FIFA Match Predictor: Technical Report

1. Introduction

The FIFA Match Predictor is a machine learning–based model designed to predict the outcomes of international football matches. The project aims to leverage historical match data, team statistics, and FIFA rankings to forecast the likelihood of a team winning a match.

The model focuses on team-level data and historical patterns to generate predictions for matches involving the top 48 teams expected to participate in the FIFA 2026 World Cup.

---

2. Data Sources

The predictor was built using multiple datasets covering different aspects of football performance:

1. `matches_1930_2022.csv` – Contains results of international matches from 1930 to 2022, including goals scored, match outcomes, and basic metadata such as tournament and date.

2. `world_cup.csv` – Records details of FIFA World Cup tournaments across years.

3. `fifa_ranking_2022-10-06.csv` – Provides the official FIFA ranking points and positions of teams as of October 2022.

4. `transfermarkt_team_stats_all-time.csv` – Includes long-term team statistics like total goals, wins, and losses across all competitions. This part of the data was **scraped** using a custom scraping tool built by me.

These files were cleaned, standardized, and combined into modeling datasets containing essential features such as team ranking differences, expected goals, attendance, and match types.

---

3. Data Cleaning and Preparation

Each dataset underwent a multi-step cleaning process:

- **Standardization:** All country names were standardized to ensure consistency across datasets (e.g., “USA” and “United States” were unified).
- **Handling Missing Data:** Rows with missing or incomplete country names or scores were removed.
- **Feature Engineering:** Additional columns were created to capture:
    - Rank difference between teams
    - Points difference from FIFA rankings
    - Expected goals (xG) estimates
    - Match context indicators (home, away, or neutral)
    - Knockout vs. group-stage identifiers

The cleaned datasets were saved in modular form for future retraining or expansion.

---

4. Model Selection

After experimentation with several algorithms, the **Random Forest Classifier** was selected as the primary prediction model due to its robustness, interpretability, and good baseline performance without extensive hyperparameter tuning.

The model was trained using historical match features, with the target variable representing the match outcome (win or non-win for the home team).

A standard data-splitting approach was used:

- 80% training data
- 20% test data

The model achieved consistent accuracy across multiple runs and handled non-linear feature relationships effectively.

---

5. Model Limitations and Considerations

a. Lack of Player-Level Data

While team-level statistics provide a general overview of performance, **player-specific data** such as injuries, form, individual ratings, or lineup strength significantly influence match outcomes.
Due to insufficient and inconsistent player-level data across years, these features were **excluded** from this version of the model. As a result, the predictor may underrepresent the real-time dynamics of team strength during tournaments.

b. Model Choice: Random Forest vs. XGBoost

Although Random Forest provided a solid baseline, there are notable drawbacks compared to **XGBoost**, which is a more advanced gradient boosting algorithm:

Aspect	Random Forest	XGBoost
Bias-Variance Tradeoff	Can produce slightly higher bias	More balanced and lower bias
Handling of Imbalanced Data	Less efficient without class weighting	Built-in optimization for imbalance
Computation	Faster to train on smaller datasets	Optimized for large-scale, distributed computation
Performance Tuning	Fewer hyperparameters; simpler to implement	Highly tunable and typically yields higher accuracy
Feature Importance	Uses mean decrease impurity (less reliable)	Provides more consistent gain-based importance metrics

The decision to use Random Forest was primarily guided by its **speed, simplicity, and reduced risk of overfitting** during early iterations. Future versions of the predictor could migrate to XGBoost or even ensemble both models for enhanced accuracy and stability.

C. Only 28 of 48 teams have been qualified so far

To handle this, we used the 28 teams that have already qualified and use our algorithm to predict the next 20 teams that will qualify

---

6. Model Deployment

A Streamlit web interface was developed to allow users to select two teams and view the predicted winner.
The interface:

- Restricts predictions to the **top 48 teams** expected for FIFA 2026.
- Displays predicted winner and confidence score.
- Uses hidden default feature values for unmodeled factors such as xG.
- Implements a **dark-themed UI** for a cleaner visual presentation.

The trained Random Forest model and data scaler were serialized using `joblib` to allow for quick loading and inference without retraining.

---

7. Results Interpretation

For each match prediction:

- The model outputs probabilities for a **home win** versus **non-home win** (draw or away win).
- The team with the higher probability is identified as the predicted winner.
- Confidence percentages are displayed to give an approximate measure of certainty based on historical data similarity.

While the predictions are consistent with general trends in international football, the absence of real-time player data, tactical shifts, or situational variables limits full predictive accuracy.

---

8. Conclusion

The FIFA Match Predictor demonstrates how machine learning can be applied to sports analytics using historical and team-level data.
Although the model captures essential patterns from over 90 years of football history, incorporating **player-level attributes**, **recent form data**, and a **more advanced algorithm such as XGBoost** would significantly improve its predictive capability.

Future work should focus on:

- Integrating up-to-date player performance metrics.

- Testing ensemble models combining Random Forest and XGBoost.
- Expanding the dataset to include newer matches and friendlies before FIFA 2026.

---

9. References

- FIFA Official Rankings Dataset
- Transfermarkt Historical Team Statistics
- Kaggle: International Football Results from 1872 to 2022
- Streamlit Documentation
- Scikit-learn Machine Learning Library
