GRUPO 1

MODELO RANDOM FOREST PREMIER LEAGUE

DIEGO LOYO

JULIAN GOMEZ

IA EXPLORADOR -G223

TALENTO TECH

INSTITUCION UNIVERSITARIA DE ENVIGADO

INTRODUCCION

El fútbol es mucho más que un deporte: es emoción, identidad y pasión compartida por millones de personas en todo el mundo. Cada partido representa una historia única llena de momentos inolvidables, decisiones estratégicas y resultados que marcan a jugadores, clubes y aficionados. Sin embargo, más allá de lo que ocurre en el césped, el fútbol moderno también se cuenta a través de los datos.

Hoy en día, cada pase, disparo, falta o tarjeta queda registrado, permitiendo que entrenadores, analistas e incluso hinchas comprendan el juego desde una nueva perspectiva: la del análisis estadístico. A medida que las herramientas tecnológicas se integran en el deporte, el estudio de los datos se ha convertido en una herramienta fundamental para entender cómo y por qué se gana o se pierde un partido.

Este proyecto nace precisamente de esa curiosidad por entender el fútbol desde los números. Usando un conjunto de datos que recopila información detallada de los partidos de la Premier League desde la temporada 2000/2001 hasta la temporada 2024/2025, nos proponemos descubrir patrones, tendencias y comportamientos que pueden pasar desapercibidos a simple vista, pero que dicen mucho sobre la evolución del juego.

La Premier League, conocida por su ritmo intenso, su competitividad y su enorme visibilidad internacional, ofrece un contexto ideal para este tipo de análisis. ¿Cuánto influye jugar en casa? ¿Qué relación hay entre los goles en el primer tiempo y el resultado final? ¿Qué equipos son más disciplinados o agresivos? Estas y muchas otras preguntas guían el desarrollo de este trabajo.

En definitiva, se trata de mirar el fútbol con otros ojos: con la mente analítica y el corazón de aficionado, combinando emoción y evidencia para conocer más a fondo este deporte que nos une.

Objetivo General

Desarrollar un modelo predictivo que permita anticipar los resultados del próximo campeonato de la Premier League, basado en el análisis estadístico de temporadas anteriores.

Objetivos Específicos

1. Analizar y preprocesar los datos históricos del campeonato de la Premier League (2000-2025) para identificar patrones relevantes que influyen en el rendimiento de los equipos.

2. Construir y Comparar distintos modelos de aprendizaje automático (regresión lineal, Random Forest, XGBoost, etc.) para seleccionar el que ofrezca la mejor capacidad predictiva sobre resultados, goles y rendimiento general.

3. . Generar predicciones para el próximo campeonato de la Premier League, incluyendo posibles campeones, rendimiento por equipo y estadísticas clave como goles anotados, partidos ganados y faltas cometidas.

DESARROLLO DEL PROYECTO

1. Descripción del Dataset

El DataSet implementado para este proyecto es una base de datos sintética, originaria de Kaggle llamada “English Premier Data” que cuenta con más de 9300 registros y 22 columnas iniciales. Información sobre cada partido y sus respectivos datos de cada uno de los equipos.

2. Columnas del Dataset

El archivo analizado contiene información de partidos oficiales de la Premier League con las siguientes columnas clave:

- Season: Temporada del torneo (2000/01-2024/25).
- Match Date: Fecha del encuentro.
- Hometeam y AwayTeam: Equipos que jugaron como local y visitante.
- FullTimeHomeGoals, FullTimeAwayGoals: Goles anotados por cada equipo.
- FullTimeResult: Resultado final (H = gana el local, A = gana el visitante, D = empate).
- HalfTimeHomeGoals, HalfTimeAwayGoals: Goles al descanso.
- HomeShotsOnTarget, AwayShotsOnTarget: Disparos a puerta.
- HomeCorners, AwayCorners: Tiros de esquina.
- HomeFouls, AwayFouls: Faltas cometidas.
- HomeYellowCards, AwayYellowCards: Tarjetas amarillas.
- HomeRedCards, AwayRedCards: Tarjetas rojas.

```
[25]    1 df.columns

Index(['Season', 'MatchDate', 'HomeTeam', 'AwayTeam', 'FullTimeHomeGoals',
       'FullTimeAwayGoals', 'FullTimeResult', 'HalfTimeHomeGoals',
       'HalfTimeAwayGoals', 'HalfTimeResult', 'HomeShots', 'AwayShots',
       'HomeShotsOnTarget', 'AwayShotsOnTarget', 'HomeCorners', 'AwayCorners',
       'HomeFouls', 'AwayFouls', 'HomeYellowCards', 'AwayYellowCards',
       'HomeRedCards', 'AwayRedCards', 'Winner', 'TotalGoals'],
      dtype='object')
```

2. Procesamiento y Limpieza de los datos

Se verificó que los datos estuvieran correctamente formateados, las fechas fueron transformadas a formato estándar y se validó que no existieran valores nulos significativos que afectaran el análisis. A continuación, se agruparon los datos por temporada y por equipos para facilitar el análisis comparativo.

Season	0
MatchDate	0
FullTimeHomeGoals	0
FullTimeAwayGoals	0
FullTimeResult	0
HalfTimeHomeGoals	0
HalfTimeAwayGoals	0
HalfTimeResult	0
HomeShots	0
AwayShots	0
HomeShotsOnTarget	0
AwayShotsOnTarget	0
HomeCorners	0
AwayCorners	0
HomeFouls	0
AwayFouls	0
HomeYellowCards	0
AwayYellowCards	0
HomeRedCards	0
AwayRedCards	0
match_year	0
match_month	0
match_day	0
HomeTeamID	0
AwayTeamID	0
local_goles_favor	0
local_goles_contra	0
local_corners	0
local_precision_tiros	0
local_victorias	0
local_derrotas	0
local_amarillas	0
local_rojas	0
visit_goles_favor	0
visit_goles_contra	0
visit_corners	0

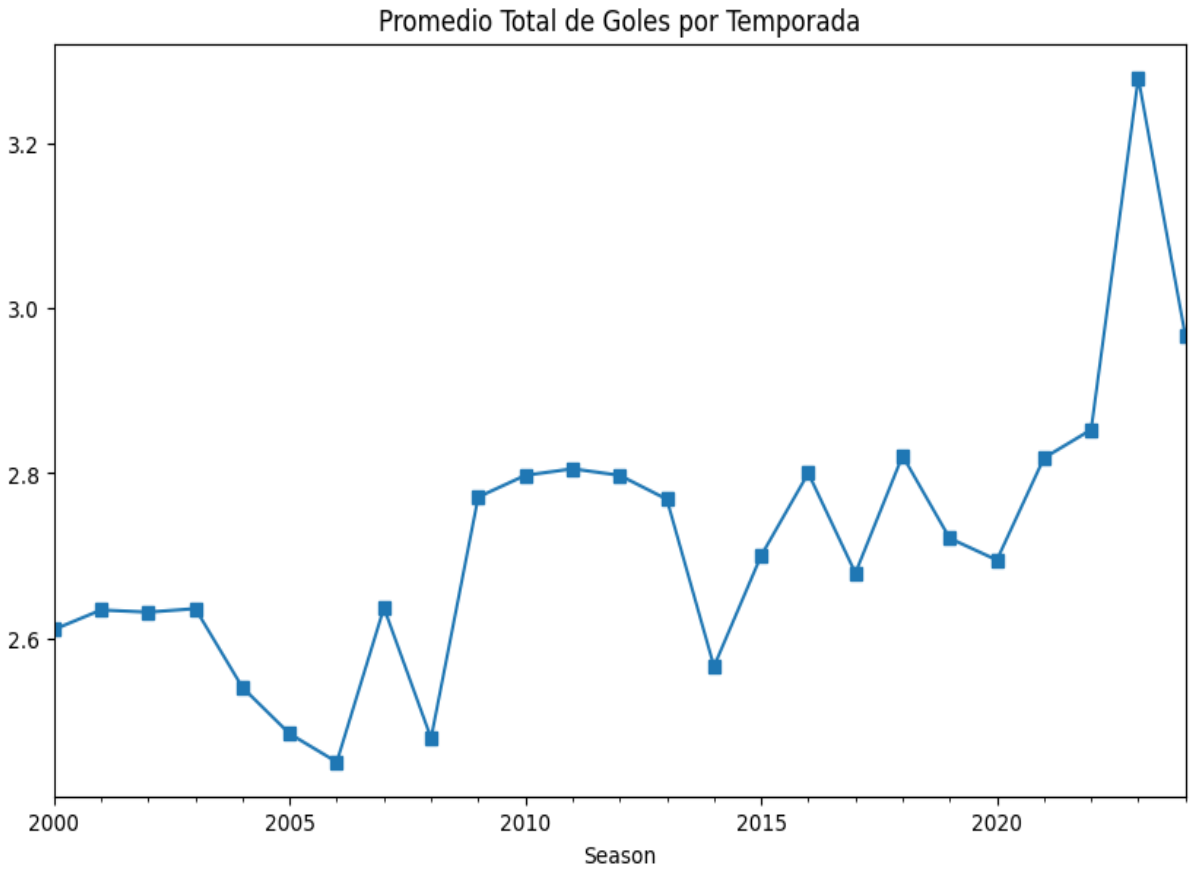
3. Objetivos del Análisis

El proyecto tiene como propósito realizar un análisis descriptivo y comparativo del rendimiento de los equipos a lo largo de múltiples temporadas, considerando:

- Promedio de victorias de los equipos locales y visitantes.

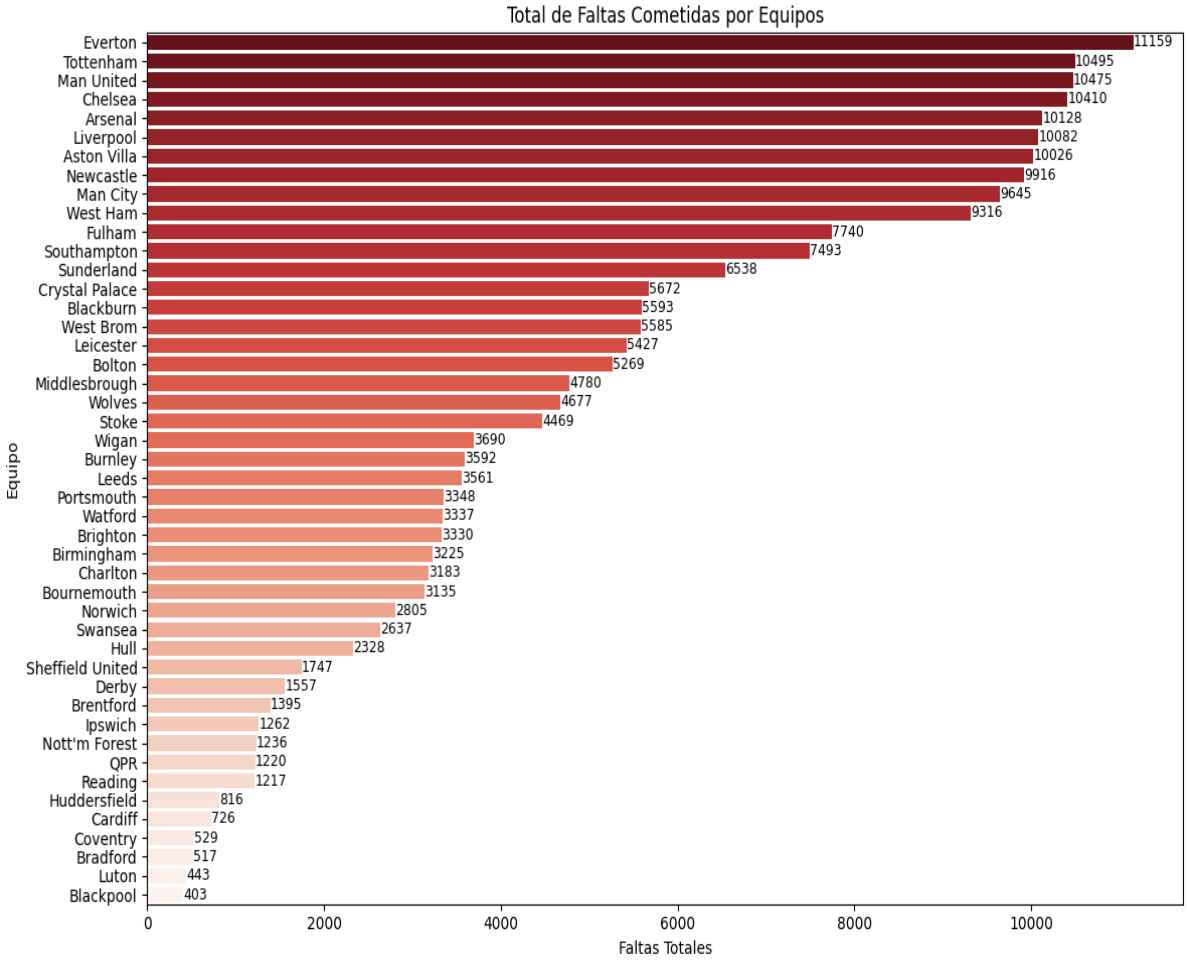
Podemos observar en el grafico que en promedio casi el doble de victorias de los equipos se da en casa, lo que nos da a entender que la Localia es un factor determinante y que afecta el resultado del juego.

- Promedio total de goles por temporada.



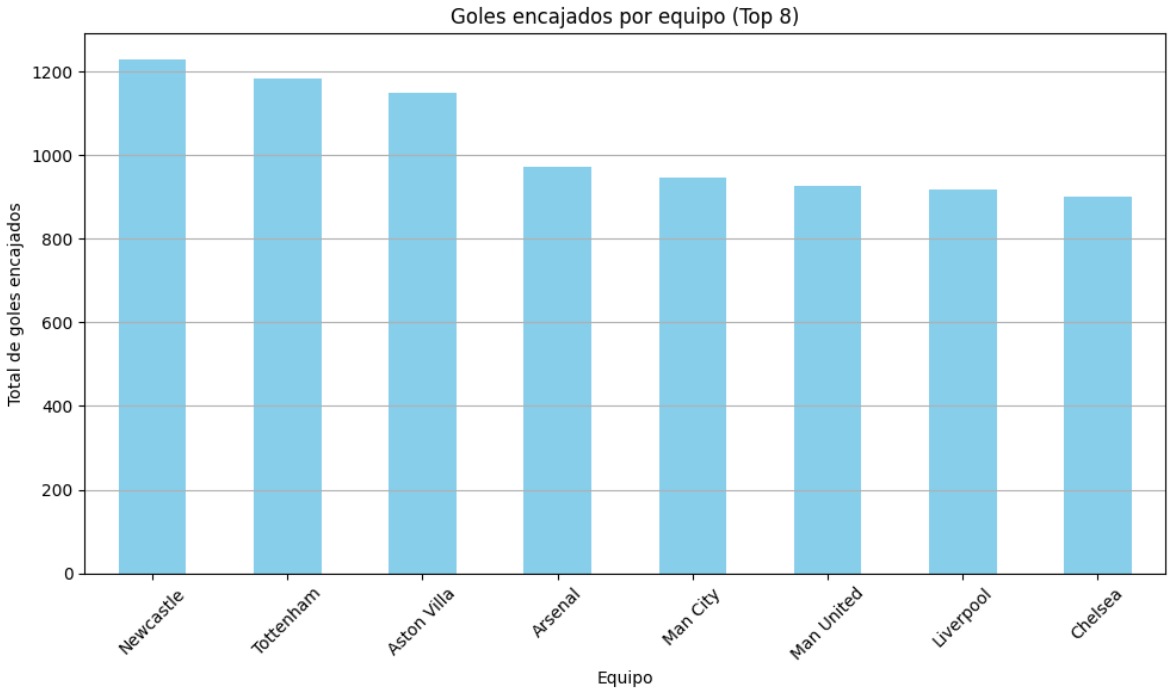
Con gráfico, interactuamos con columnas de goles. En las cuales generamos un total de goles por cada temporada y hacemos su comparación la cual nos puede relacionar eventos reales como un bajón de goles en el año 2020 (pandemia), la temporada 2005-2006 la Premier experimento un bajo de goles debido a implementación de un juego más defensivo en sus equipos y un aumento de grandes porteros que quedarían para la historia.

- Total, de faltas cometidas por equipo.



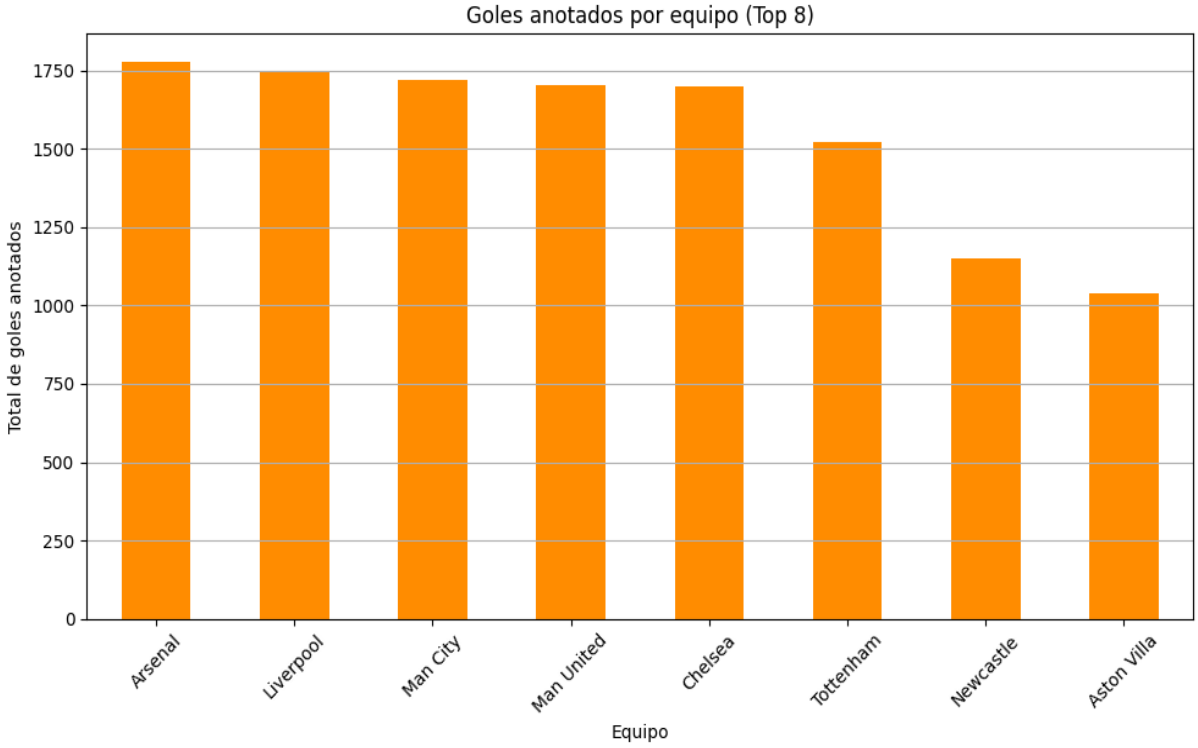
En esta grafica de valores de faltas, podemos evaluar la cantidad de faltas de los equipos en lo largo de las 2024 temporadas, dándonos a mostrar que el Everton ha cometido más de 10.000 mil faltas.

- Top 8 de goles encajados por equipo.



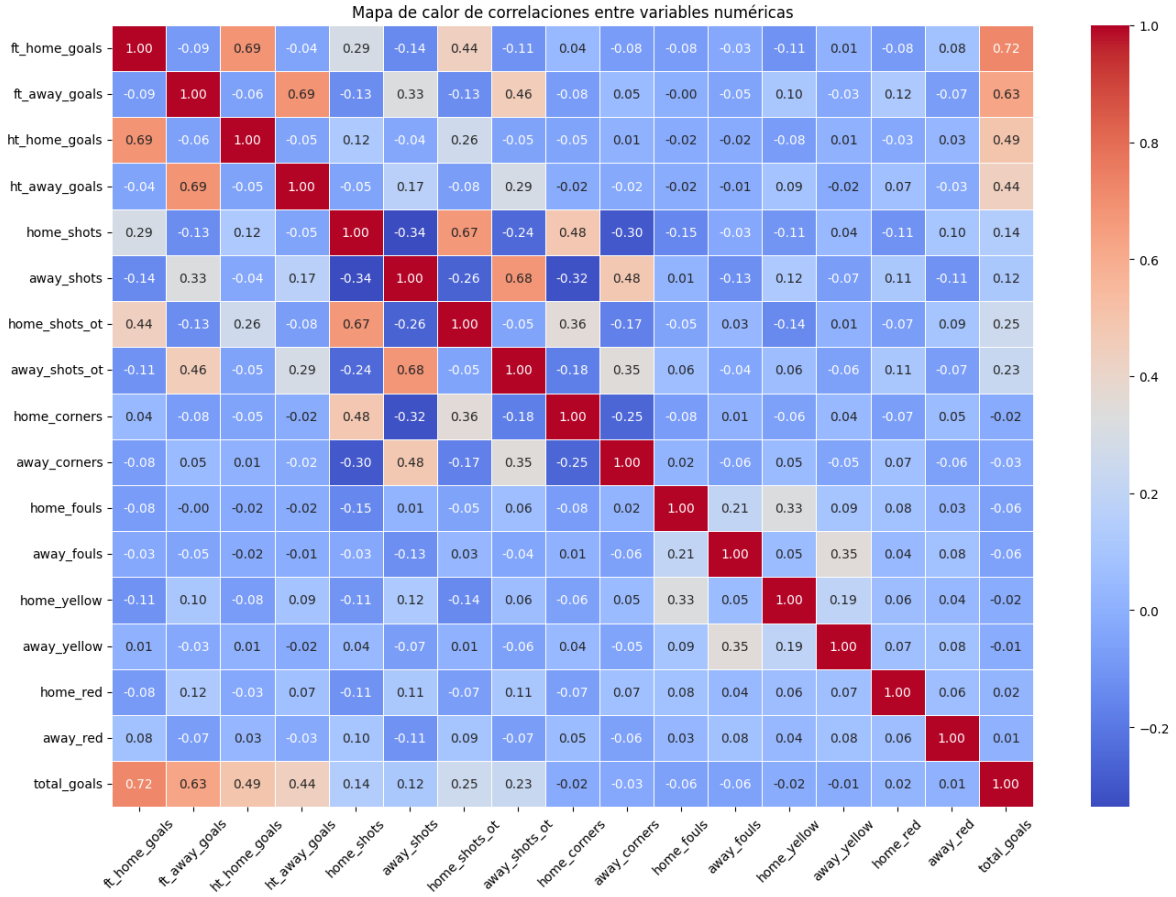
Con esta grafica buscamos ilustrar los equipos que más goles han sufrido a lo largo de las temporadas.

- Top 8 de goles anotados por equipo.



Con esta grafica buscamos ilustrar los equipos que más goles han hecho a lo largo de las temporadas.

MATRIZ DE CORRELACION



La matriz de correlación nos muestra qué tan relacionadas están las estadísticas del partido entre sí. Cuanto más intenso el color, más fuerte es la relación entre dos variables, como entre tiros y goles.

METRICAS DEL MODELO

Se evaluaron distintos modelos como (Regresión Lineal, XGBoost, RamdomForest):

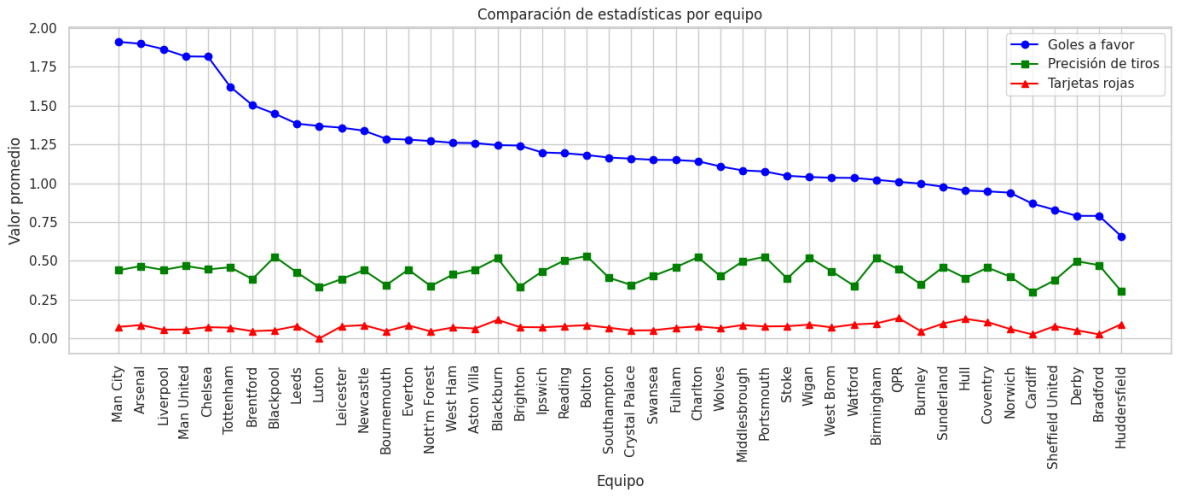
Precisión (Accuracy): 92%

MSE (Error cuadrático medio): 1.0

F1 Score: 1.00

Observando las métricas del modelo RamdomForest podemos ver una predicción del 92% esto no es muy común, sin embargo, por ser datos con fechas, tener gran abundancia de información y un modelo bien alimentado, es esperado esto. El promedio F1 y Score afectarían los porcentajes menore, serían el triunfo visitante y el local.

COMPARACION ESTADISTICA POR EQUIPO



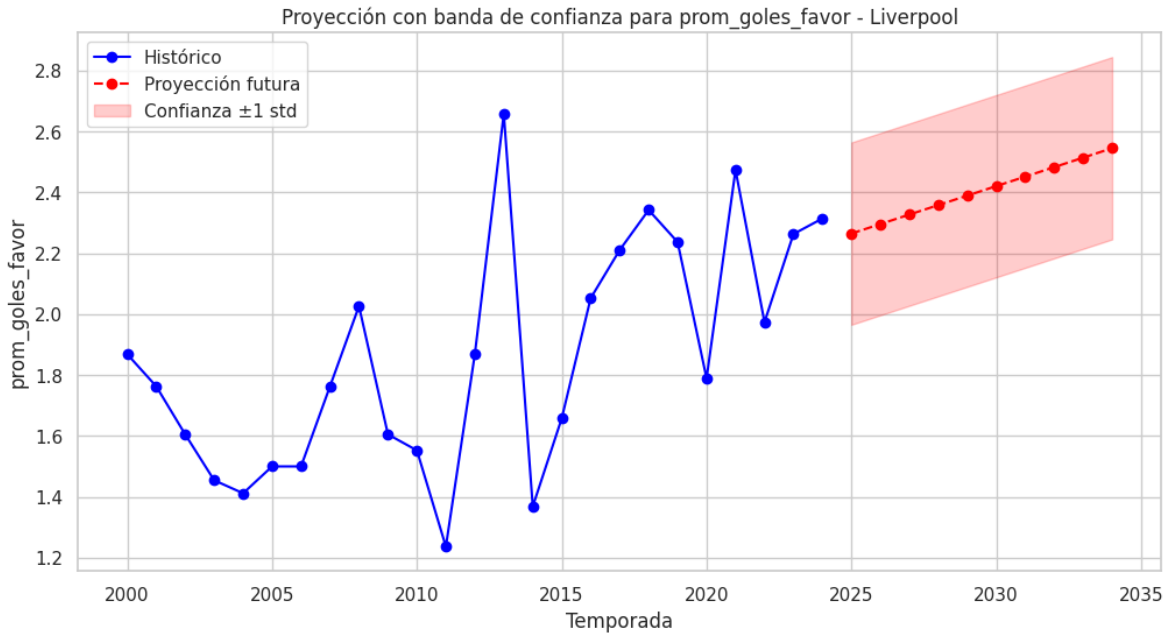
Promedio de goles a favor (prom_goles_favor) → Muestra el poder ofensivo del equipo.

Precisión de tiros al arco (precision_tiros) → Mide qué tan eficaces son los disparos (calidad de ataque).

Promedio de tarjetas rojas (prom_rojas) → Refleja el juego agresivo o indisciplina.

Estas tres métricas nos dan un panorama: ofensiva, eficacia y disciplina.

PROYECCION LIVERPOOL



A lo largo de los años, el rendimiento ofensivo del Liverpool ha tenido altibajos, pero desde 2015 se nota una clara recuperación. El modelo proyecta que el equipo mantendrá una tendencia al alza en sus goles a favor durante la próxima década. Aunque hay incertidumbre (representada por la banda roja), se espera que Liverpool siga siendo un equipo fuerte en ataque. Esto no solo habla de estadísticas, sino de una identidad ofensiva que parece consolidarse con el tiempo.

Recomendaciones y Proyecciones futuras

Por ultimó queremos resaltar los conocimientos adquiridos en este Bootcamp y la implementación de los conceptos vistos en clase por medio de los docentes. A futuro, la continuación del proyecto es mejorar cada día más las métricas para acercarnos a un resultado real.

CONCLUSIONES DEL PROYECTO

- Jugar en casa sigue marcando la diferencia, a lo largo del análisis, quedó claro que el hecho de jugar como local influye de manera significativa en los resultados. Más allá de las estadísticas, se nota que el apoyo de la hinchada, el conocimiento del campo y la energía emocional que se siente al estar “en casa” pueden inclinar la balanza a favor del equipo anfitrión.

- Los datos nos ayudan a ver lo que los ojos no siempre notan, Este proyecto demuestra que el fútbol es mucho más que lo que ocurre en 90 minutos. Al mirar hacia atrás, temporada tras temporada, los números revelan tendencias, cambios y aprendizajes que enriquecen nuestra forma de entender el juego. Analizar fútbol con datos no le quita magia, al contrario: nos ayuda a admirarlo desde otra perspectiva.

CONCLUSION GENERAL

Analizar los datos de la Premier League fue más que revisar estadísticas o contar goles. Fue una manera de redescubrir el fútbol desde otra perspectiva, más profunda y objetiva, pero sin perder la emoción que lo hace único. Cada número, cada resultado, cada tarjeta o disparo, nos cuenta una parte de la historia que se vive cada fin de semana en los estadios.

Este proyecto nos permitió ver cómo ciertos patrones se repiten, cómo la localía sigue siendo poderosa, cómo un buen arranque puede marcar el rumbo de un partido, o cómo hay equipos que, más allá de ganar o perder, mantienen una identidad clara en su forma de jugar.

REPOSITORIO GID

https://github.com/DiegoLoy0/TalentoTech

BIBLIOGRAFIA

1. Kaggle. (2024). *English Premier League Dataset (2000–2024)*. Recuperado de:

(https://www.kaggle.com/datasets/irkaal/english-football-results)

2. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

(https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html)

3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

https://www.statlearning.com/

4. Raschka, S. (2015). *Python Machine Learning: Unlock Deeper Insights into Machine Learning with Python*. Packt Publishing.

https://sebastianraschka.com/books.html

5. Premier League. (2024). *Official Statistics and Historical Data*.

https://www.premierleague.com/stats