**Informe sobre análisis de estadísticas de futbolistas**

1. **Descripción del Dataset**

El Dataset utilizado corresponde a las estadísticas de jugadores de fútbol de la temporada 2024-2025 en las cinco grandes ligas europeas.

Este conjunto de datos fue obtenido de FBref y contiene un total de 2752 registros con 267 variables.

* 1. **Columnas relevantes**

Las columnas relevantes para este Dataset están clasificadas de acuerdo a las siguientes categorías:

* Información básica del jugador.
* Tiempo de juego y apariciones.
* Estadísticas de ataque.
* Estadísticas de defensa.
* Estadísticas de pases y creatividad.
* Estadísticas de porteros.
* Posesión y control de pelota.
* Estadísticas diversas.

1. **Metodología KDD**

Para la elaboración del proyecto se usó la metodología Knowledge Discovery in Databases (KDD), o Descubrimiento de Conocimiento en Bases de Datos, para analizar el rendimiento de futbolistas en función de diversas métricas. Las fases KDD que se han abordado incluyen la selección de datos, preprocesamiento, transformación, minería de datos, evaluación de modelos e interpretación de los resultados.

* 1. **Fase 1: Selección de los Datos**

El dataset contiene varias columnas que proporcionan información básica sobre cada jugador, como su nombre, nacionalidad, posición, edad, equipo, entre otros. Además, incluye estadísticas avanzadas relacionadas con el rendimiento de los jugadores, como goles, asistencias, tiros a puerta, entre otros.

También, en esta fase, se agregaron todas las librerías necesarias para la elaboración de los modelos posteriores. Se cargó se visualizó la estructura del Dataset para luego comenzar con la fase de preprocesamiento de los datos.

* 1. **Fase 2: Preprocesamiento de Datos**

En esta etapa se realizaron transformaciones para preparar los datos para el análisis. Se filtraron las columnas útiles y se gestionaron los valores nulos utilizando técnicas de imputación (media). Además, se normalizaron las características numéricas utilizando el estándar de escalado (StandardScaler).

Con el fin de eliminar variables irrelevantes o poco informativas, se eligieron las columnas que contenían al menos el 5% de valores no nulos, asegurando la calidad de las métricas para cada posición.

El motivo de este filtro por determinado porcentaje de valores no nulos se debe a que en Datasets deportivos como este (con estadísticas de jugadores de fútbol), es normal que muchas columnas tengan valores nulos porque no aplican a todos los jugadores. Cada jugador desempeña una posición distinta en el campo (portero, defensa, mediocampista, delantero) y las métricas que se recolectan varían mucho según esa posición.

* 1. **Fase 3: Transformación**

En esta fase se buscó convertir y adaptar los datos a un formato adecuado para la minería de datos. Las tareas u objetivos planteados en esta fase fueron:

* Dividir el Dataset en subconjuntos por rol (porteros, defensas, mediocampistas y delanteros).
* Reducir columnas irrelevantes para cada subconjunto.
* Preparar cada grupo para un análisis más específico y efectivo.

La decisión de establecer estos objetivos nace principalmente de lo mencionado en la fase 2, las columnas irrelevantes según determinada posición de un jugador, para ello se decidió crear eso cuatro subconjuntos y, a cada uno, establecerles columnas relevantes o útiles según la cantidad de valores no nulos que tengan. Esto ayudó a preparar los datos para la siguiente fase.

* 1. **Fase 4: Minería de datos**

El objetivo principal de esta fase es aplicar algoritmos de minería de datos para encontrar patrones y relaciones ocultas. Se aplicaron modelos de aprendizaje supervisado y no supervisado para extraer patrones de los datos.

Esto permitió observar patrones en el rendimiento de los jugadores según su posición. Además, se entrenaron modelos predictivos, como regresión lineal, Random Forest y redes neuronales (MLP), para predecir métricas como goles y asistencias. Los modelos usados fueron los siguientes:

* **K-Means (Clustering):** Se utilizaron los clústeres generados para identificar subgrupos de jugadores con características similares, lo que permitió una segmentación más precisa de los futbolistas en función de sus estadísticas de rendimiento.
* **Regresión Lineal:** Se aplicó un modelo de regresión lineal para predecir métricas como los goles de los futbolistas en función de sus características. Los resultados fueron muy precisos, con un R² cercano a 1.0, lo que indica una predicción precisa de los goles de los jugadores.
* **Redes Neuronales (MLP):** Se entrenó un modelo de red neuronal (MLP) para la misma tarea de predicción de goles. Aunque los resultados fueron buenos, hubo algunas diferencias con respecto a la regresión lineal, especialmente en el caso de las métricas de error, como el MSE y MAE.
* **Random Forest:** También se implementó un modelo de Random Forest, que mostró una excelente capacidad de predicción, especialmente para predicciones con grandes volúmenes de datos y variables no lineales.
  1. **Fase 5: Evaluación e Interpretación**

Se evaluaron los modelos predictivos utilizando métricas como MSE, MAE y R². Los resultados mostraron un buen rendimiento en los modelos de regresión lineal y Random Forest, con un MSE cercano a cero, mientras que la red neuronal presentó resultados más variados.

Se logró interpretar lo siguiente de cada modelo:

* **Regresión Lineal:** Logró un MSE cercano a 0 y un R² de 1.0, indicando que el modelo fue capaz de predecir los goles con gran precisión.
* **Random Forest:** También obtuvo resultados muy positivos, con un R² cercano a 1.0 y un MSE bajo, lo que lo convierte en una excelente opción para predicciones no lineales.
* **Red Neuronal (MLP):** Aunque los resultados fueron buenos, no alcanzaron el nivel de precisión de los modelos anteriores, con un R² más bajo y un MSE relativamente más alto.

Respecto a los clústeres, en la fase de minería de datos, se agruparon a los jugadores en clústeres según sus características clave (por ejemplo, goles, asistencias, etc.). Cada posición de jugador (porteros, defensas, mediocampistas, delanteros) mostró características específicas.

* **Porteros:** Para el caso del subconjunto de porteros, estos mostraron diferencias claras en el rendimiento de los porteros, particularmente en variables como *Saves* y goles concedidos. Los porteros más destacados en estos clústeres son probablemente aquellos con mejores estadísticas defensivas y de reacción.
* **Defensas:** Los defensores se agruparon principalmente según su participación en *tackles* y pases clave. Los defensores más activos en el juego también creaban más oportunidades para sus compañeros.
* **Mediocampistas:** La segmentación de mediocampistas reveló dos tipos de jugadores: uno más ofensivo (con altos valores de *xAG* y pases clave) y otro más defensivo.
* **Delanteros:** Para los delanteros, los clústeres indicaron la distinción entre jugadores más enfocados en goles y aquellos que también tienen una buena capacidad de asistencia.

Finalmente, se concluyó que, con la fase de evaluación e interpretación, se logró identificar modelos que predicen bien el rendimiento de los jugadores, segmentado jugadores en clústeres según sus características y extraído conclusiones que pueden ser útiles tanto para entrenadores como para analistas de datos deportivos. Ahora, con esta información, se pueden tomar decisiones más informadas sobre la estrategia del equipo, tácticas, y evaluaciones individuales de los jugadores.

* 1. **Fase 6: Presentación del conocimiento**

Los resultados obtenidos fueron presentados mediante gráficos y análisis de rendimiento, como los gráficos de dispersión para la regresión, matrices de confusión para la clasificación binaria y los diagramas de clústeres para los grupos de jugadores. Esta visualización permitió interpretar los hallazgos de manera más clara y comunicarlos de manera comprensible.

A partir de la validación de los modelos, se concluyó lo siguiente:

* **Modelo más eficiente:** El modelo de Regresión Lineal resultó ser el más preciso para predecir goles, con un R² de 1.0, lo que sugiere que existe una relación lineal fuerte entre las características de los jugadores y sus goles.
* **Modelo alternativo:** El Random Forest también mostró un rendimiento impresionante, especialmente cuando se trabaja con datos no lineales y de mayor complejidad, lo que lo convierte en una excelente opción en escenarios más complejos.
* **Desempeño de la red neuronal:** Aunque las redes neuronales (MLP) no lograron el mismo nivel de precisión que los modelos anteriores, se observó que, para tareas de predicción más complejas, su rendimiento podría mejorar con un ajuste adecuado de parámetros.

El análisis llevado a cabo a lo largo de este proyecto ha proporcionado una visión integral y detallada sobre el rendimiento de los futbolistas, permitiendo identificar patrones clave y realizar predicciones precisas sobre diversas métricas relevantes, como los goles y las asistencias. El enfoque adoptado, que incluye el uso de técnicas avanzadas de minería de datos y aprendizaje automático, ha demostrado ser altamente eficaz en la mejora de la toma de decisiones dentro del ámbito deportivo, especialmente en lo que respecta al análisis y evaluación del rendimiento de los jugadores.

En primer lugar, el proceso de minería de datos, que incluyó métodos como el clustering con K-Means, permitió segmentar a los jugadores en diferentes grupos según sus características clave, como la capacidad de generar goles o su rol en la defensa. Esta segmentación permitió entender mejor cómo las diferentes posiciones y estilos de juego impactan el rendimiento individual y colectivo. Además, la aplicación de modelos predictivos, tales como la regresión lineal, Random Forest y las redes neuronales (MLP), brindó una visión más profunda de las relaciones entre las variables y su capacidad para predecir con precisión métricas clave del rendimiento, como los goles, asistencias y otras estadísticas relacionadas.

El uso de técnicas de aprendizaje automático permitió validar la calidad de los modelos y la precisión de las predicciones. En particular, la regresión lineal y el Random Forest demostraron ser altamente efectivos al proporcionar predicciones con errores mínimos y un ajuste preciso a los datos. Aunque las redes neuronales también mostraron potencial, su rendimiento fue comparativamente menor, lo que nos lleva a la conclusión de que, para el caso específico de este análisis, modelos más simples y directos pueden ser igualmente efectivos.

Finalmente, los resultados obtenidos no solo validan la efectividad de los modelos aplicados, sino que también proporcionan una base sólida para la futura investigación y el desarrollo de aplicaciones prácticas en el campo del análisis de rendimiento deportivo. Las técnicas empleadas ofrecen una herramienta poderosa para entender mejor el impacto de las características individuales y colectivas en el rendimiento de los jugadores, lo que puede resultar crucial para la toma de decisiones tanto en la planificación de estrategias como en la evaluación del talento en el ámbito futbolístico. Estos enfoques pueden ser la base de análisis más profundos y aplicaciones futuras en la mejora continua del rendimiento de los jugadores y equipos.