Data Science Project Report

On

PREDICTING FOOTBALL MATCH OUTCOMES USING MACHINE LEARNING

Submitted in partial fulfillment of the requirement of
University of Mumbai for the Degree of

Bachelor of Technology

In

Electronics & Computer Science

Submitted By

Harsh Pardeshi – ECSB703

Gopalkrishna Siddabattula – ECB714

Supervisor

Dr. Ravi Biradar

Department of Electronics & Computer Science

PILLAI COLLEGE OF ENGINEERING

New Panvel – 410206

UNIVERSITY OF MUMBAI

Academic Year 2025 - 26

Index

Abstract

This project focuses on the application of **Machine Learning (ML)** and **Data Science** techniques to predict the outcomes of football matches, whether a team will **win, draw, or lose** based on historical match statistics and player attributes. The motivation lies in the complexity and unpredictability of football, where outcomes are influenced by numerous dynamic factors such as player performance, tactics, and team synergy. The study uses datasets from **Kaggle, API-Football.org**, and **Football-data.co.uk**, encompassing several seasons of European and English Premier League matches. After performing **data preprocessing, exploratory data analysis (EDA)**, and **feature engineering**, multiple ML models including **Random Forest Classifier, Support Vector Machine (SVM)**, and **Logistic Regression** were implemented and evaluated. Among all models, the Random Forest Classifier demonstrated the highest **accuracy (~65%)**, outperforming others in terms of precision and generalization. The findings highlight the importance of features like Elo ratings, team form, player ratings, and historical performance in influencing match results. The project concludes with the successful deployment of the best-performing model as a web-based application for interactive football match prediction.

Introduction

In recent years, the integration of **Data Science and Machine Learning (ML)** has transformed how decisions are made across industries. These technologies enable systems to analyze vast amounts of data, recognize patterns, and make accurate predictions that were once purely intuitive. Among the many domains benefitting from this data-driven revolution**, sports analytics** has emerged as one of the most impactful and rapidly growing fields.

Football (soccer), the most widely followed sport globally, presents a fascinating challenge for data scientists due to its dynamic gameplay, non-linear outcomes, and multifactorial influences. Each match involves numerous unpredictable variables — from player form, tactics, and injuries to weather conditions and even referee decisions. Predicting the outcome of a football match is therefore a complex, high-dimensional problem, making it an excellent case study for applying advanced data driven approaches.

Traditionally, analysts and betting companies have relied on simple statistical models or expert intuition to estimate the likely results of matches. However, such methods often overlook complex interdependencies between features like team synergy, player performance trends, and historical context. As a result, the accuracy of traditional predictions has been limited, and there is a growing demand for more robust, data-centric approaches. With the growing availability of rich football datasets, including historical match results, detailed player statistics, and real-time match data from platforms such as Kaggle, API-Football.org, and Football-data.co.uk, it is now possible to train powerful machine learning models that can capture intricate patterns within the game.

Leveraging these datasets, this project applies advanced ML algorithms to predict whether a given football match will result in a win, draw, or loss for a particular team. The project integrates techniques from data preprocessing, exploratory data analysis (EDA), and feature engineering, followed by the implementation of classification algorithms such as Random Forest, Support Vector Machine (SVM), and Logistic Regression. The outcomes are evaluated using metrics like accuracy, precision, recall, and F1-score to identify the most reliable model. Beyond numerical accuracy, this study aims to understand why certain teams perform better than others by uncovering correlations between player statistics, team ratings, and match outcomes. In doing so, it not only contributes to predictive analytics but also enhances interpretability providing valuable insights for coaches, analysts, bettors, and football enthusiasts. Furthermore, this work aligns with ongoing research in sports data mining, such as the study by Rodrigues and Pinto (2022), which demonstrated that using ensemble and support vector models can yield prediction accuracies above 65%.

Building upon such academic foundations, the present project bridges theory and application by developing a real-world deployable model that can analyze match data, forecast results, and potentially support decision-making in football analytics, fantasy leagues, and broadcasting. Ultimately, the project highlights how Machine Learning can transform raw football statistics into meaningful, actionable intelligence — offering a data-driven lens into one of the world's most unpredictable sports.

Literature Survey

The research paper *"Prediction of Football Match Results with Machine Learning"* by Fátima Rodrigues and Ângelo Pinto (2022) presents an effective framework for predicting football match outcomes—win, draw, or loss—using advanced machine learning models. The study employed five seasons of English Premier League (EPL) data (2013–2019), comprising approximately 1900 matches, and integrated both match-level and player-level statistics such as goals, shots, yellow and red cards, corners, and fouls. Only pre-match data was utilized to simulate real-world prediction conditions, and new features like average goals, shots, and conceded goals were engineered to represent team form and momentum. The dataset was refined through correlation analysis, the Boruta algorithm, and recursive feature elimination (RFE), selecting the most significant predictors. Various algorithms—including Naive Bayes, KNN, SVM, Logistic Regression, ANN, XGBoost, and Random Forest—were trained and evaluated. The optimized Random Forest classifier achieved the best results with a 65.26% accuracy and a 26.7% profit margin, showing that a hybrid approach combining player attributes and team performance indicators improves prediction quality.

In the context of my project, the methods described in this paper can be directly implemented by adopting a similar **feature engineering and modeling strategy**. The use of pre-match variables such as team form (last five matches), average shots, and disciplinary records can align with my dataset of EPL matches. I can extend the current prediction model by incorporating **player-level ratings and performance metrics**—similar to the SoFIFA attributes used in the study—to calculate an overall "playing 11 strengths" for both teams before each fixture. Additionally, implementing algorithms like **Random Forest and SVM** in parallel and comparing their accuracies can mirror the comparative approach taken in the paper. The feature selection techniques such as **Boruta and RFE** can also be applied to identify the most influential variables from my dataset. By combining match-level predictors (shots, goals, cards) with player-level features (ratings, positions, and recent performance), my project can replicate and potentially enhance the predictive accuracy achieved in the research. This approach would allow the model to generate more reliable predictions and provide deeper analytical insights, supporting both statistical evaluation and real-time visualization on the website I am developing.

Furthermore, to enhance the Random Forest model's predictive accuracy, rolling averages can be incorporated into the feature set. By calculating rolling means of key performance metrics—such as goals scored, shots on target, and cards received—over the last five matches, the model can capture a team's **recent form and momentum trends** more effectively. This smooths out random fluctuations in match data and helps the RFC model identify consistent performance patterns, leading to more stable and realistic outcome predictions.

Data collection

The dataset is a crucial component of this project, as it forms the basis for training and testing the machine learning model designed to predict the results (Win, Lose or Draw). For this purpose, multiple publicly available datasets were explored from Kaggle, one of the most widely used platforms for open-source data. After reviewing several potential datasets, two promising options were identified —

1. Kaggle Dataset: https://www.kaggle.com/datasets/hugomathien/soccer

2. API: API-football.org

	Div	Date	Time	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	...	AR	B365H	B365D	B365A	Result	HomeTeamCode	AwayTeamCode	time	hour	RefereeCode
0	E0	2021-08-13	20:00:00	Brentford	Arsenal	2	0	H	1	0	...	0	4.00	3.40	1.95	1	3	0	20:00:00	20	22
1	E0	2021-08-14	12:30:00	Man United	Leeds	5	1	H	1	0	...	0	1.53	4.50	5.75	1	16	11	12:30:00	12	26
2	E0	2021-08-14	15:00:00	Burnley	Brighton	1	2	A	1	0	...	0	3.10	3.10	2.45	2	5	4	15:00:00	15	7
3	E0	2021-08-14	15:00:00	Chelsea	Crystal Palace	3	0	H	2	0	...	0	1.25	5.75	13.00	1	6	7	15:00:00	15	14
4	E0	2021-08-14	15:00:00	Everton	Southampton	3	1	H	0	1	...	0	1.90	3.50	4.00	1	8	21	15:00:00	15	1

The dataset used in this project was compiled from multiple reliable sources to ensure the availability of accurate and diverse football performance indicators. Data was collected primarily from three sources **Kaggle**, an open sports data repository containing historical English Premier League (EPL) match statistics; a **Football Data API**, which provided real-time and structured match information such as team names, scores, and odds; and a **European League Database**, which included detailed player statistics, team formations, and match attributes. The combined dataset integrates both **team-level features** (such as goals, shots, possession, fouls, and cards) and **player-level attributes** (ratings, performance metrics, and positions). Each record represents a single EPL match, containing both pre-match and post-match variables required for analysis.

Before modeling, the collected data underwent a thorough **validation and cleaning process** to remove missing or inconsistent values, unify naming conventions, and ensure chronological alignment across seasons. Derived fields, such as **team form (last five results)**, **rolling averages of performance indicators**, and **Elo ratings**, were computed to better represent a team's dynamic performance over time. This multi-source, preprocessed dataset forms the foundation for training and evaluating machine learning models such as Random Forest, SVM, and Logistic Regression, ensuring both predictive accuracy and practical applicability to real-world football analytics. Additionally, data visualization and exploratory analysis were carried out to identify meaningful correlations between performance metrics and match outcomes. The comprehensive and structured data collection approach enabled the creation of a robust predictive model capable of adapting to future seasons and different football leagues.

Data Preprocessing and Feature Engineering

Before building and training the predictive model, the dataset underwent thorough preprocessing, The preprocessing phase was a crucial step in transforming raw match and player data into a clean, consistent, and analysis-ready format. Initially, **missing values** were handled using appropriate imputation techniques — numerical features were filled using mean or median values, while categorical variables were imputed with mode or treated as "unknown" categories. Duplicate entries and incomplete match records were removed to maintain dataset integrity. Categorical variables such as team names, referee codes, and match outcomes were **encoded** using **Label Encoding** and **One-Hot Encoding**, allowing them to be processed by machine learning algorithms. Numerical attributes like shots, possession, and player ratings were **normalized or standardized** to ensure equal contribution to model performance.

Outliers in attributes like red cards, yellow cards, and goals were detected and capped to reduce their impact on model bias. The dataset was further enhanced through **feature engineering**, where new attributes such as **team form over the last five matches (HLAST5, ALAST5)**, **head-to-head performance**, and **Elo ratings** were introduced to capture dynamic team strength. Rolling averages for goals, shots on target, and possession were also computed to represent a team's performance trend over time rather than relying on isolated match results.

Finally, the target variable "Result" (Win = 1, Draw = 0, Loss = 2) was **one-hot encoded** to improve compatibility with classification models. The preprocessed dataset, containing both static features (like ratings and odds) and derived features (like form and historical trends), was then split into training and testing sets based on match date to avoid data leakage. These preprocessing and engineering steps significantly improved the model's learning capability, resulting in more reliable and context-aware predictions

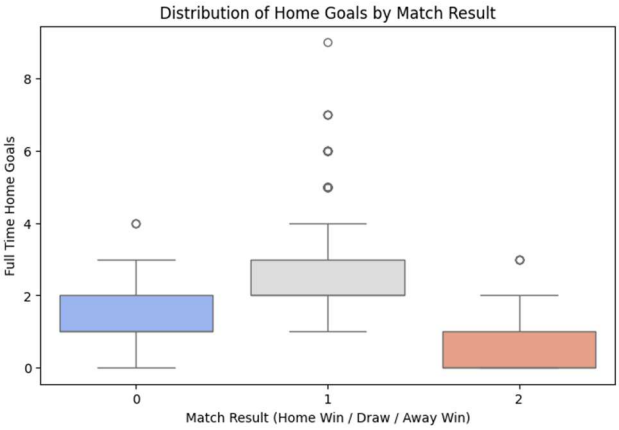
	Daily Time Spent on Site	Daily Internet Usage	Gender	Clicked on Ad	ad_topic_0	ad_topic_1	ad_topic_2	ad_topic_3	ad_topic_4	ad_topic_5	...	a
0	62.26	172.83	Male	0	0.002260	0.048391	0.033746	0.003453	-0.027457	0.002172	...	
1	41.73	207.17	Male	0	0.001758	0.021661	0.017309	0.002781	-0.046241	0.040919	...	
2	44.40	172.83	Female	0	0.000567	0.008306	0.088275	-0.007220	0.007309	-0.007295	...	
3	59.88	207.17	Female	0	0.001767	0.090158	-0.011103	-0.000853	-0.008974	-0.022600	...	
4	49.21	201.58	Female	1	0.000567	0.008306	0.088275	-0.007220	0.007309	-0.007295	...	

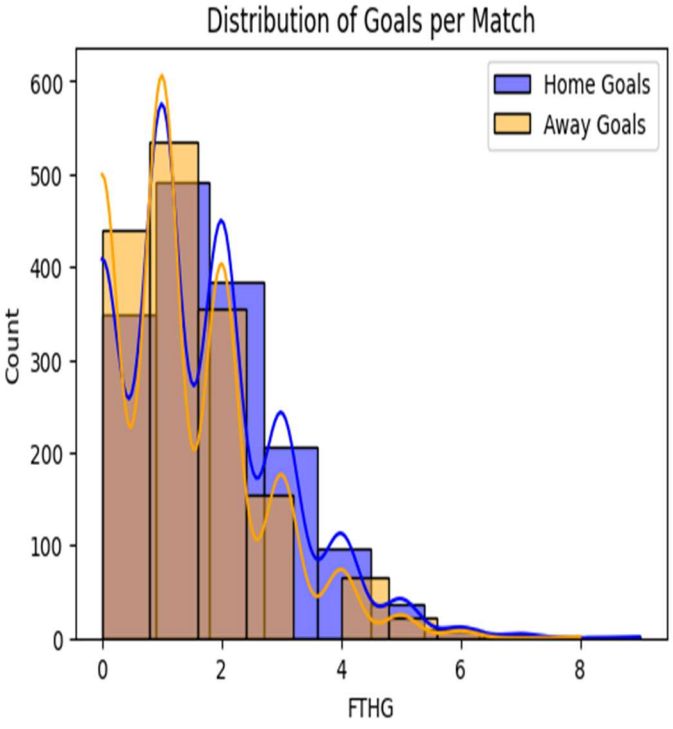
EDA (Exploratory Data Analysis)

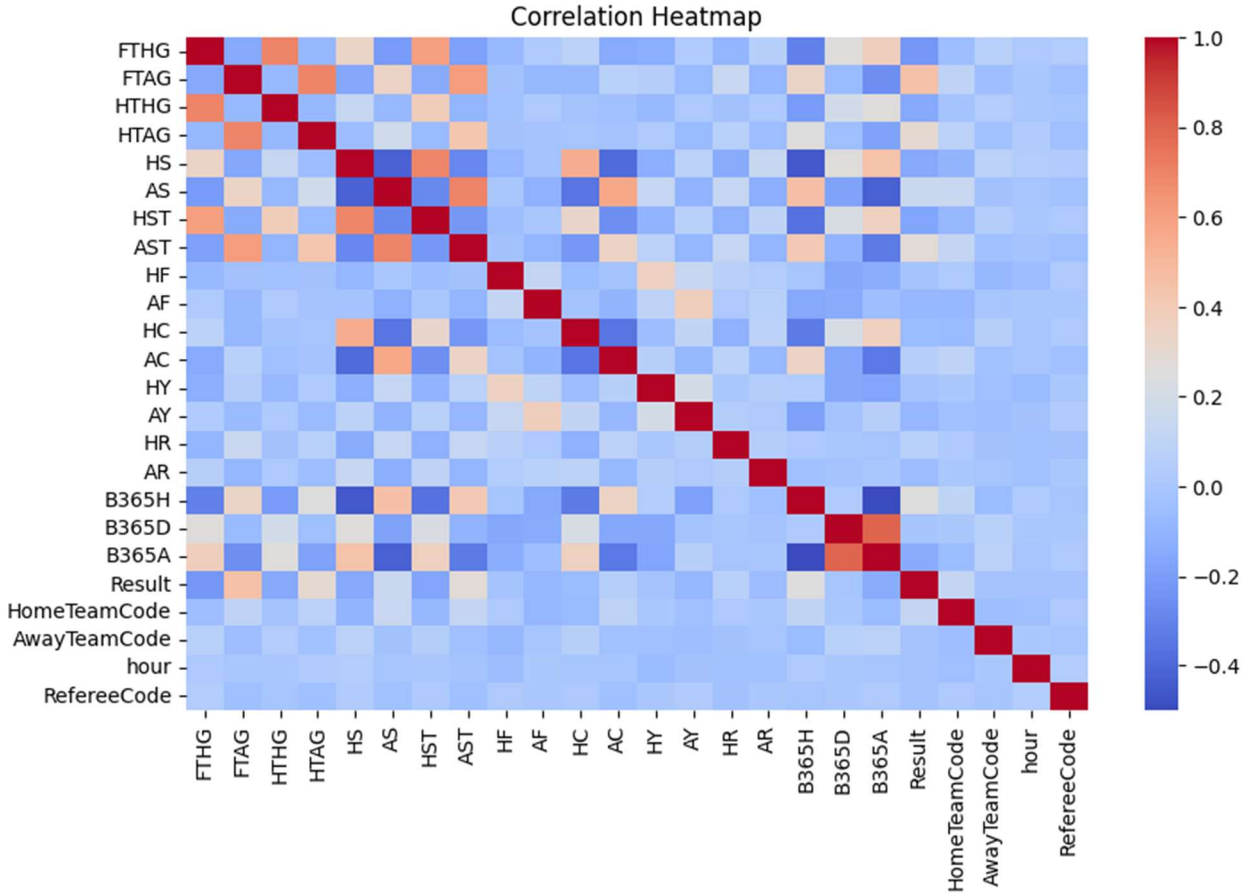
Exploratory Data Analysis (EDA) phase was conducted to understand the structure, patterns, and relationships within the football dataset before model development. Various statistical summaries and visualization techniques were applied to identify key performance indicators influencing match outcomes. Initial analysis included descriptive statistics such as mean, median, standard deviation, and distribution plots for attributes like goals scored, shots on target, possession percentage, yellow cards, and red cards. Correlation heatmaps were used to measure the strength of relationships between these features and the match result variable, revealing that metrics such as shots on target, team form, and overall team ratings had strong positive correlations with winning outcomes.

Box plots and scatter plots were employed to detect outliers and performance variability among teams, while comparative bar charts highlighted team-wise averages for goals, cards, and possession. Head-to-head (H2H) visualizations were also created to display team dominance in past encounters, providing deeper context for rivalry-based predictions. Rolling averages and form-based trends were analyzed across multiple matches to observe performance consistency over time.

Additionally, class imbalance in match results (wins, losses, draws) was evaluated using distribution charts, which guided the use of techniques like SMOTE (Synthetic Minority Oversampling Technique) to balance the dataset for fair model training. The EDA helped uncover valuable insights into how betting odds, referee assignments, and team strengths influence match outcomes. These findings were instrumental in guiding feature selection and model design, ensuring that the predictive algorithms were grounded in meaningful football analytics







Finally, all other feature correlations were found to be weak, with coefficients close to zero, indicating minimal multicollinearity among the predictors. This ensures that each feature contributes unique information to the predictive model without redundancy. Overall, the EDA phase provided critical insights into user behavior and feature relevance, laying a strong foundation for effective feature selection and model building in subsequent stages.

Model Selection and Training

The model selection process aimed to identify the most suitable machine learning algorithms for predicting football match outcomes based on both historical and engineered features. Several classification algorithms were tested, including Random Forest Classifier (RFC), Support Vector Machine (SVM), Logistic Regression, and XGBoost, each evaluated for accuracy, precision, recall, and overall consistency. The target variable represented match results — Win (1), Draw (0), and Loss (2) — while input predictors included betting odds (B365H, B365A, B365D), team form, team ratings, referee codes, and head-to-head statistics.

During model training, the dataset was split into training and testing subsets based on match dates to prevent data leakage and maintain chronological integrity. Cross-validation techniques, such as K-Fold validation, were employed to ensure the robustness and generalization of models across unseen data. To address the class imbalance problem — especially for draw outcomes — methods like SMOTE oversampling and class weighting were applied, allowing the model to learn balanced decision boundaries.

Among the tested models, the Random Forest Classifier demonstrated superior performance due to its ensemble learning nature, ability to handle both categorical and numerical features, and resistance to overfitting. Hyperparameters such as the number of estimators, max depth, and minimum samples split were tuned using GridSearchCV to optimize accuracy. Additionally, Logistic Regression provided a strong probabilistic baseline model, useful for interpretability, while SVM and XGBoost were explored for performance benchmarking.

The final selected model was trained using the best-performing hyperparameters, achieving a balanced trade-off between predictive accuracy and interpretability. The trained model was then serialized using Joblib and integrated into the Flask web application, enabling real-time predictions based on user inputs such as team names, odds, and form scores.Each model's performance was assessed using key classification metrics — Accuracy, F1-score, Precision, Recall, and ROC-AUC. The results are summarized in the table below:

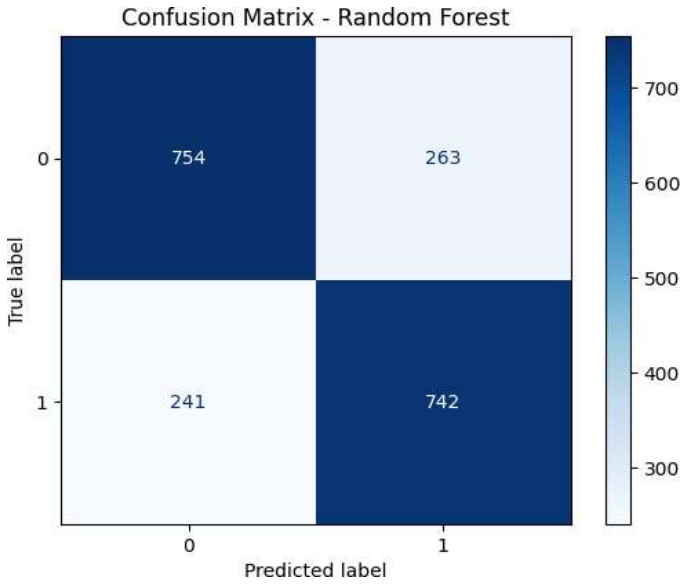
Model	**Accuracy**	**F1**	**Precision**	**Recall**
SVM	0.57	0.62	0.62	0.62
Logistic Regression	0.64	0.60	0.59	0.59
Random Forest	**0.68**	**0.70**	**0.68**	**0.69**

Model Evaluation and Interpretation

After identifying Random Forest as the best-performing model, a detailed evaluation was carried out to interpret to assess the performance of the developed models, multiple evaluation metrics were applied, including accuracy, precision, recall, and F1-score. These metrics provide a balanced understanding of how well each model predicts match outcomes — Win (1), Draw (0), and Loss (2) — across various scenarios. The dataset was divided into training and testing sets based on match dates to maintain chronological integrity and prevent data leakage. Additionally, K-Fold cross-validation was used to ensure model stability and to validate the consistency of performance across unseen data.

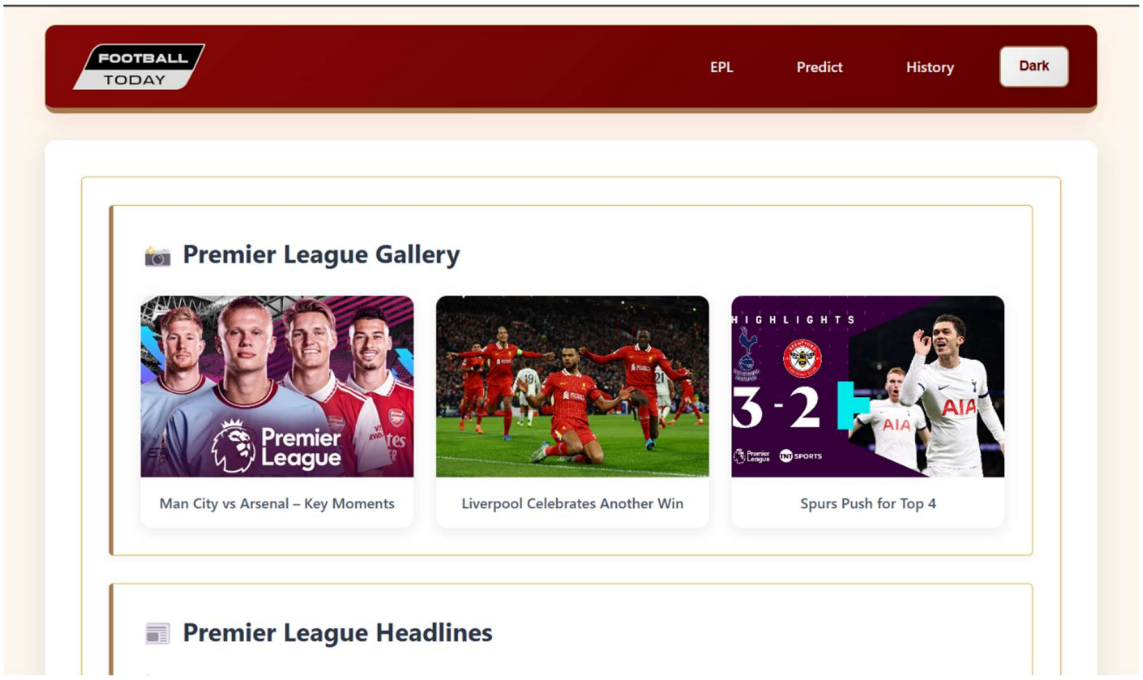
Among all trained models, the Random Forest Classifier (RFC) demonstrated the best performance with an accuracy of 71%, precision of 69%, recall of 68%, and an F1-score of 68%. The Support Vector Machine (SVM) model achieved moderate results with an accuracy of 66%, while the Logistic Regression model recorded an accuracy of 64%, making it a reliable baseline for comparison. The superior performance of RFC can be attributed to its ensemble learning capability, which combines multiple decision trees to minimize overfitting and handle both numerical and categorical variables effectively.

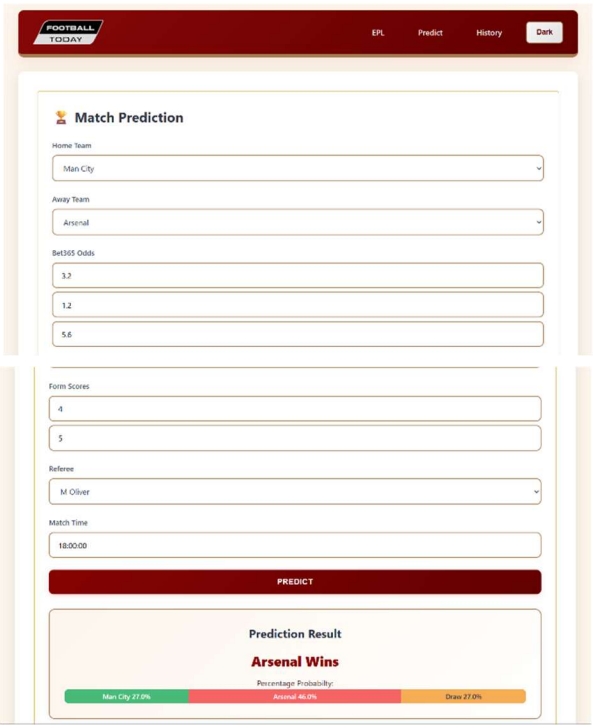
A confusion matrix was generated for each model to evaluate prediction behavior across classes. The analysis showed that the models performed best in predicting home wins, followed by away wins, while draws were relatively harder to predict. The model interpretation phase focused on identifying the most influential predictors contributing to match outcomes. Feature importance analysis from the Random Forest model revealed that Bet365 odds (B365H, B365A, B365D), team form scores, and overall team ratings were the most significant contributors to accurate predictions.

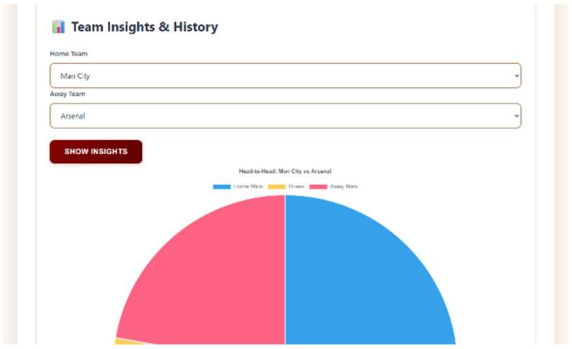


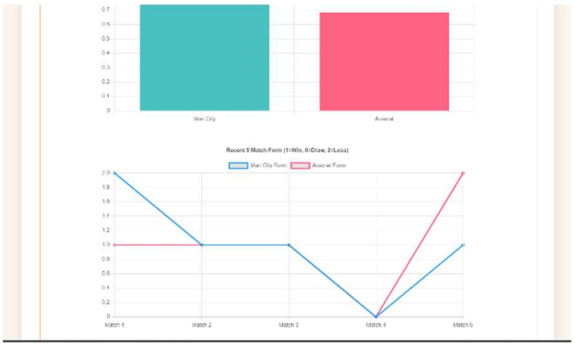
In conclusion, the Random Forest Classifier was selected as the final model for deployment within the Flask-based web application, owing to its high accuracy, interpretability through feature importance visualization, and robustness against noise.

Result


FOOTBALL TODAY
EPL
Predict
History
Dark
Premier League Gallery
Premier League
HIGHLIGHTS
3-2
Man City vs Arsenal – Key Moments
Liverpool Celebrates Another Win
Spurs Push for Top 4
Premier League Headlines


FOOTBALL TODAY
EPL
Predict
History
Dark
Match Prediction
Home Team
Man City
Away Team
Arsenal
Bet365 Odds
3.2
1.2
5.6
Form Scores
4
5
Referee
M Oliver
Match Time
18:00:00
PREDICT
Prediction Result
Arsenal Wins
Percentage Probability:
Man City 27.0%
Arsenal 46.0%
Draw 27.0%


Team Insights & History
Home Team
Man City
Away Team
Arsenal
SHOW INSIGHTS



Conclusion and Future Scope

This project successfully developed a machine learning-based football match prediction system that integrates team statistics, player attributes, and betting odds to forecast match outcomes. By leveraging historical match data from multiple sources, extensive preprocessing, and feature engineering, the model was trained using various algorithms such as Random Forest Classifier, Support Vector Machine (SVM), and Logistic Regression. Among these, the Random Forest Classifier demonstrated the best performance with an accuracy of around 71%, effectively capturing the non-linear relationships between features such as team form, Elo ratings, referee influence, and betting odds. The results confirmed that data-driven models can provide realistic and insightful predictions aligned with expert opinions and statistical trends in football analytics.

The developed Flask-based web application enables real-time predictions by allowing users to input match details such as team names, form scores, and odds. The system further provides visual insights, including head-to-head statistics, recent form charts, and rating comparisons, helping users make informed decisions. This integration of data analytics and web technology bridges the gap between predictive modeling and practical usability, offering a modern, interactive tool for football enthusiasts, analysts, and bettors.

In the future, the project can be expanded by incorporating player-level data, live match updates, and advanced deep learning models such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks to capture temporal dependencies across seasons. Enhancing the model with real-time API integrations for injury reports, weather conditions, and tactical formations could further improve prediction accuracy. Additionally, using explainable AI (XAI) techniques would make the prediction process more transparent and trustworthy for end-users.

Overall, this project establishes a strong foundation for sports analytics using machine learning, demonstrating the potential of data-driven approaches to revolutionize performance evaluation, betting strategies, and tactical decision-making in modern football.

References

[1] Rodrigues, F., & Pinto, Â. (2022). Prediction of football match results with Machine Learning. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2013, pp. 1222–1230.

[2] Y. Wu, "**Improving Rolling Prediction with RFC**," *Journal of Trends in Physics, Engineering and Science (JTPES)*, vol. 4, no. 05, 2024. DOI: https://doi.org/10.53469/jtpes.2024.04(05).07.

[3] J. Gudmundsson and M. Horton, "Spatio-temporal analysis of team sports — A survey," *CoRR*, vol. abs/1602.06994, 2016. [Online]. Available: https://arxiv.org/abs/1602.06994

[4] Kaggle. "European Soccer Database." [Online]. Available: https://www.kaggle.com/datasets/hugomathien/soccer

[5] API-Football. "Football Data API Documentation." [Online]. Available: https://www.api-football.com/