# Informe - Ejercicio: Implementación del Algoritmo K-means

Andres Sebastian Guevara Ortiz - Thomas Jaramillo Aguirre

Octubre 2024

## 1 Introducción

El proyecto se basa en un aprendizaje no supervisado de la base de datos del juego FC25 (FIFA), para poder clasificar a sus jugadores en 4 clusters. Se utiliza el algoritmo K-means con 4 variaciones del mismo aplicando 3 tipos diferentes de distancias: Euclidiana, Mahalanobis y L1. Además, se logra hacer una visualización mediante un Análisis de Componentes Principales (PCA).

## 2 Preparación de Datos

La base de datos pasó por un Análisis Exploratorio de Datos (EDA) para un buen procesamiento, asegurando que las variables seleccionadas sean de provecho para el algoritmo. En este caso, se utilizarán 45 variables para cada jugador, con una escala que irá de 0 a 100 para que cada una tenga un peso similar en el procesamiento.

### 2.1 Transformación de Variables

Algunas variables clasificadas como string, como el pie preferido, han sido transformadas en variables binarias. Además, en las categorías que contenían datos NaN (datos nulos), se han imputado valores en 0, como en la categoría de arqueros donde estos no afectan la estadística.

## 3 Implementación del Algoritmo

Para la primera implementación del algoritmo se ha utilizado la librería de `sklearn`, la cual ya incluye K-means para su uso rápido, además del K-means++ que es una forma mejorada de este algoritmo. Se ha utilizado la técnica del codo para determinar la cantidad ideal de clusters para una correcta clasificación de los datos.

### 3.1 K-means Manual

En el K-means desarrollado a mano, se han seguido los pasos del algoritmo que constan de:

1. Asignar la cantidad de clusters necesarios (utilizando el método del codo).
2. Ubicar una cantidad de centroides iguales a la cantidad de clusters.
3. Asignar cada dato al clúster al que se adecúa.
4. Recalcular el centroide para seguir buscando datos.
5. Generar un parámetro de finalización que será la tolerancia, que indica la diferencia de posición entre el clúster inicializado y a la distancia que se encuentra de su antecesor.

se debe tener en cuenta que este algoritmo es sensible a la inicializacion como tambien, no sirver para estructuras convexas y todos los datos deben de estar escalados.

## 4 Análisis de Resultados

Después de la función general, se realizarán varianzas entre las 3 distancias para encontrar nuevas distribuciones de la clusterización. Todo esto pasará por un PCA para retener los componentes principales, que proporcionarán una varianza de solo un 80% de cada dato. Para su próxima visualización, se presentarán tablas en 2D y 3D, incluyendo los centroides en términos del PCA.

## 5 Visualización de Resultados

En la Figura 1 se presenta una visualización de los clusters obtenidos:

## 6 Visualizacion del Algorimo K-means en distancia euclidiana

En la Figura 2 se muestra la visualización de los clusters y centroides utilizando la distancia euclidiana:

## 7 Visualización 3D de Clusters y Centroides

En la Figura 3 se muestra la visualización 3D de los clusters y centroides utilizando la distancia euclidiana:

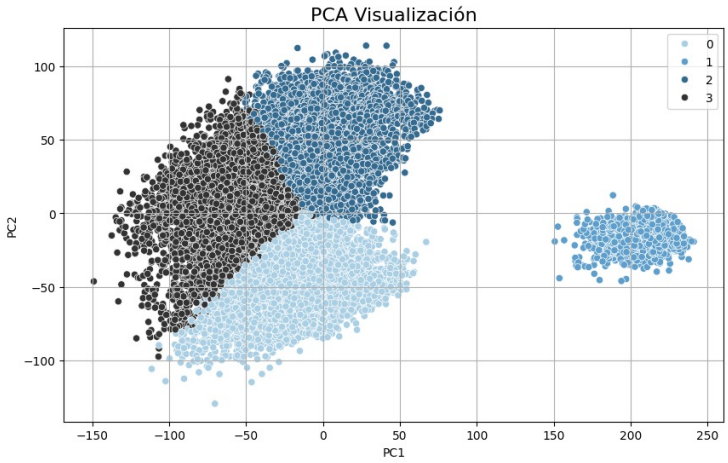


Figure 1: Clusters de jugadores del juego FC25.

# 8 Visualización de K-means para la distancia mahalanobis

En la Figura 4 se muestra la visualización del algoritmo K-means utilizando la distancia Mahalanobis:

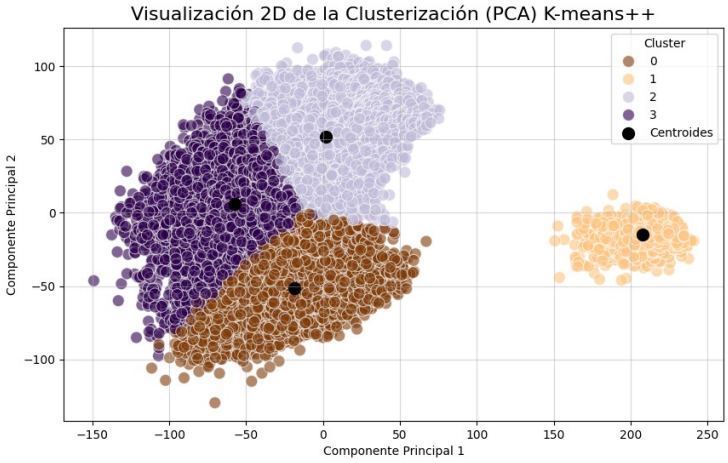


Figure 2: Visualización de los clusters y centroides utilizando distancia euclidiana.

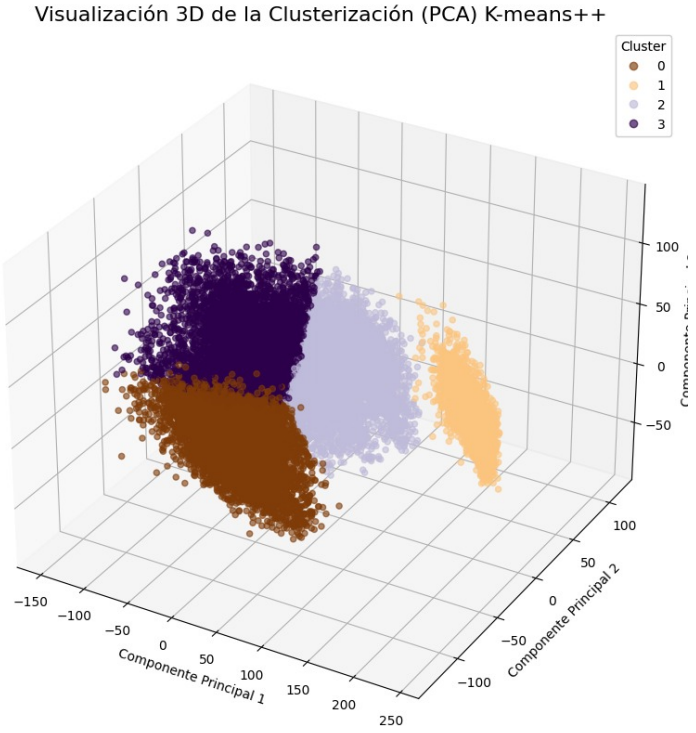


Figure 3: Visualización 3D de los clusters y centroides utilizando distancia euclidiana.

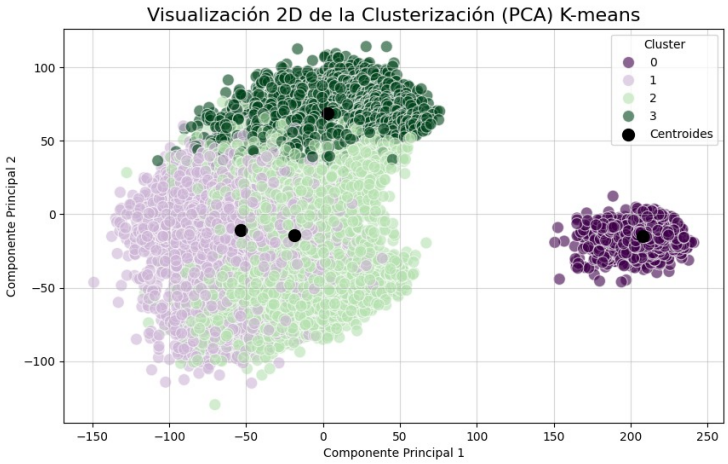


Figure 4: Visualización del algoritmo K-means con distribución de distancia Mahalanobis.