**Análisis de Rendimiento Futbolístico.**

**Un Enfoque Basado en Inteligencia Artificial.**



**HUGO FRAILE MARTÍNEZ**

**Índice**

1. **Objetivos y consideraciones previas.**
   * Se explicará el objetivo del trabajo a realizar y la forma de proceder.
2. **Análisis del Dataset original.**
   * Se realizará un análisis exhaustivo del Dataset, para valorar que información es relevante para el experimento.
3. **Preprocesamiento del Dataset.**
   * Se realiza la limpieza y preparación de un Dataset de fútbol, eliminando datos innecesarios y gestionando valores faltantes para garantizar su uso en análisis posteriores.
4. **Análisis Estadístico de Jugadores.**
   * Se exploran las métricas clave de rendimiento de los jugadores, observando estadísticas como goles, asistencias y minutos jugados, proporcionando un panorama general de los datos.
5. **Comparación de Jugadores.**
   * Se comparan jugadores seleccionados utilizando estadísticas avanzadas, destacando similitudes y diferencias en su desempeño mediante visualizaciones y tablas comparativas.
6. **Búsqueda de Jugadores Similares.**
   * Se aplica un algoritmo de Clustering para agrupar a los jugadores según su rendimiento en varias categorías, identificando grupos de jugadores con perfiles similares.
7. **Scouting de Talento.**
   * Se crearán y valorarán distintas aproximaciones de modelo, a la hora de crear un modelo que, en función de las características de un nuevo jugador, sea capaz de predecir otra característica futura del mismo.

**1 Objetivos y consideraciones previas**

-El objetivo de este trabajo es desarrollar un modelo predictivo basado en inteligencia artificial que permita evaluar y predecir el rendimiento de los jugadores de fútbol en tiempo real y a lo largo de una temporada. Optimizar la toma de decisiones de cara a fichar a un jugador para reemplazar a otro que se marcha del equipo y poder predecir el rendimiento en función de los datos medios por partido de un nuevo jugador que emerge en el panorama futbolístico. Mediante el análisis de datos detallados de jugadores y equipos. Identificaremos patrones ocultos en el rendimiento detectables con métodos convencionales y crearemos modelos que sean capaces de predecir de forma exacta indicadores de rendimiento como pueden ser el número de goles.

- Es preciso considerar que nuestro trabajo estará muy limitado a los datos que tenemos, sin embargo, este está hecho para ser escalable, es decir con unos datos de mayor calidad se podrían realizar predicciones exactas, a su vez estas predicciones podrían usarse de apoyo a la hora de tomar decisiones profesionales.

**2 Análisis del Dataset original.**

El Dataset que utilizaremos en este análisis está disponible en el siguiente enlace: [Kaggle Football Player Stats 2022-2023](https://www.kaggle.com/datasets/vivovinco/20222023-football-player-stats). Este contiene más de 2500 filas y 124 columnas, ofreciendo una gran cantidad de información detallada sobre el rendimiento de los jugadores de las 5 grandes ligas europeas de fútbol a lo largo de la primera mitad de la temporada 2022-2023. Incluye variables clave como el nombre del jugador, la nación, la posición en el campo, el equipo en el que juega, la liga a la que pertenece, y estadísticas detalladas como los goles, minutos jugados, asistencias, tiros, pases, regates y acciones defensivas. Este conjunto de datos es extremadamente útil para realizar análisis profundos sobre el rendimiento de los jugadores, identificando patrones y tendencias que pueden ayudar a mejorar la toma de decisiones tácticas y estratégicas.

Con un total de 124 columnas, este Dataset ofrece una perspectiva integral del rendimiento futbolístico, abordando desde estadísticas ofensivas (goles, asistencias, tiros a puerta) hasta métricas defensivas (entradas, intercepciones, bloqueos). También incluye información avanzada sobre los pases, como la distancia total y progresiva de los pases, pases cortos, medianos y largos completados, y pases que terminan en la zona del tercio ofensivo o en el área penal. Además, se analizan aspectos de la participación en jugadas clave como acciones que crean tiros y goles, y estadísticas relacionadas con los regates, los balones recuperados y las acciones aéreas ganadas.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

**3 Preprocesamiento del Dataset.**

-El preprocesamiento del Dataset es una fase crucial y necesaria a la hora de encarar el proyecto, porque la calidad y utilidad de nuestros datos va a depender de esta. El proceso comenzó con la eliminación de aquellas columnas que consideramos innecesarias para el análisis, en este caso hemos eliminado las columnas [‘Rk’ y ‘Born’], una por ser irrelevante y otra por tener bastante similitud con la columna ‘Age’.

-Después hicimos una breve vista a los valores faltantes dentro de las propias filas de Dataset, este paso fue especialmente delicado, ya que, entendiendo la propia naturaleza de los datos, algunos jugadores que tengan posiciones en el campo más ofensivas, van a tender a tener valores nulos o casi nulos en determinados parámetros y no por ello tenemos que descartarlos. Finalmente acabamos descartando una única fila por una duplicidad.

-Tras ello con el propósito de simplificar y mejorar los resultados del proyecto, decidimos agrupar a los jugadores en 4 subgrupos, en función de su posición en el campo. Originalmente los jugadores estaban agrupados en 10 posiciones: [‘DF’, ‘MF’, ‘FW’,’GK’, ‘FWMF’, ‘MFFW’, ‘MFDF’, ‘DFMF, ‘DFFW, ‘FWDF’], sin embargo, vamos a agruparlo en 4 subgrupos: [‘GK’, ‘DF’, ‘MF’, ‘FW’]. Tras crear la función pertinente que se encargue de agrupar a los jugadores, obtenemos el siguiente histograma que mide la cantidad de jugadores por grupo.

Gráfico, Gráfico de barras

Descripción generada automáticamente

-Finalmente, guardaremos el Dataset, para poder utilizarlo en nuestro proyecto.

**4 Análisis Estadístico de Jugadores.**

-La primera parte de nuestro proyecto va a consistir en hacer un análisis detallado del rendimiento de cada jugador. Para ello hemos creado una función **individual\_chart**, que no sirve para visualizar el rendimiento individual de cada jugador. La función muestra en que percentil se encuentra ese jugador en una estadística determinada respecto al resto de jugadores de su misma posición de las 5 grandes ligas europeas. Para visualizar la información hemos decidido utilizar un pizza chart, que ilustra perfectamente esos percentiles y genera un gráfico muy visual y fácil de entender para cualquier persona posea está o no conocimientos técnicos.

-A continuación, podemos ver dos ejemplos de uso, en dos jugadores de nuestra liga, **Karim Benzema** y **Koke**, visualizando los datos de ambos futbolistas de esta manera, y con cierto conocimiento técnico de la materia, podríamos catalogarlos como un perfil de futbolista determinado, ya que el gráfico nos ha permitido saber en qué características son buenos estos, y como de buenos son respecto al resto de sus compañeros.

Gráfico, Gráfico de proyección solar

Descripción generada automáticamente

Gráfico, Gráfico de proyección solar

Descripción generada automáticamente

**5 Comparación de Jugadores.**

**-**Ahora de cara a realizar una comparación más específica hemos creado otra herramienta para nuestro proyecto, y es que a veces de cara a encontrar el recambio óptimo para un jugador que se nos marcha, queremos realizar una comparativa más específica entre nuestra nueva incorporación y el jugador que deja el equipo. La función **compare\_players,** se encarga de generar gráficos visualmente atractivos, que nos permiten realizar una comparación directa entre ambos jugadores, mostrando sus fortalezas y debilidades respectivamente en cada categoría. Con el fin de hacer lo más visual posible esta comparación, hemos un polígono de radar por cada jugador, con el fin de que la diferencia visual entre sus respectivas áreas nos muestre en que características son fuertes respecto al otro. Además, en este gráfico, las estadísticas de tipo continuo se utiliza el percentil 95 como máximo valor, hemos seleccionado este, ya que al probar a seleccionar un percentil más alto daba pie a que el gráfico tuviese un escalado mucho más inusual, por otra parte, los radios de las estadísticas enteras toman el valor máximo del dataset. Este gráfico está diseñado para ser adaptable a cualquier conjunto de datos y métricas dependiendo de las necesidades del análisis, además de permitirnos obtener insights, sobre los aspectos fuertes del jugador.

Gráfico, Gráfico radial

Descripción generada automáticamente

Gráfico, Gráfico radial

Descripción generada automáticamente

**6 Búsqueda de Jugadores Similares**

-La tercera parte de nuestro proyecto consiste en un modelo de Machine Learning que se encargará de realizar un análisis de los jugadores que tienen un mayor porcentaje de similitud a otro dado, para ello hemos creado un modelo de Clustering basado en las características de cada jugador. Tras la previa división en 4 grupos de los jugadores, ahora hemos seleccionado las características más relevantes de cada posición, para ello me he basado en mi propio juicio y conocimiento de la materia. Nuestro modelo previamente normalizara los datos, ya que sin esta normalización algunas características que no están expresadas en promedio por 90 minutos (como los goles y los pases), no se podrían comparar de manera equitativa. Luego se implementa un algoritmo de distancia euclídea, que mide la similitud entre un jugador especifico y el resto de los jugadores de su misma posición, esto nos permite detectar jugadores con habilidades y roles parecidos. Nuestra función **scout\_similar\_players**, realiza un sistema de scouting el cual permitiría a un director deportivo encontrar los jugadores más parecidos a otros de cara a encontrar un potencial reemplazo para una marcha del club, finalmente la función devuelve un gráfico que nos permite observar de manera más visual las similitudes entre jugadores.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Gráfico

Descripción generada automáticamente

**7 Scouting de Talento.**

**-**La última parte de nuestro proyecto, consistirá en la creación de un sistema de scouting de talento mediante a modelos de inteligencia artificial. Nuestro modelo se encargará de predecir una característica determinada en función de unos datos de entrada que le proporcionaremos. Con los datos proporcionados, la función **model\_player\_scouting\_train**, se encargará de entrenar nuestro modelo, utilizando únicamente esas columnas del Dataframe original. Tras este entrenamiento, la función **load\_model\_and\_test**, se encargará de utilizar los datos propios de un nuevo jugador, que no está en el Dataframe, Finalmente el modelo hará una predicción de la característica que deseábamos predecir.

-Un ejemplo de aplicación real, sería por ejemplo un club de fútbol que detecta una joven promesa, que tiene unas determinadas estadísticas media por partido, del cual no se tienen sus registros completos de la temporada y se quiere saber cuántos goles aspira a meter en una cantidad de minutos determinada, para ello nuestro modelo puede dar con un alto porcentaje de fiabilidad los goles aproximados que podría meter este nuevo jugador.

-Para poner a prueba nuestro modelo, le hemos metido datos de jugadores reales (jugadores cuyos datos figuran en el Dataset) y hemos obtenido los siguientes resultados:

Miguel Almirón (Objetivo :10 goles):

-Hicimos una primera aproximación dando una cantidad mínima de datos y obtuvimos lo siguiente:

Error cuadrático medio: 0.14783659943811792

Modelo entrenado y guardado correctamente

Modelo eliminado

Predicción de goles para el jugador ficticio utilizando nuestro modelo: 8.711994938463992

-Después tras introducir más datos, obtuvimos, mejores resultados, ya que nuestro modelo tenía más datos con los que aprender.

Error cuadrático medio: 0.2160854809991706

Modelo entrenado y guardado correctamente

Modelo eliminado

Predicción de goles para el jugador ficticio con Regresión Lineal: 9.876800542519687

-Folarín Balogun (14 goles):

Error cuadrático medio: 0.2160854809991706

Modelo entrenado y guardado correctamente

Modelo eliminado

Predicción de goles para el jugador ficticio con Regresión Lineal: 14.179066465107738

-Erling Haaland (25 goles):

Error cuadrático medio: 0.2160854809991706

Modelo entrenado y guardado correctamente

Modelo eliminado

Predicción de goles para el jugador ficticio con Regresión Lineal: 22.748804138321212

-Finalmente hicimos pruebas para calcular el número de goles que marcarían ciertos jugadores ficticios, con datos generados aleatoriamente, pero manteniendo cierto consenso y lógica.

Error cuadrático medio: 0.2160854809991706

Modelo entrenado y guardado correctamente

Modelo eliminado

Predicción de goles para el jugador Carlinhos Brown con Regresión Lineal: 15.434898306919452

Error cuadrático medio: 0.2160854809991706

Modelo entrenado y guardado correctamente

Modelo eliminado

Predicción de goles para el jugador Francois Gallardo con Regresión Lineal: 3.095542838169453

-El primer modelo usaba Regresión Lineal para predecir la característica deseada, dándonos resultados bastante buenos y rápidos, sin embargo, me parece interesante verificar, si haciendo un modelo con redes neuronales, podemos obtener mejores resultados manteniendo un tiempo de computación moderados, es por ello por lo que creamos una nueva aproximación a nuestro proyecto.

-La función **model\_player\_scouting\_train\_nn**, entrena el modelo con datos filtrados por posición y ajustados por minutos jugados, utilizando una red con varias capas densas. Finalmente, tras guardar el modelo entrenado, la función **load\_model\_and\_test\_nn**, nos permite cargar el modelo previo y predecir la característica deseada para el nyevo jugador. Para verificar la capacidad del modelo, vamos a probar nuevamente a predecir los datos del jugador Erling Haaland, ya que como habíamos visto previamente, el modelo anterior no era del todo exacto, al tratarse de un jugador que en el concepto futbolístico de goleador se sale de cualquier parámetro.

Predicción de goles para el jugador ficticio con Red Neuronal: [25.148357]

-Como podemos ver, tras unos segundos más de tiempo obtenemos una predicción mucho más acercada a la realidad del jugador, sin embargo, creo que el modelo de Regresión Lineal es mucho óptimo para este trabajo ya que en término de equilibrio entre la calidad de las predicciones y el coste computacional de estas, los resultados no son tan distantes a la aproximación por Redes Neuronales.