



FIFA WORLD CUP 2026

PREDICTION ANALYSIS

Giggling Platypus Co.

27 December, 2025

OBJECTIVE:

- Analyze FIFA rankings to predict top-performing teams for World Cup 2026.
- Compare Logistic Regression and Random Forest performance.
- Visualize insights from FIFA data.
- Provide predictions for Top 10 and Top 2 finalists.





DATA SOURCES

USING DATA SOURCES

- FIFA RANKINGS SCRAPER: LATEST FIFA MEN'S WORLD RANKINGS.
- WORLD CUP 2026 TEAMS: 48 QUALIFIED TEAMS.
- HISTORICAL MATCHES DATASET: PREVIOUS WORLD CUPS AND OTHER TOURNAMENTS.
- FEATURE ENGINEERED DATASET: COMBINED AND PROCESSED FEATURES FOR MODELING.



Features Used

- Rank: FIFA rank of the team.
- Points: FIFA points based on performance.
- Source Code: Encoded category – Qualified / Probable / Added.
- Derived Features: Top_Team (1 if team is in top 10), Rank_Category (Strong / Weak).



EDA:

- Top 10 Teams Among Final 48: Bar chart of points.
- Correlation Analysis: Scatter plot showing Rank vs Points – negative correlation (higher rank → lower points).

DATA PREPROCESSING:

- Cleaned team names (capitalization, stripping spaces).
- Converted numeric columns (Rank, Points).
- Merged datasets to align rankings with qualified teams.
- Handled missing values in Points & Rank.



FEATURE SELECTION

Feature Selection

- Used Logistic Regression coefficients to determine importance:
 - Points → most influential.
 - Rank → also important.
- Random Forest also gives feature importance.
- Helps understand which features contribute most to prediction.



Modeling

- Logistic Regression: Predict probability of being top-performing.
- Random Forest: Ensemble method to predict top teams.
- Both models trained on Rank, Points, Source_Code.
- Output: Predicted top 10 teams and top 2 finalists.

MODEL EVALUATION:



- Metrics Used:
- Accuracy
- ROC-AUC score
- Confusion Matrix
- Classification Report
- ROC Curve Interpretation: Measures model's ability to distinguish top teams from others.
- AUC Score: Higher → better distinction.

A photograph of a soccer coach in a dark blue tracksuit with white stripes on the sleeves, pointing his right index finger directly at the viewer. He has short dark hair and a serious expression. In the foreground, the back of another person's head and shoulders are visible, wearing a red and white striped shirt.

PREDICTIONS:

Predictions

- Top 10 Teams (Example Table or Chart): Display predicted top 10 by points.
- Top 2 Finalists: Teams predicted with highest probability by models.

Insights & Interpretation

- Teams with higher FIFA points more likely to be top performers.
- Rank inversely related to performance (lower rank → stronger team).
- Random Forest captures non-linear patterns; Logistic Regression captures linear relationships.

CONCLUSIONS:

Conclusion

- Successfully predicted top teams using FIFA data and ML models.
- Logistic Regression and Random Forest both provide insights but with different perspectives.
- Feature importance shows which factors influence team success.
- Visualization helps in understanding trends and ranking distributions.



TOOLS AND TECHNOLOGIES:



Tools & Technologies

- Python: Pandas, Matplotlib, Seaborn, Scikit-learn, Joblib.
- Data Scraping: Custom web scraper for FIFA rankings.
- Visualization: Bar charts, scatter plots, ROC curve, confusion matrix.

