Variaciones del K-means aplicado al juego FC25

Carlos Manuel Orrego Franco

October 2024

1 Introducción

Proyecto de aprendizaje no supervisado que busca clasificar los jugadores del popular juego FC25 (Fifa) en 4 clusters, usando el algoritmo k-means con 4 variantes del mismo, aplicando 3 tipos de métricas diferentes para medir la similitud (Euclideana, Mahalanobis y L1). Se hace la visualización usando PCA.

2 Descripción del Conjunto de Datos

Descripción del conjunto de datos (2 renglones).

3 Exploración de Datos (EDA)

Se lleva a cabo un análisis exploratorio de los datos (EDA), seguido de una descripción del workflow:

- Número de variables seleccionadas y usadas en el análisis.
- Proceso de escalado de las variables.

Además, se realiza la imputación de los datos faltantes y se codifica la variable del pie preferido en una variable binaria.

4 Implementación del Algoritmo k-means

4.1 Implementación Usando Librerías

Se utiliza la implementación del algoritmo k-means proporcionada por librerías estándar.

4.2 Implementación Manual

Se implementa manualmente el algoritmo k-means, utilizando tres tipos diferentes de distancias: Euclideana, Mahalanobis y L1.

5 Descripción del Algoritmo k-means

El algoritmo k-means es un método de clasificación no supervisada, utilizado para particionar un conjunto de datos en k clusters. Algunas de sus características clave son:

- Es un ejercicio de clasificación no supervisado.
- No es adecuado para estructuras de datos no convexas.
- Es sensible a la inicialización de los centroides.
- Los datos deben ser escalados adecuadamente para su correcto funcionamiento.

6 Visualización y Reducción de Dimensionalidad

Para visualizar la clusterización, se utiliza PCA (Análisis de Componentes Principales) para reducir la dimensionalidad del conjunto de datos. Se eligen tres componentes principales que preservan el 80% de la varianza de los datos. Posteriormente, los centroides se transforman en el espacio de dichas componentes para su visualización en 2D y 3D.