# Premier League Match Outcome Prediction

## 1. Introduction

This project aims to predict the results of English Premier League (EPL) matches based on historical data and team performance metrics. The predictions classify matches as either a "Win" for the home team or "Draw or Lose".

## 2. Dataset

The dataset comprises historical English Premier League (EPL) match data, featuring key performance indicators for both the home and away teams. These include performance metrics derived from the last five matches, such as:

- Wins (Last5Wins) of Home Team and Away Team
- Average shots (Last5AvgSh) of Home Team and Away Team
- Average shots on target (Last5AvgSot) of Home Team and Away Team
- Average goals scored (Last5AvgGf) of Home Team and Away Team
- Average goals conceded (Last5AvgGa) of Home Team and Away Team

This data has been sourced by scraping publicly available football databases FBref. The dataset spans from the 2020 season to the present, ensuring a comprehensive representation of team performances and trends over multiple seasons. All "last five matches" metrics (e.g., wins, average shots, goals scored/conceded) were derived from the scraped data through additional calculations performed by us.

## 3. Methodology

In this project, data preprocessing involved encoding match results as 0 (Win) or 1 (Draw or Lose), venues as 0 (Home) or 1 (Away). A custom function was used to calculate the "last five matches" metrics for each team prior to every match, such as wins, average shots, and goals scored/conceded. For feature engineering, the dataset was structured to pair each home team with its corresponding away team, integrating their respective performance metrics. The model used for prediction was a Random Forest Classifier, trained on features derived from the last five matches for both teams. The dataset was split into an 80/20 train-test split to evaluate performance. Finally, the trained model was applied to predict the outcomes of upcoming fixtures, with the results saved in a tab-delimited file.

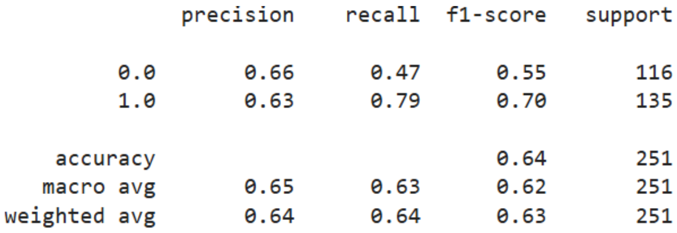
As a stand-alone serverless ML system, we utilized GitHub Actions to automate the daily process of scraping the latest data and predicting the outcomes of the next 10 matches. This workflow ensures the system remains up-to-date with real-time data, providing timely and accurate predictions without the need for manual intervention or dedicated server infrastructure.

## 4. Results

- **Model Performance**

  Test accuracy: 64%

  Detailed classification report:



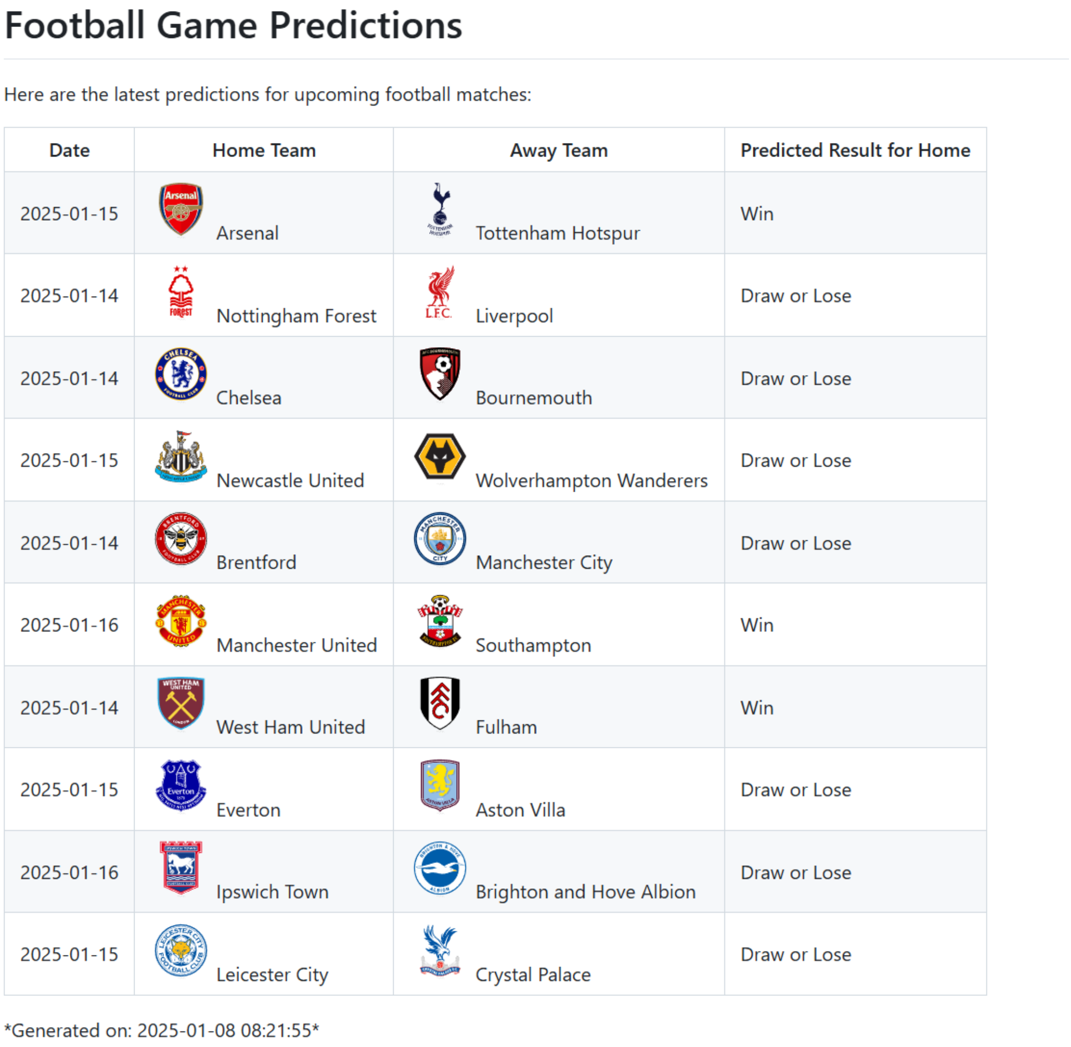
```
              precision    recall  f1-score   support

         0.0       0.66      0.47      0.55       116
         1.0       0.63      0.79      0.70       135

    accuracy                           0.64       251
   macro avg       0.65      0.63      0.62       251
weighted avg       0.64      0.64      0.63       251
```

- **Predictions:**

Upcoming fixtures and their predicted outcomes https://github.com/hydrixer/premierleague_predict/blob/main/code/README.md



# Football Game Predictions

Here are the latest predictions for upcoming football matches:

| Date | Home Team | Away Team | Predicted Result for Home |
|---|---|---|---|
| 2025-01-15 | Arsenal | Tottenham Hotspur | Win |
| 2025-01-14 | Nottingham Forest | Liverpool | Draw or Lose |
| 2025-01-14 | Chelsea | Bournemouth | Draw or Lose |
| 2025-01-15 | Newcastle United | Wolverhampton Wanderers | Draw or Lose |
| 2025-01-14 | Brentford | Manchester City | Draw or Lose |
| 2025-01-16 | Manchester United | Southampton | Win |
| 2025-01-14 | West Ham United | Fulham | Win |
| 2025-01-15 | Everton | Aston Villa | Draw or Lose |
| 2025-01-16 | Ipswich Town | Brighton and Hove Albion | Draw or Lose |
| 2025-01-15 | Leicester City | Crystal Palace | Draw or Lose |

*Generated on: 2025-01-08 08:21:55*

### 5. How to Run the Code

The code is deployed using github action and is configured to perform daily predictions and fetch data every 2 days. To run it from the start, first run scraping.ipynb to fetch data from fbref.com, then run forecast1.ipynb, it will train the model, then collect the match fixtures and predict results for the next matchday.