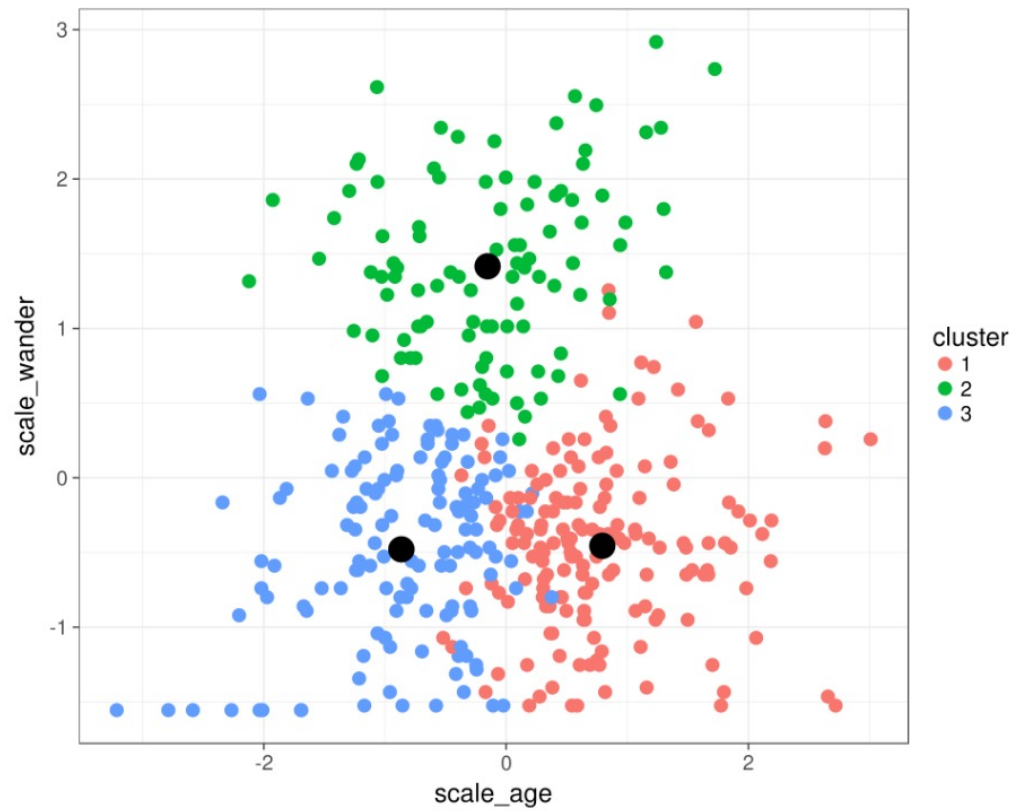
Machine Learning Models

Machine Learning Models

No Supervisado – K-Means

Aprendizaje No Supervisado: K-Means Clustering

Es un tipo de aprendizaje no supervisado, que se utiliza cuando tiene datos no etiquetados es decir, datos sin categorías o grupos definidos. El objetivo de este algoritmo es encontrar grupos en los datos. Los puntos de datos se agrupan según la similitud de características



FUNDAMENTOS BÁSICOS

1. Objetivo Principal:

- El objetivo principal del K-Means es agrupar los datos de manera que los elementos dentro de un mismo grupo sean más similares entre sí que con los elementos de otros grupos.

2. Número de Clústeres (K):

- Antes de aplicar el algoritmo, debes especificar el número de clústeres (K) que deseas formar. Este parámetro es crítico y afecta directamente la calidad de los resultados.

3. Proceso Iterativo:

- El algoritmo opera de manera iterativa. Inicia ubicando aleatoriamente K centroides (puntos iniciales representativos) en el espacio de características.

4. Asignación de Puntos:

- Cada punto de datos se asigna al clúster cuyo centroide está más cercano. La cercanía se mide típicamente utilizando la distancia euclidiana.

La fórmula de la distancia euclidiana entre dos puntos (x1, y1) y (x2,y2) en un espacio bidimensional es:

$$\text{Distancia Euclidiana} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

En un espacio de mayor dimensión, la fórmula generalizada para la distancia euclidiana entre dos puntos $P = (p_1, p_2, \ldots, p_n)$ y $Q = (q_1, q_2, \ldots, q_n)$ es:

$$\text{Distancia Euclidiana} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

5. Actualización de Centroides:
   - Los centroides se actualizan tomando el promedio de todos los puntos asignados a cada clúster. Este proceso se repite hasta que los centroides dejan de cambiar significativamente o se alcanza un número máximo de iteraciones.

6. Convergencia:
   - El algoritmo converge cuando los centroides se estabilizan y los puntos ya no cambian de clúster de manera significativa.

7. Resultados:
   - Una vez que el algoritmo converge, obtienes los clústeres finales y cada punto de datos pertenece a un clúster específico.

8. Sensibilidad a la Inicialización:
   - La calidad de los resultados puede depender de la inicialización de los centroides. Por lo tanto, a veces, se ejecuta el algoritmo varias veces con diferentes inicializaciones y se selecciona el mejor resultado.

9. Usos Comunes:
   - K-Means se utiliza en diversas aplicaciones, como segmentación de clientes, compresión de imágenes, análisis de texto, entre otros.

10. Limitaciones:
   - K-Means asume que los clústeres son convexos y de forma esférica, lo que puede limitar su rendimiento en conjuntos de datos con formas de clúster más complejas. Además, es sensible a los valores atípicos.

En resumen, K-Means es un algoritmo eficaz y rápido para la agrupación de datos, pero la elección adecuada del número de clústeres y la inicialización de los centroides son aspectos críticos para obtener resultados significativos.

Formas de calcular las distancias de los puntos a los centroides

Distancia Manhattan (o distancia de la ciudad): Esta distancia se calcula sumando las diferencias absolutas entre las coordenadas de los puntos en lugar de los cuadrados de las diferencias. En un espacio bidimensional, la fórmula es:

$$\text{Distancia Manhattan} = |x_2 - x_1| + |y_2 - y_1|$$

Distancia Minkowski: Es una generalización que incluye tanto la distancia euclidiana como la distancia Manhattan. La fórmula generalizada es:

$$\text{Distancia Minkowski} = \left(\sum_{i=1}^{n} |q_i - p_i|^p\right)^{1/p}$$

Donde p es un parámetro que controla la "norma" de la distancia (1 para Manhattan, 2 para Euclidiana).

Distancia de Chebyshev: Esta distancia se calcula tomando el máximo de las diferencias absolutas a lo largo de todas las dimensiones. En un espacio bidimensional:

$$\text{Distancia Chebyshev} = \max(|x_2 - x_1|, |y_2 - y_1|)$$

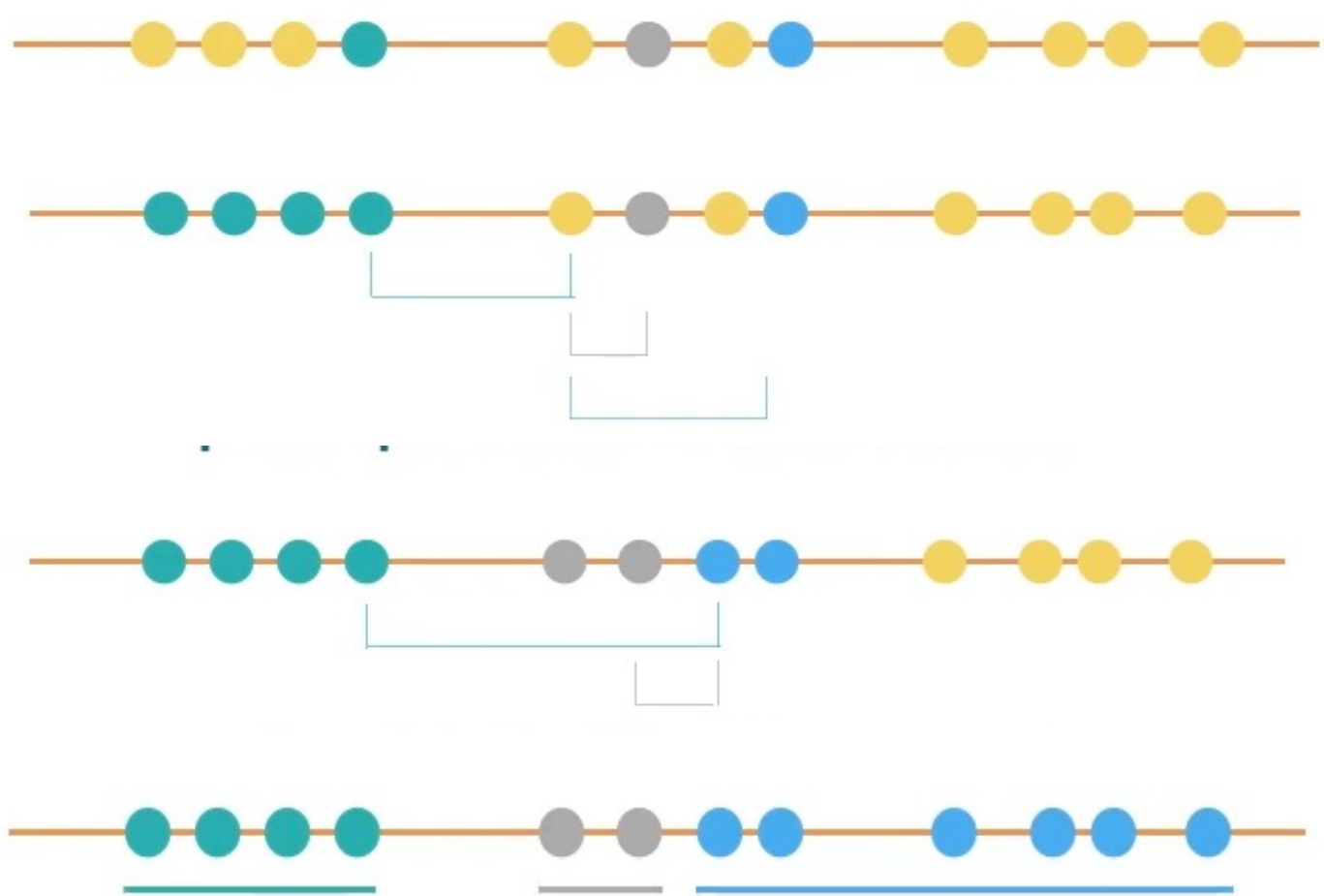
Aprendizaje No Supervisado: K-Means Clustering

Paso 1: Seleccionar el número de cluster (K) que deseas identificar en los datos. En este caso K = 3

Aprendizaje No Supervisado: K-Means Clustering

Paso 2: Seleccionar al azar K puntos. No necesariamente deben ser puntos de nuestros datos pueden ser puntos nuevos

Paso 3: Medimos la distancia entre cada uno de los datos y los puntos seleccionados, asignándole el punto que se encuentre más cerca



Paso 4: Colocamos nuevos K puntos y repetimos el procedimiento

Seleccionamos al azar nuevos centros de grupos

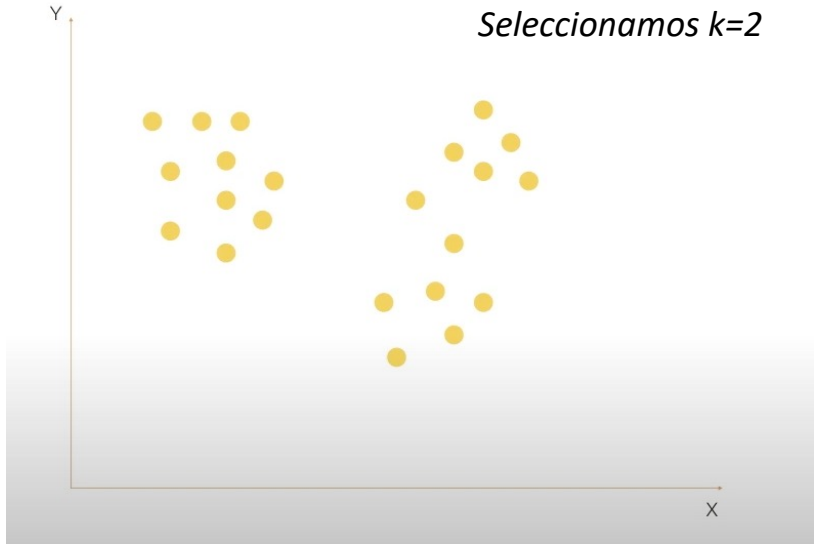
Obtenemos resultado distinto al anterior

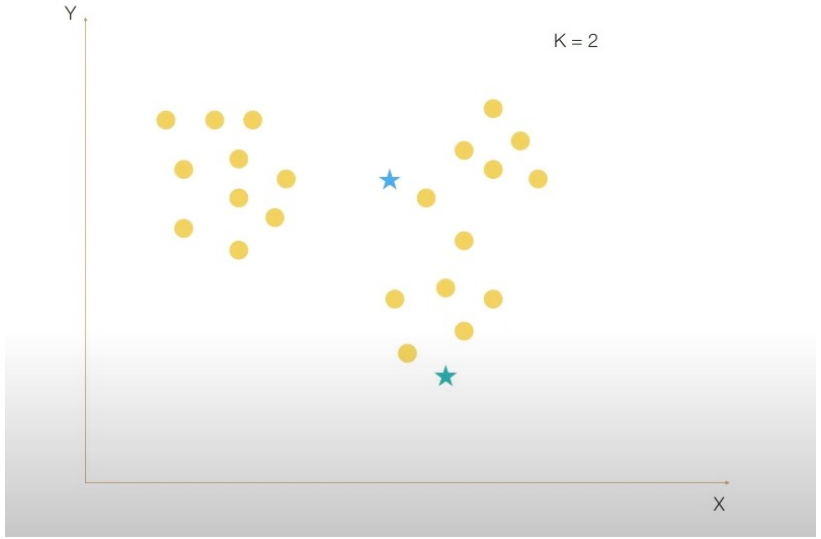
Repetimos el proceso y asignamos los puntos al centroide más cercano

El algoritmo repetirá el proceso hasta determinar que esta es la mejor distribución al obtenerse en dos iteraciones el mismo resultado

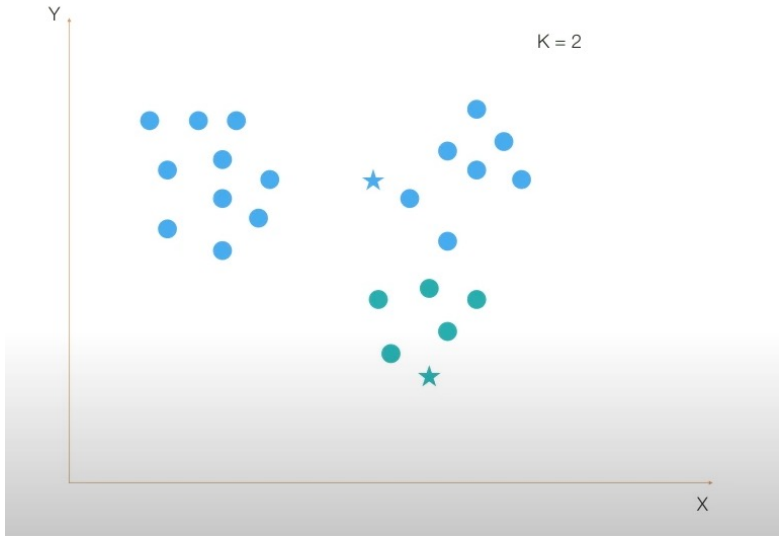
Datos en dos dimensiones

Seleccionamos k=2



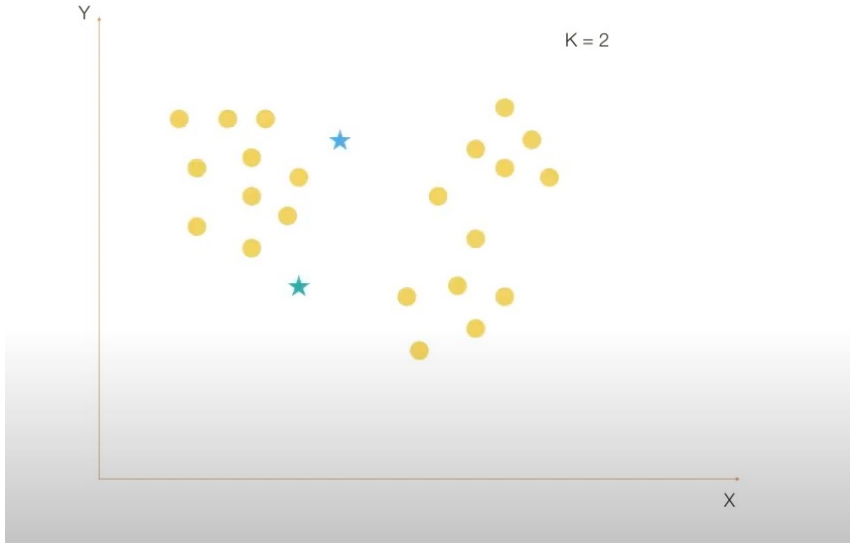


Ubicamos los centroides de forma aleatoria



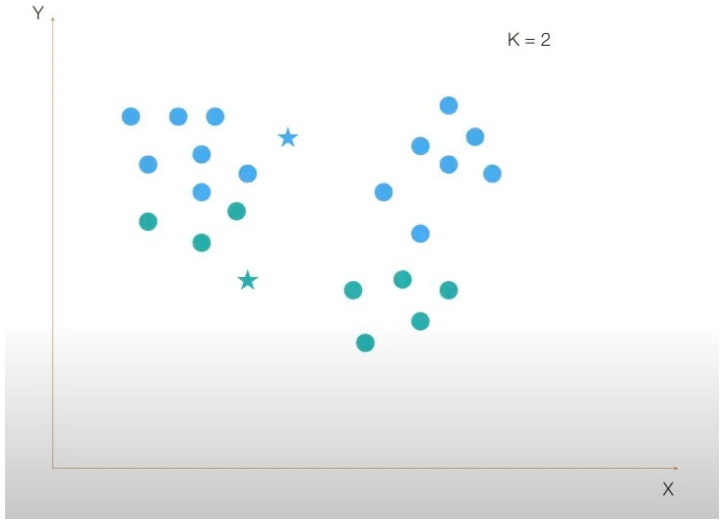
Medimos distancias

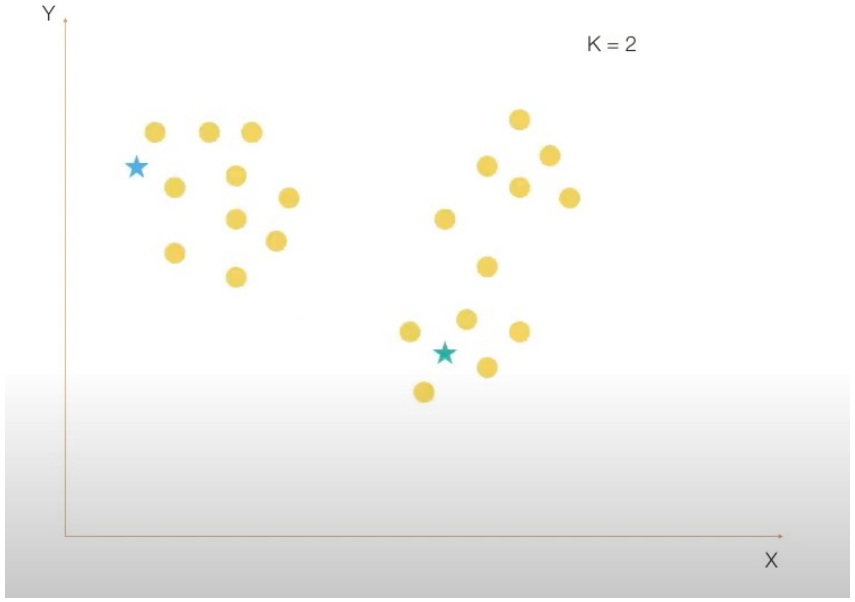
Resultado de primer análisis

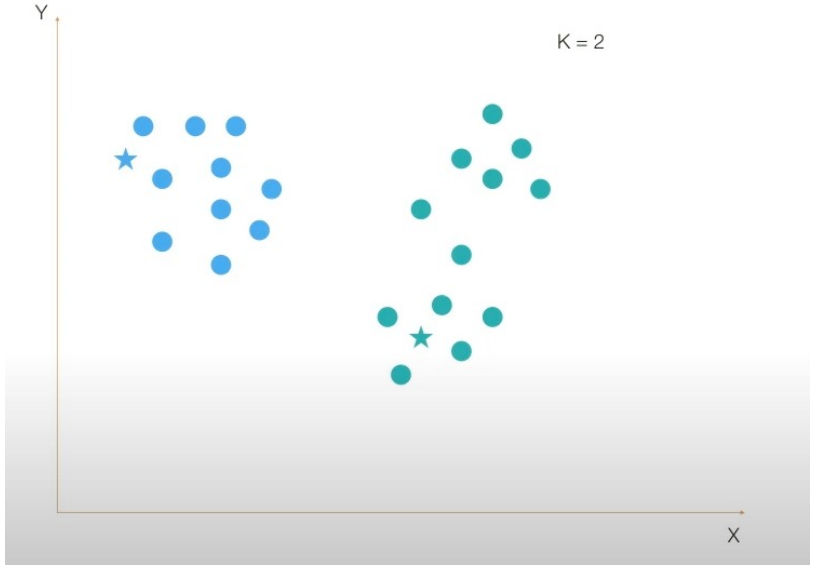


Volvemos a ubicar los centros aleatoriamente

Medimos distancias



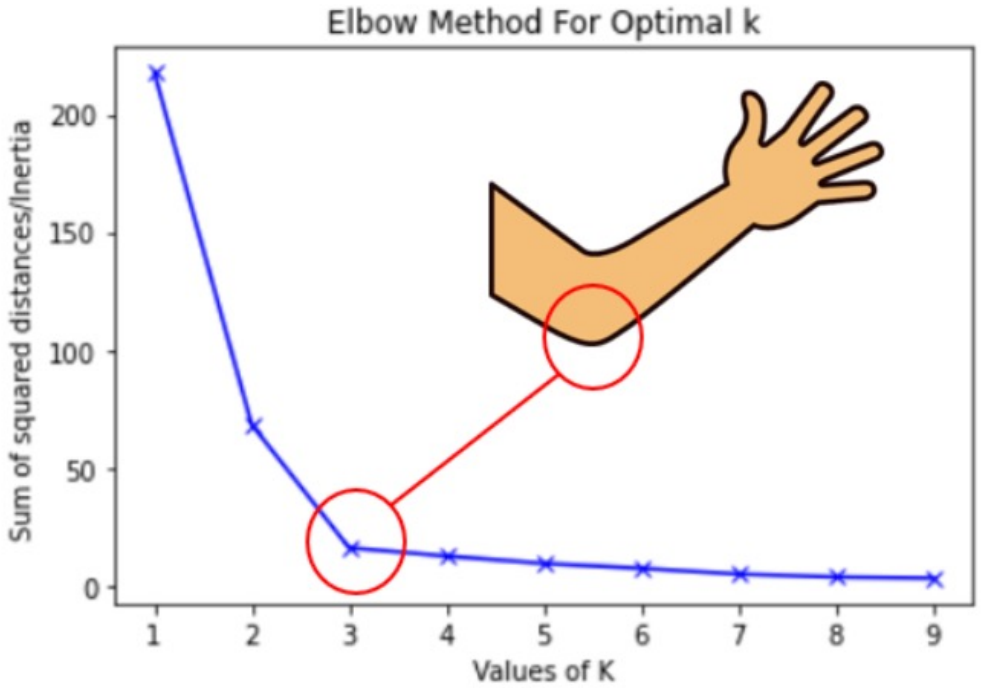




Repetimos el proceso hasta que se forman los mismos grupos en iteraciones consecutivas

- Método del “Codo” y el concepto de “Inercia”

- La "inercia" en el contexto del algoritmo K-Means se refiere a la suma de las distancias al cuadrado entre cada punto de datos y el centroide de su clúster asignado. En otras palabras, la inercia mide cuánto se extienden los puntos de datos dentro de su propio clúster.
- El objetivo del algoritmo K-Means es minimizar esta inercia. Cada vez que se ajusta el modelo con un número diferente de clusters (k), la inercia se calcula y se utiliza como métrica para evaluar el rendimiento del modelo.
- Un modelo K-Means con un menor valor de inercia se considera mejor porque significa que los puntos dentro de cada clúster están más cerca entre sí y, por lo tanto, el modelo es más coherente y representa mejor las estructuras en los datos.
- Cuando se realiza la visualización de codo para determinar el número óptimo de clusters, estás observando cómo la inercia cambia con diferentes valores de k. Buscar el "codo" implica identificar el punto donde la reducción en la inercia comienza a disminuir significativamente, lo que sugiere que agregar más clusters ya no proporcionará una mejora sustancial en la compactación de los datos. Este punto es crucial para elegir un número óptimo de clusters que equilibre la precisión del modelo con su complejidad.



Line plot between K and inertia

Aprendizaje No Supervisado: K-Means Clustering

1. Es un algoritmo rápido, robusto y simple que proporciona resultados confiables cuando los conjuntos de datos son distintos o bien separados entre sí de forma lineal
2. Debemos especificar el número de clústeres, de antemano, y los resultados finales son sensibles a la inicialización
3. La agrupación puede no funcionar bien si contiene datos muy superpuestos o si los datos son ruidosos o están llenos de valores atípicos