 son aquellos cuyos centroides tienen un mayor valor en los accidentes fatales. Se puede también observar que los clusters 7 y 15 tienen un valor bajo en los demás tipos de accidentes, mientras que los demás clusters tienen valores mayores en accidentes serios y leves, por lo que no son tan buen indicativo de los accidentes fatales como los clusters 7 y 15.



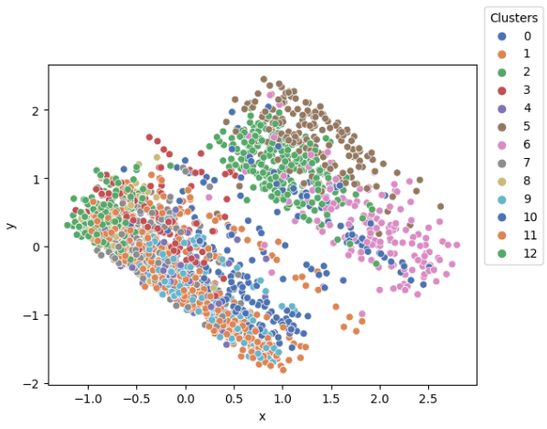
Se puede observar que, la comparación de la media de los dos clusters sobre la media general de todos los clusters nos da que las características que se repiten con respecto a cada cluster y que podrían ser la razón de los accidentes fatales y serios son las siguientes (en general):

- Hubo 3 víctimas, lo que indica que para ser considerado fatal o serio debe haber a partir de este número de personas afectadas
- El tipo de vía era calzada
- El límite de velocidad estaba por encima de los 30 km/h
- El área era rural
- No había cruce

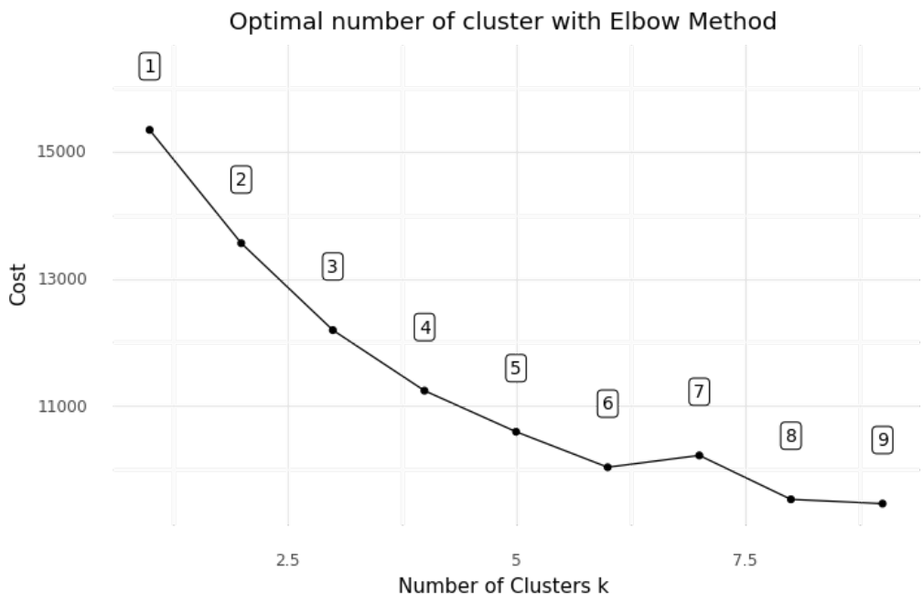


4.1.2 Gaussian Mixture

Vemos que tanto el Akaike como el Bayesiano se minimizan en 13 clusters. Tomaremos esta medida.



4.1.3 K-modes



5. Visualización

GMM (Gaussian Mixture Model) es un modelo probabilístico que asume que los datos provienen de una combinación de varias distribuciones gaussianas (normales). En este algoritmo, los clusters se generan mediante la identificación de diferentes distribuciones gaussianas que representan diferentes grupos de datos. Cada cluster tiene una distribución gaussiana, que se caracteriza por su media y varianza. Como resultado, los clusters de GMM pueden tener formas irregulares.

Por otro lado, K-means es un algoritmo de agrupamiento que utiliza la distancia euclidiana para definir la similitud entre los puntos de datos. Este algoritmo comienza por asignar los datos aleatoriamente a k clusters y luego ajusta los centros de los clusters para minimizar la distancia media entre los puntos y los centros de los clusters. Debido a la naturaleza de la distancia euclidiana, los clusters de K-means suelen tener similar tamaño.

Finalmente, K-modes es un algoritmo de agrupamiento que se utiliza para datos categóricos o nominales. En este algoritmo, los clusters se generan mediante la identificación de patrones de similitud entre los atributos categóricos de los datos. K-modes utiliza una medida de distancia basada en la similitud de los atributos, que difiere de la distancia euclidiana utilizada en K-means. Debido a la naturaleza de los datos categóricos, los clusters de K-modes suelen ser menos definidos.