FIFA 2026 FINALIST PREDICTION - FINAL REPORT

Student: Naveen Kumar B N

Register Number: 24UG00321

Course: B.Tech CSE (AI)

University: Chanakya University

1. Project Overview

The FIFA 2026 Finalist Prediction project is an end-to-end machine learning system designed to predict which teams are most likely to reach and win the FIFA World Cup 2026. The project follows a four-week workflow:

• Week 1 – Data Collection and Feature Engineering

• Week 2 – Model Building and Evaluation

• Week 3 – Model Optimization and Final Prediction

• Week 4 – Final Prediction and Flask Application Deployment

The project integrates historical FIFA World Cup match data (1930–2022) with FIFA world rankings (1992–2024). Machine learning models were developed, optimized, and deployed through a Flask web application for interactive predictions.

2. Week 1 – Data Collection and Feature Engineering

Week 1 focused on building a clean and structured dataset combining World Cup match data and FIFA ranking data. The datasets used were 'WorldCupMatches (1).csv' and 'fifa_ranking-2024-06-20.csv'. Both datasets were cleaned, preprocessed, and merged using 'Year' and 'Team' as common keys.

Key Steps:

• Selected relevant columns such as Year, Stage, Teams, Goals, and Win Conditions.

• Removed missing team entries and renamed columns for consistency.

• Determined winners and identified finalist teams.

• Computed yearly statistics like Goals_For, Goals_Against, Goal_Difference, Win_Rate, and Matches_Played.

• Merged with FIFA ranking data to include FIFA_Rank, FIFA_Points, and Confederation details.

• Added a binary target variable 'Is_Finalist' (1 = reached final, 0 = not).

The cleaned dataset 'fifa_1930_2022_with_rank.csv' contained approximately 1000 records and 11 columns, covering team performance and ranking data from 1930 to 2022.

3. Week 2 – Model Building and Evaluation

Week 2 involved developing predictive models using Logistic Regression and Random Forest Classifier. Data was split into 80% training and 20% testing sets with standardized scaling applied to features.

Performance Metrics:

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	0.85	0.80	0.76	0.78	0.86
Random Forest	0.91	0.88	0.85	0.86	0.93

Random Forest outperformed Logistic Regression, achieving higher accuracy and ROC-AUC scores. Feature importance analysis revealed FIFA_Rank, Goal_Difference, and FIFA_Points as top predictors.

4. Week 3 – Model Optimization and Final Prediction

Week 3 emphasized optimizing the models using GridSearchCV for hyperparameter tuning and 10-fold cross-validation. The Random Forest and Gradient Boosting models were compared based on multiple evaluation metrics.

Optimized Model Performance:

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Optimized Random Forest	0.87	0.87	0.87	0.87	0.87
Gradient Boosting	0.74	0.82	0.80	0.80	0.82

Feature importance analysis ranked FIFA_Rank, Goal_Difference, and FIFA_Points as the top influencing features. The optimized Random Forest model proved to be the most accurate and stable.

5. Week 4 – Final Prediction and Deployment

In the final phase, both Random Forest and Gradient Boosting models were combined using ensemble averaging. Predictions were made to determine semifinalists, finalists, and the probable winner.

Predicted Results:

• Top 4 Semifinalists: Argentina, Germany, Brazil, France

• Predicted Final Match: Germany vs Argentina

• Predicted Winner: Argentina

The results were visualized through probability comparison charts and bar graphs showing semifinalist performance. Argentina was consistently identified as the team with the highest ensemble probability of winning the FIFA 2026 World Cup.

6. Flask Application Integration

A Flask web application ('app.py') was developed to deploy the trained models. The app allows users to select two teams and a location to predict match outcomes interactively. The backend computes probabilities using the ensemble model and adjusts results based on host country and confederation advantages.

Key Features of the Flask App:
• Loads merged dataset and trains Random Forest and Gradient Boosting models.
• Accepts user input (Team A, Team B, Host Location).
• Computes win probabilities and identifies the most probable winner.
• Provides a clean and user-friendly interface using index.html.

7. Conclusion

The FIFA 2026 Finalist Prediction project successfully demonstrates how machine learning can be used to analyze historical sports data and make reliable tournament forecasts. Each stage of the project—data preparation, model development, optimization, and deployment—was completed systematically.

Final Outcome: Argentina is predicted as the most probable winner of the FIFA 2026 World Cup. The ensemble model, combined with web deployment, provides an interactive, data-driven approach to sports prediction.