

AI-Powered Football Match Prediction and Analytics Dashboard

Hifzaan Siraj Mukadam

Department of Artificial intelligence and data science

Anjuman-I-Islam Abdul Razak Kalsekar Polytechnic Campus

New Panvel Navi Mumbai Maharashtra, India

Guide : Prof. Ali Karim Sayed

Abstract - Football analytics has become an important application area for machine learning, enabling data-driven decision-making for coaches, analysts, and fans. This paper presents an AI-Powered Football Match Prediction and Analytics Dashboard that combines a supervised machine learning model with an interactive web-based dashboard. The system uses historical match data to engineer pre-match rolling-form features and trains a RandomForest classifier to predict match outcomes as Home Win, Draw, or Away Win, achieving reasonable accuracy on the evaluation dataset. Beyond prediction, the platform provides a Monte Carlo match simulator, an explainable AI module that reports feature importance and confusion matrices, a rule-based AI Coach Assistant that recommends formations and tactics with reasoning, and a league-wide analytics dashboard covering team statistics, standings, and comparisons. Additional tools such as a Dream XI generator, transfer recommendation engine, opponent weakness detector, momentum tracker, and match difficulty rating extend the platform's practical utility. The system is built using Python, scikit-learn, Streamlit, and Plotly, and is designed to be modular, extensible, and easy to deploy for both academic and practical use.

Key Words: Football Analytics, Machine Learning, Random Forest, Match Prediction, Streamlit Dashboard, Explainable AI, Monte Carlo Simulation.

1. INTRODUCTION

Football is the world's most popular sport, generating vast amounts of match data every season. Traditionally, match analysis and outcome prediction relied heavily on expert judgment and simple statistical summaries. With the growth of machine learning, it has become possible to systematically learn patterns from historical match data and generate data-driven predictions and tactical insights. The objective of this project is to design and

implement a complete, end-to-end football analytics platform that predicts match outcomes using machine learning while also providing rich analytics, simulation, and coaching tools through an interactive dashboard. The system is intended to serve as both an educational demonstration of applied machine learning and a practical decision-support tool for football enthusiasts, students, and analysts.

2. LITERATURE REVIEW

Statistical modelling of football match outcomes has a long history. Dixon and Coles proposed one of the earliest widely-used approaches, modelling goals scored by each team as independent Poisson-distributed random variables calibrated from historical attacking and defensive strength, an approach that still underlies many modern match simulators, including the one used in this project.

More recent work has shifted toward supervised machine learning classifiers such as Random Forests, Support Vector Machines, and Gradient Boosting to directly predict match outcomes (Home Win, Draw, Away Win) from engineered features such as recent form, head-to-head history, and squad strength. Random Forests, introduced by Breiman, remain popular for tabular sports data due to their robustness to noisy features and their ability to report feature importance for interpretability.

Existing football analytics tools are often either purely statistical dashboards with no predictive capability, or black-box prediction APIs with no explanation or tactical guidance. This project addresses that gap by combining a transparent, explainable machine learning model with simulation, analytics, and a rule-based tactical assistant in a single, self-contained application.

3. SYSTEM ARCHITECTURE AND METHODOLOGY

The system follows a modular architecture that separates data handling, model training, prediction, analytics, simulation, and the user interface into independent Python modules. This design keeps the machine learning pipeline reusable and testable independently of the Streamlit front end.

The data pipeline (utils.py) automatically detects and maps differently-named columns to a canonical schema, cleans missing or invalid records, and engineers pre-match rolling-form features for each team so that the model never has access to information from the match being predicted, avoiding data leakage. train_model.py trains and evaluates the RandomForest classifier; predict.py loads the trained artifacts to serve predictions; analytics.py computes team and league statistics; simulation.py runs Monte Carlo simulations; coach.py implements the rule-based tactical assistant; and app.py ties every module together into the Streamlit dashboard.

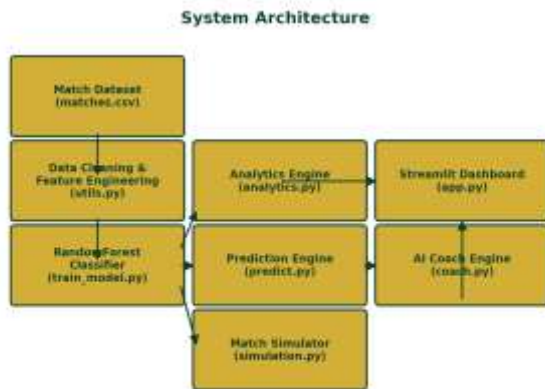


Fig -1: System Architecture of the Proposed Platform

4. MACHINE LEARNING MODEL

The prediction target is a three-class variable: Home Win, Draw, or Away Win, derived directly from the final score of each historical match. The feature set includes each team's win rate, average goals scored, and average goals conceded over its last five matches, along with encoded team and league identifiers. A RandomForestClassifier from scikit-learn (300 estimators, balanced class weights) was selected for its robustness on tabular data and its built-in feature importance reporting, which supports the explainability goals of the project.

The dataset was split into training (80%) and testing (20%) sets using stratified sampling to preserve class balance. Model performance was evaluated using Accuracy, Precision, Recall, and F1 Score (weighted average across classes), and a confusion matrix was generated to inspect class-wise performance in detail.

5. RESULTS AND DISCUSSION

Table -1: Model Evaluation Metrics

Metric	Value
Accuracy	47.3%
Precision	45.3%
Recall	47.3%
F1 Score	45.9%

The trained model achieved an accuracy of 47.3% on the held-out test set, noticeably above the 33.3% baseline expected from random guessing across three balanced classes. Since the bundled dataset is synthetically generated with injected randomness for demonstration purposes, real-world datasets with richer features (player-level data, injuries, head-to-head history) are expected to yield higher predictive accuracy. Figure 2 shows the confusion matrix produced during evaluation.

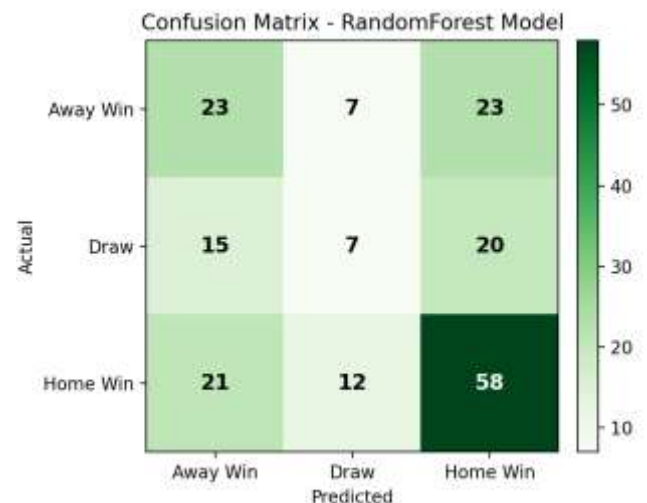


Fig -2: Confusion Matrix of the RandomForest Classifier

Figure 3 shows the relative feature importance reported by the trained model. Recent team form (win rate over the last five matches) and average goals scored/conceded were consistently the most influential features, confirming that recent momentum is a strong predictor of match outcome, consistent with footballing intuition and prior literature.

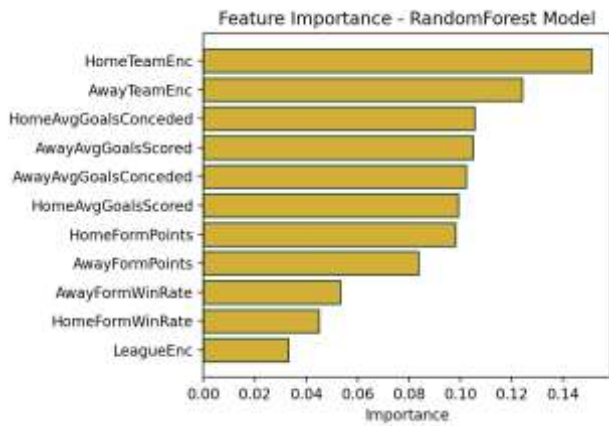


Fig -3: Feature Importance from the RandomForest Model

6. USER INTERFACE AND DASHBOARD DESIGN



Fig -4: Home Page of the Streamlit Dashboard

The Home page (Figure 4) presents a dataset overview with key metrics such as total matches, teams, and leagues covered, along with quick navigation to every module. The Match Prediction page (Figure 5) lets the user select a home and away team and instantly view the predicted outcome along with win, draw, and loss probabilities, a bar chart of the probability distribution, and a plain-English explanation of why the model made that prediction, directly supporting the explainable AI goals of the project.

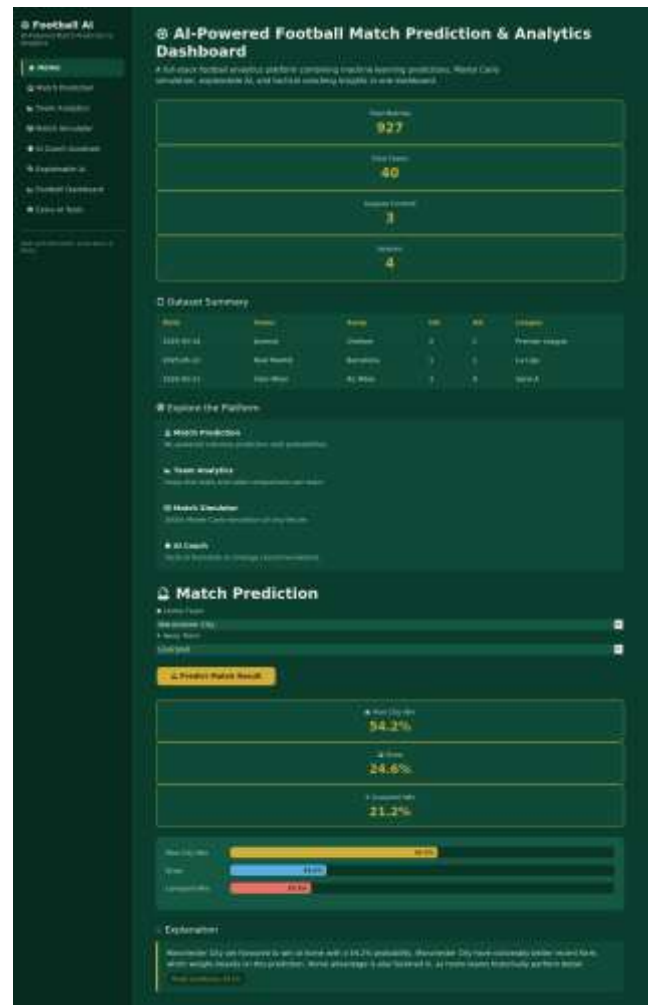


Fig -5: Match Prediction Interface with Probability Distribution

7. MATCH SIMULATION AND EXTRA AI FEATURES

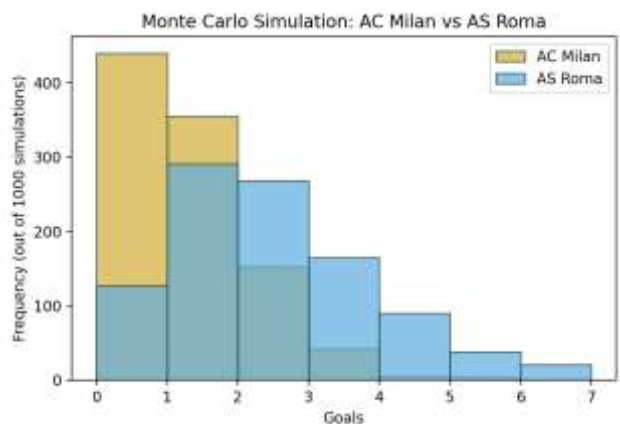


Fig -6: Monte Carlo Match Simulation Results (1000 Simulations)

The Match Simulator runs a fixture 1000 times using Poisson-distributed goal counts calibrated from each team's historical attacking and defensive strength, reporting win/draw/loss percentages, the most likely scoreline, and a full probability distribution, as shown in Figure 6. The AI Coach Assistant is a transparent, rule-based engine that takes team form, opponent form, and home/away status as input and recommends a formation, tactical style, pressing strategy, defensive

and attacking advice, and set-piece suggestions, explicitly explaining the reasoning behind every recommendation so it remains fully interpretable to coaches and students alike.

Several additional analytical tools extend the platform's practical value: a Momentum Tracker that scores a team's recent run of results from -100 to +100; a Match Difficulty Rating that scores an upcoming fixture from 1 (easy) to 10 (very hard) based on relative team strength; an Opponent Weakness Detector that highlights statistical weaknesses such as a leaky defense or poor away form; a Dream XI Generator that suggests a formation and lineup based on a team's attacking/defensive profile; and a Transfer Recommendation Engine that identifies the type of player signing (defensive reinforcement, attacking depth, etc.) a squad most needs based on its statistical profile.

8. CONCLUSION AND FUTURE SCOPE

This project successfully demonstrates a complete, modular football analytics platform that unifies machine learning prediction, explainable AI, Monte Carlo simulation, and rule-based tactical coaching in a single interactive dashboard. The RandomForest model achieved meaningful predictive performance while remaining fully interpretable through feature importance and confusion matrix analysis. Future work includes integrating real-time fixture and odds data via a sports API, incorporating player-level statistics for a fully data-driven Dream XI generator, experimenting with gradient boosting models such as XGBoost, adding a persistent database backend for storing prediction history, and deploying the platform publicly for broader use by football analysts and enthusiasts.

ACKNOWLEDGEMENT

The authors would like to thank their project guide and the faculty of the Department of Computer Engineering for their continuous guidance and support throughout the development of this project.

REFERENCES

1. Breiman, L.: Random Forests. Machine Learning, Vol. 45, No. 1, 5-32 (2001)
Springer-VerlagBerlinHeidelberg New York
2. Dixon, M.J., Coles, S.G.: Modelling Association Football Scores and Inefficiencies in the Football Betting Market. Journal of the Royal Statistical Society: Series C, Vol. 46, No. 2, 265-280 (1997)
Springer-VerlagBerlinHeidelberg New York
3. Pedregosa, F. et al.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, Vol. 12, 2825-2830 (2011)

Springer-VerlagBerlinHeidelberg New York4. Streamlit Inc.: Streamlit Documentation. <https://docs.streamlit.io> (2024)