Informe Trabajo final Analítico de datos

Juan Andrés García Jiménez - Jaime Andrés Ocampo - Jacobo Giraldo

Octubre 2024

Partimos de un exhaustivo análisis exploratorio de datos (EDA) para examinar y comprender en detalle las características y la distribución del conjunto de datos, basándonos en la base de jugadores del videojuego EA FC 25.

- El flujo de trabajo comienza importando las librerías necesarias y cargando el archivo que contiene las estadísticas de los jugadores. Este archivo será la base para calcular diferentes estadísticas y evaluar el rendimiento de cada jugador. En este proceso, se implementan técnicas de análisis y visualización de datos que nos permiten profundizar en el comportamiento de los jugadores, con el objetivo de representar estos hallazgos gráficamente. A continuación, repetimos este proceso, pero introducimos diferentes métricas de distancia, como la euclidiana, la de Mahalanobis y la de Manhattan, para estudiar cómo afectan los resultados.

- Además, se preparó el archivo de datos para su análisis eliminando la información irrelevante, ajustándolo para el procedimiento correcto, y luego, mediante una URL, se cargaron los datos listos para su procesamiento.

- En cuanto a la imputación de datos faltantes, se emplearon técnicas adecuadas para completar los valores ausentes, asegurando así que los datos estén completos y evitando posibles errores. También se transformó la variable "pie preferido" de los jugadores en un formato binario para simplificar su tratamiento en los modelos posteriores.

- Para la implementación del algoritmo K-means, primero se aplicó utilizando librerías predefinidas como sklearn, lo que optimizó el proceso y aseguró una correcta agrupación de los datos. Posteriormente, se realizó una implementación manual del algoritmo, explorando tres tipos de distancias: la euclidiana, que mide la distancia directa entre puntos; la Manhattan, que sigue una estructura de cuadrícula; y la Mahalanobis, que considera la correlación entre variables. Esto nos permitió evaluar cómo cada métrica impacta en la formación de los clusters.

- El K-means es un algoritmo diseñado para detectar patrones ocultos en los datos sin etiquetas. Sin embargo, presenta algunas limitaciones, por ejemplo no es eficaz para datos con estructuras no convexas y es sensible

a la inicialización de los centroides, lo que puede alterar los resultados. Además, es crucial que los datos estén correctamente escalados, ya que el algoritmo depende de las distancias para funcionar adecuadamente.

- Finalmente, para mejorar la visualización de los clusters, se empleó el análisis de componentes principales (PCA). Este método reduce la dimensionalidad del conjunto de datos, permitiendo proyectar los resultados en dos y tres dimensiones. Se seleccionaron tres componentes principales que preservan el 80% de la varianza, y los centroides fueron transformados a este nuevo espacio para facilitar una representación visual clara y comprensible de las agrupaciones formadas.