Informe Implementacion K-means con distintas distancias

Jusn Jose Valencia - Maycol Anderson Ruales - Matias Silva

Octubre 2024

1 Descripción del Conjunto de Datos

El conjunto de datos utilizado en este análisis contiene atributos de jugadores de fútbol, incluyendo características físicas, habilidades técnicas y el pie preferido de cada jugador. En total, se trabajó con un conjunto de 10 variables que abarcan tanto aspectos físicos como técnicos de los jugadores, luego de realizar una selección basada en su relevancia para la clusterización.

2 Análisis Exploratorio de Datos

Antes de implementar los algoritmos de K-Means, se realizó un análisis exploratorio de datos para comprender mejor la estructura de los datos y su distribución. Primero se estudio el tamaño de los datos, luego se estudiaron las 10 mejores ligas en graficas. Luego se hizo una distribucion Overral para ver dato atipicos en los datos y hacer un histograma de ello. Como agregado se hizo la tecnica de funcion codo para determinar la cantidad de clusters para implementar K-means. Y finalmente se hizo un grafico boxplot para determinar los distintos atributos de cada jugador.

3 Workflow del Análisis

El proceso completo siguió el siguiente flujo de trabajo:

1. Selección de variables: Se escogieron 10 variables relevantes relacionadas con características físicas y técnicas de los jugadores.

2. Escalado de los datos: Se aplicó la estandarización de los datos, ya que K-Means es sensible a las diferencias de escala entre las variables.

3. Imputación de valores faltantes: Los valores nulos fueron imputados utilizando la mediana para variables numéricas y una codificación binaria para el "pie preferido".

4. Codificación binaria: La variable "pie preferido" se transformó en una variable binaria para su uso en el análisis.

4 Version 1: Implementación del K-Means Usando Librerías

Para la primera versión del algoritmo K-Means, se utilizó la librería `sklearn`. La implementación de K-Means con inicialización estándar se realizó utilizando las funciones predefinidas de la librería, que permite una fácil implementación y evaluación. Esta versión sirve como una línea base para comparar las otras implementaciones.

5 Version 2: Implementación con K-Means++

En esta segunda versión, se utilizó el algoritmo K-Means++ para mejorar la inicialización de los centroides. K-Means++ selecciona los centroides iniciales apartir de la mediana, evitando así el problema de la aleatoriedad en la inicialización que puede afectar la convergencia del algoritmo estándar.

6 Implementación Manual del Algoritmo K-Means con Distancias Euclidiana, Mahalanobis y Manhattan

Se implementaron manualmente para las tres versiones del algoritmo K-Means, cada una utilizando una métrica de distancia diferente:

1. **Distancia Euclidiana:**

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

2. **Distancia de Mahalanobis:**

$$d(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$

3. **Distancia de Manhattan:**

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

7 Descripción del Algoritmo K-Means

K-Means es un método de clasificación no supervisada que agrupa datos en k clusters basándose en la cercanía de los puntos a los centroides. Asi como fue implementado en las versiones con el siguiente proceso:

1. Inicializar k centroides.
2. Asignar puntos a los centroides más cercanos.

3. Recalcular los centroides como el promedio de los puntos asignados.
4. Repetir hasta la convergencia.

Es crucial escalar los datos para evitar que diferencias en las escalas afecten el resultado.

8 Reducción de Dimensionalidad con PCA y Visualización

Se utilizó Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos a 2 y 3 componentes principales para asi facilitar la visualización de los clusters en 2D y 3D. Al seleccionar tres componentes principales obtenemos aproximadamente el 80% de la varianza total de los datos, para asi evitar cesgos. De esta manera los centroides obtenidos fueron transformados a este espacio reducido, lo que permite visualizar la clusterización en gráficos bidimensionales y tridimensionales.