

# Predicting FIFA World Cup 2026 Finalists Using Machine Learning

**Aditya Padwalkar** | ID: 24UG00304

Section B | Computer Science & Artificial Intelligence

College Engineering Project

# Introduction

This project leverages historical World Cup data and machine learning models to predict the finalists of the FIFA World Cup 2026. By analyzing patterns in team performance metrics—including win rates, goal differences, and international rankings—we build predictive models capable of identifying which teams are most likely to advance to the final match.

**Goal:** Develop accurate predictive models and deploy them in an interactive web application for real-time analysis.



# Project Objectives

## Predict Finalists

Identify the two teams most likely to reach the World Cup 2026 final match using historical data patterns.

## Apply ML Models

Implement and compare Logistic Regression and Random Forest classifiers for binary classification tasks.

## Interpret Results

Analyze feature importance to understand which performance metrics drive finalist predictions.

## Deploy Application

Build an interactive Streamlit app enabling users to input team data and receive real-time predictions.

# Data Collection & Preprocessing

01

## Data Sources

Integrated three datasets: historical match records, World Cup tournament data, and FIFA team rankings. Combined data spans multiple tournament cycles for comprehensive analysis.

02

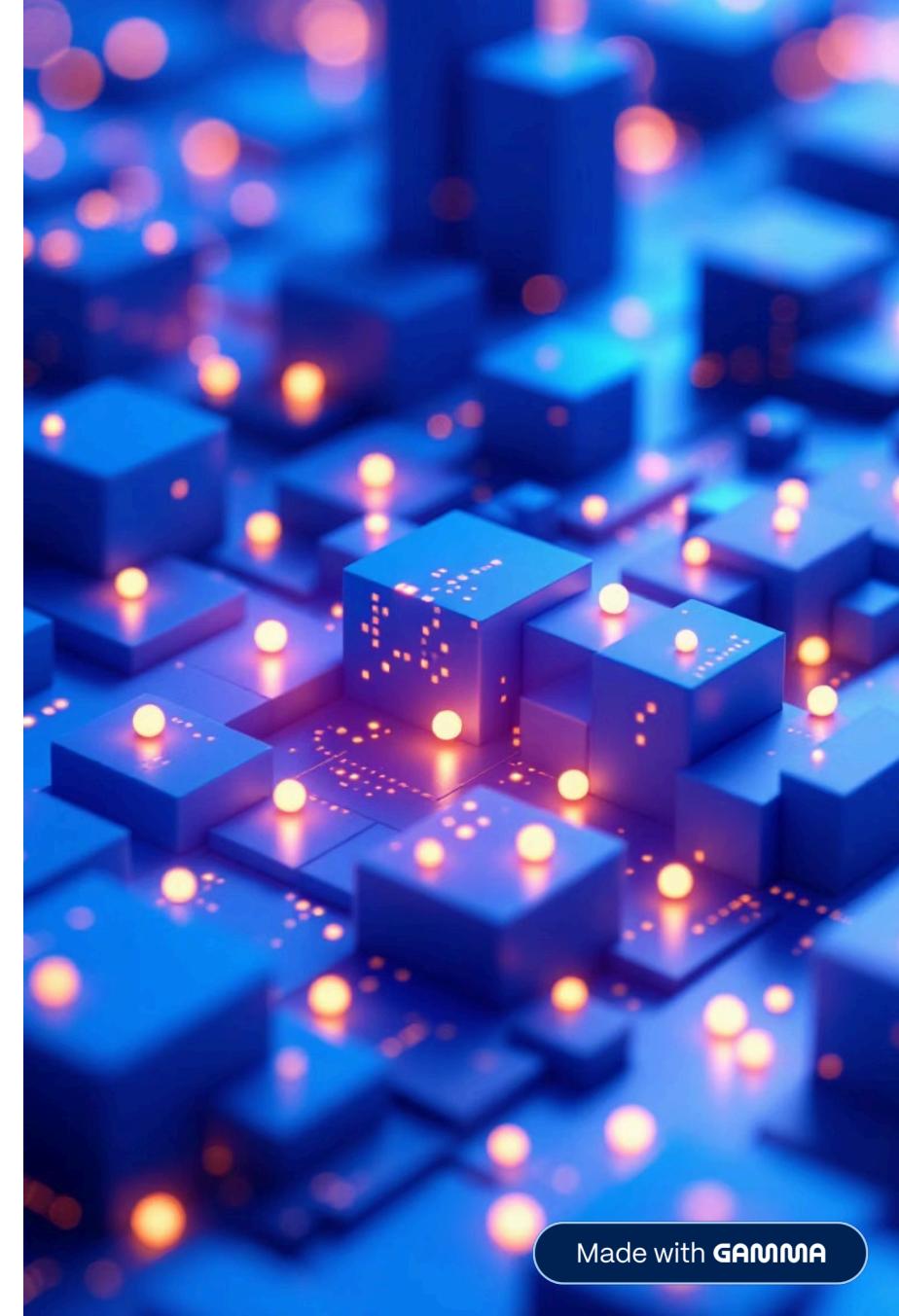
## Cleaning Process

Removed duplicate entries and standardized team names across datasets. Extracted tournament years and aligned temporal data to ensure consistency.

03

## Quality Assurance

Verified data integrity, handled missing values, and validated team rankings. Ensured all records aligned with official FIFA records and tournament dates.



# Feature Engineering

Four engineered features capture team performance and competitive strength:

Feature	Definition & Significance
Win Rate	Percentage of matches won by a team. Higher win rates indicate consistent performance and competitive strength.
Goal Diff Avg	Average goal differential (goals scored minus conceded) per match. Reflects offensive dominance and defensive stability.
FIFA Rank	Official international ranking score. Captures global competitive standing and recent performance trends.
Participations	Number of World Cup tournaments a team has participated in. Reflects tournament experience and institutional knowledge.

These features were engineered to capture multidimensional aspects of team performance: form (win rate, goal differential), current status (FIFA rank), and experience (tournament participations).



# Model Building Strategy

## Logistic Regression

Probabilistic classifier optimized for binary classification. Provides interpretable probability estimates and clear decision boundaries.

## Random Forest

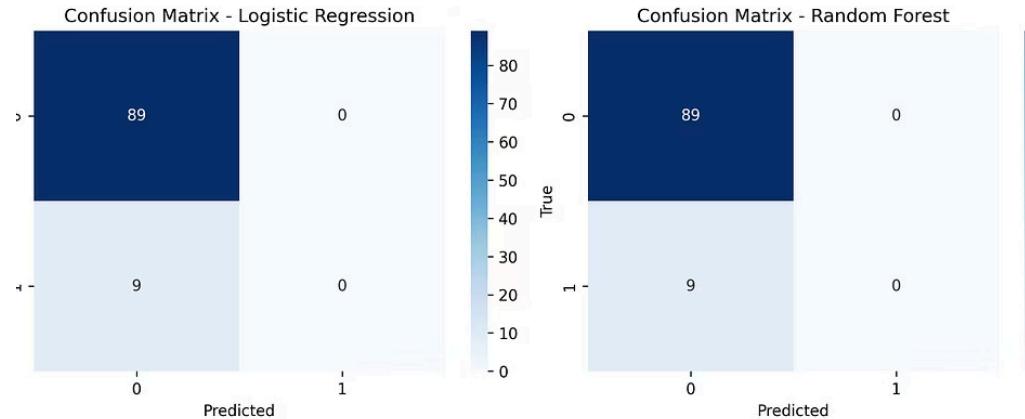
Ensemble method combining multiple decision trees. Captures non-linear relationships and feature interactions in complex datasets.

## Optimization

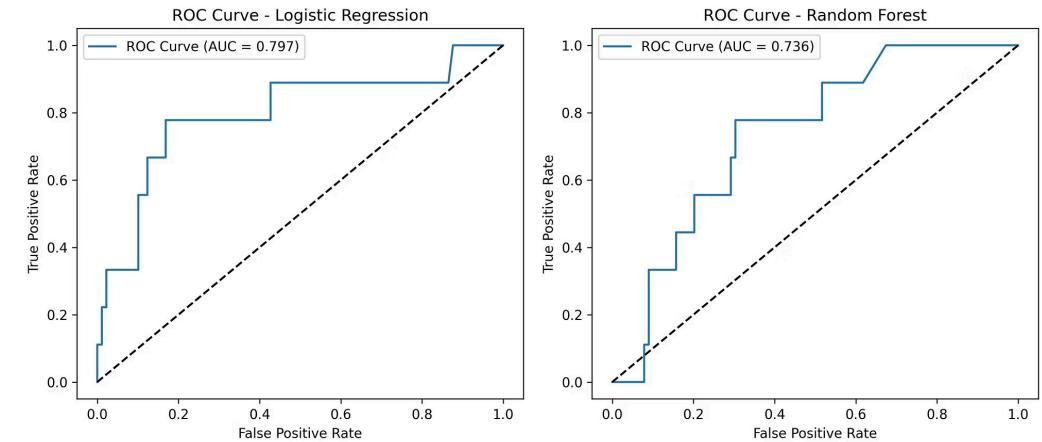
GridSearchCV performed hyperparameter tuning. StandardScaler normalized features to equalize feature contribution.

# Model Performance Evaluation

## Confusion Matrices



## ROC Curves Comparison



AUC Scores: Logistic Regression: 0.797 | Random Forest: 0.736



# Feature Importance Analysis

## Goal Differential Average (Highest Impact)

Most influential predictor of finalist qualification. Teams with higher goal differential margins demonstrate superior match control and offensive efficiency.

## Win Rate (Strong Secondary Factor)

Second-most important feature. Consistent winning performance strongly correlates with finalist advancement.

## Ranking & Experience (Supporting Factors)

FIFA ranking and tournament participation provide context but carry less predictive weight than direct performance metrics.



# Final Predictions: 2026 World Cup Finalists

1

Predicted Winner

France

2

Predicted Runner-Up

Spain

Top 10 finalist probability rankings determined by Logistic Regression model (AUC = 0.797). Predictions based on historical performance patterns and engineered feature analysis.

# Conclusion & Future Work

## Key Findings

- Logistic Regression outperformed Random Forest with AUC of 0.797
- Goal differential and win rate are primary prediction drivers
- Historical data patterns effectively predict tournament outcomes
- France and Spain emerge as 2026 finalist predictions

## Future Enhancements

- Address class imbalance using SMOTE techniques
- Integrate player-level performance statistics
- Implement ensemble and deep learning models
- Expand real-time data collection and model retraining
- Enhance Streamlit UI with advanced visualizations