Proyecto de programación para análisis de datos.

Definición del Problema

En el mundo del deporte es trascendental poder saber lo que va a ocurrir en el futuro. Lo que yo busco lograr en este proyecto es poder predecir, tomando en cuenta las últimas tres temporadas (2020-2021, 2021-2022 y 2022-2023), que jugadores van a mejorar, empeorar o mantener su rendimiento. Esta información es vital para los directivos deportivos y entrenadores, ya que sabiendo lo que va a pasar en el futuro, pueden tomar mejores decisiones para el futuro. Recolección y preparación de datos

Recolecté las bases de datos para 2020-2021, 2021-2022 y 2022-2023 desde Kaggle, importándolas a mi computadora. Hubo varias cosas que tuve que hacer durante el proceso de limpieza y preparación de los datos, entre las cosas que se hicieron fue: - Eliminar los jugadores duplicados y en la posición "TOT" poner el último equipo en el que jugó el jugador - Eliminar los jugadores que no aparecen en las tres temporadas - Poner el nombre de los jugadores como el índice - Convertir las variables categóricas a numéricas - Cambiarle el nombre a las columnas para poder identificar de qué año son - Normalizar los valores de los Datasets - Discretizar algunas estadísticas

Análisis exploratorio de datos

Hice varias comparaciones de estadísticas para ver si existía cierta relación entre dos estadísticas. Algunos insights interesantes que encontré fueron los siguientes:

- Existe un relación directa entre los puntos anotados y las pérdidas generadas. Entre mas puntos genera un jugador, mas perdidas genera.
- Existe una relación entre los puntos anotados y la cantidad de tiros libres intentados. Entre más puntos mete un jugador, más tiros libres intenta.
- Otra relación interesante es que los rebotes de un jugador tiene relación con los tiros libres intentados.

Al terminar el uso del algoritmo KMeans, utilicé gráficas de barras para poder entender de una manera más sencilla cual es la tendencia de los jugadores dentro de ese cluster.

Metodología

Las librerías utilizadas fueron: - Pandas: para el manejo con los dataframes - Numpy: Para realizar operaciones con np.arrays - Matplotlib.pyplot: Para realizar la visualización de los datos
- Sklearn.cluster.KMeans: Para realizar la agrupación de los jugadores con tendencias similares

Durante el desarrollo del proyecto saltaron ciertas cosas que no tenía contempladas: - La manera en la que se iba a evaluar la efectividad del modelo. - La

manera de convertir a numérica la variable “Tm” - El descubrimiento de que el dataset de la temporada 2020-2021, no estaba completo y solo contenía los datos de una porción de la temporada.

Análisis Predictivo

Tomé la decisión de usar el algoritmo de KMeans, este es un algoritmo de clustering, no supervisado. Este se acomoda de gran manera a mi proyecto, ya que yo no tengo una “Y” que me diga los resultados, solo tengo columnas “X” mas no columnas “Y”.

Para la preparación de los datos, se realizaron las operaciones especificadas en el apartado de recolección y preparación de datos. Además de eso, se concatenaron los tres dataframes, para que fuera uno solo tomando como punto de referencia el nombre del jugador.

Los datos se entrenaron y se predijeron en una sola línea con el operador de. fit_predict() en base al Dataset ya con los datos y después se creó una nueva columna “Cluster” a partir del resultado del .fit_predict().

El modelo se evaluó tomando las estadísticas de la temporada 2023-2024 y comparando el desempeño de los jugadores con las predicciones del KMEANS. Se realizó el mismo proceso de limpieza de datos y feature engineering con esta base de datos que que lo que se hizo con las bases de datos iniciales.