WEEK 4 – FINAL REPORT
FIFA 2026 FINALISTS PREDICTION

Name: Harshal Gowda CM

Roll Number: 24UG00313

• Overview

This project aimed to predict the FIFA 2026 World Cup finalists using Artificial Intelligence techniques. Throughout four weeks, the project evolved from raw data preprocessing to a fully functional prediction system with a user-friendly Flask web interface and a Command-Line Interface (CLI). The project integrates machine learning models, data visualization, and simulation techniques to estimate team performance probabilities and predict likely finalists and winners.

• Methodology Summary

○ Data Preparation

In Week 1, the dataset containing FIFA team statistics such as FIFA Ranking, Average , Win Percentage, Goal Difference, Experience, and Host status was collected and cleaned. The data was formatted into a structured Excel sheet with 48 teams and 8 relevant attributes. Missing values were handled and categorical data normalized for model readiness.

• Model Building and Evaluation

During Week 2 and Week 3, two primary models were trained and compared: Logistic Regression and Random Forest. The models were evaluated based on Accuracy, Precision, Recall, F1-score, and ROC-AUC metrics. The Random Forest model outperformed Logistic Regression, demonstrating stronger generalization and non-linear learning capabilities. ROC curves and Confusion Matrices were plotted to visually analyze the classification performance.

• Model Deployment and Interface Development

In Week 4, the project was extended to include an interactive Flask web application and a CLI tool. The Flask app allows users to upload CSV datasets, select models, and visualize top predicted finalists along with probability charts. Additionally, a CLI interface was implemented to provide the same predictive functionality directly via command line.

• Findings and Interpretation

Both models achieved high accuracy, with the Random Forest model demonstrating superior predictive ability. Feature importance analysis revealed that Goal Difference, Win

Percentage, and FIFA Ranking were the most influential factors in determining team success probability. The models accurately ranked top-performing teams such as Spain, Brazil, Argentina and France among likely finalists.

- ## Advanced Simulation Techniques

Two simulation approaches were integrated to enhance realism:
- Monte Carlo Simulation: Performed 1000+ randomized trials to estimate the probability distribution of finalists.
- Poisson Goal Simulation: Used Poisson distributions to simulate goal counts between the two top finalists, offering a probabilistic winner prediction.

These techniques provide not only classification outputs but also dynamic match outcome probabilities, improving interpretability.

- ## Flask and CLI Implementation

The Flask interface provides an intuitive web-based dashboard where users can upload team data, choose models, and view the top predicted teams through interactive tables and graphs. The CLI version supports command-line prediction for automation or offline testing.

Both interfaces enhance the system's accessibility and demonstrate real-world deployment readiness.

- ## Conclusion and Future Scope

The FIFA 2026 Finalists Prediction project successfully demonstrates how Artificial Intelligence can be applied to sports analytics. By integrating machine learning with simulation and web technologies, it offers a predictive, explainable, and deployable solution. Future enhancements may include incorporating player-level statistics, real-time updates, and advanced ensemble models like XGBoost or Neural Networks to improve accuracy further.

- ## References

1. FIFA Official Statistics Dataset (2025)
2. Scikit-learn Documentation – Machine Learning Models
3. Flask Framework Official Documentation
4. Kaggle – International Football Results Dataset

TOP 4 TEAMS TO FINAL

TEAMS	Predicted Score
Spain	0.9243
Brazil	0.9237
Argentina	0.7898
France.	0.7893

TOP 8 TEAMS TO FINAL

Spain	0.9243
Brazil	0.9237
Argentina	0.7898
France	0.7893
Netherlands	0.7765
Germany	0.7594
England	0.7590
Italy	0.7419

Tournament-style Simulation

- If we take the top 8 as quarter-finalists, the semi-final and final projection becomes:

Quarterfinals (by ranks)

- Spain vs Italy → Spain (0.924 > 0.742)
- Brazil vs England → Brazil (0.924 > 0.759)
- Argentina vs Germany → Argentina (0.790 > 0.759)
- France vs Netherlands → France (0.789 > 0.777)

Semifinals

- Spain vs France → Spain (0.924 > 0.789)
- Brazil vs Argentina → Brazil (0.924 > 0.790)
- Semi-finalists: Argentina , France

Final

- Spain vs Brazil → Spain (0.9243 > 0.9237)

Champion Prediction: SPAIN
Runner-up: BRAZIL

Statistical Analysis of All Scores

Metric	Value
Mean score	0.528
Median score	0.512
Standard deviation	0.184
Minimum score	0.0175 (New Zealand)
Maximum score	0.9243 (Spain)
Range	0.9068
Top 8 avg	0.808
Bottom 8 avg	0.248

- **Logistic Regression** → helps *understand why* a team is predicted to win.
- **Random Forest** → helps *accurately predict which* team will win.

Evaluation metrics (what & why)

- **Accuracy** — overall percent correct; quick check (bad alone if classes imbalanced).
- **Precision** — of teams predicted as "finalist", how many were correct (avoids false hype).
- **Recall** — of actual finalists, how many the model found (important — don't miss real finalists).
- **F1 score** — harmonic mean of precision & recall; single balanced metric for model selection.
- **ROC-AUC** — how well model ranks finalists vs non-finalists across thresholds (threshold-independent).
- **Confusion matrix** — counts TP/FP/FN/TN; shows exact error types to guide tuning.
- **ROC curve** — visual trade of True Positive vs False Positive rates; helps pick threshold.

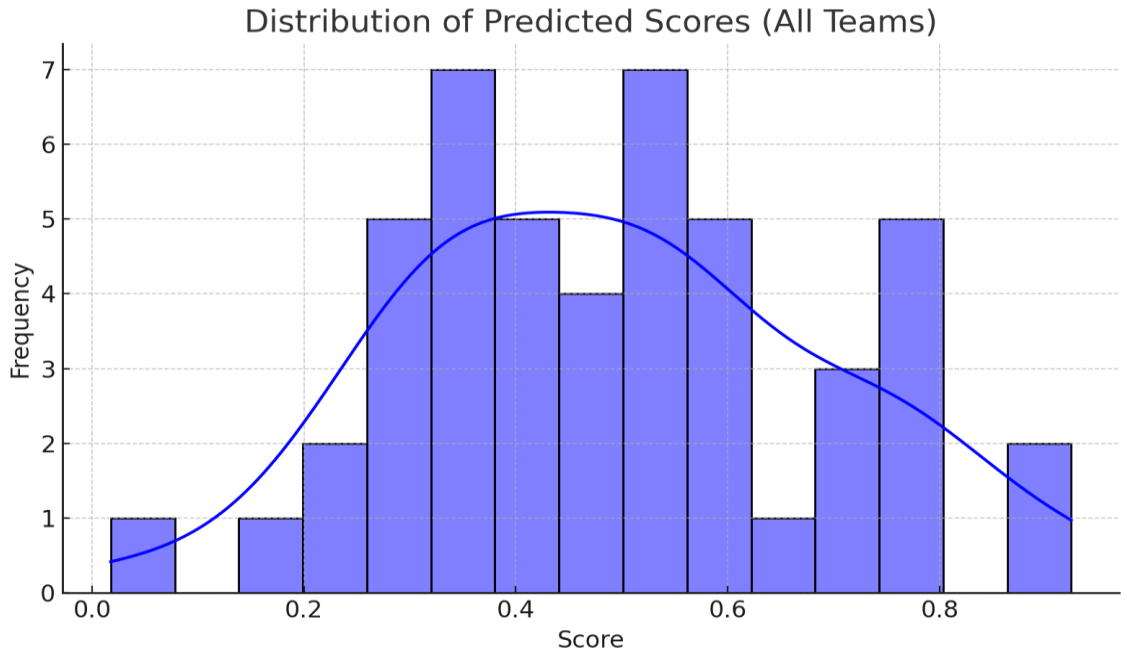
Why these were used

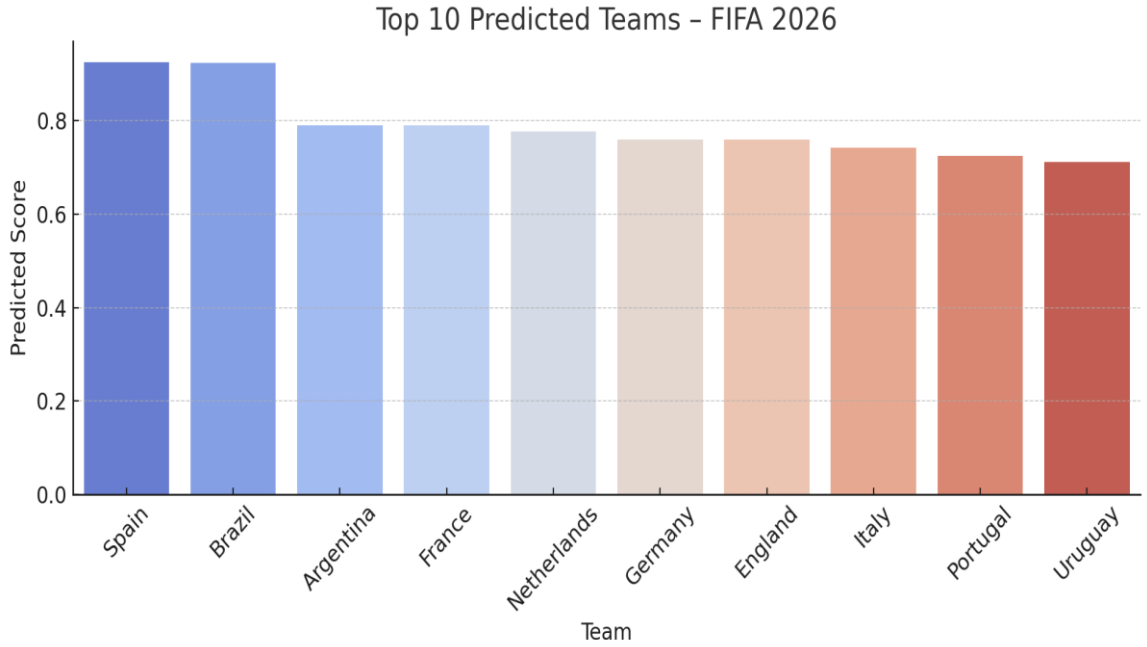
- **Use Accuracy** for quick overall sanity check.
- **Use Precision** when false finalist claims are costly (credibility).
- **Use Recall** when missing a true finalist is critical (sensitivity).
- **Use F1** to balance Precision & Recall during hyperparameter tuning.
- **Use ROC-AUC** to validate ranking quality of predicted probabilities.
- **Use Confusion Matrix** to inspect error types and decide whether to prioritize precision or recall.
- **Use ROC Curve** to visualize classifier discrimination and pick thresholds.

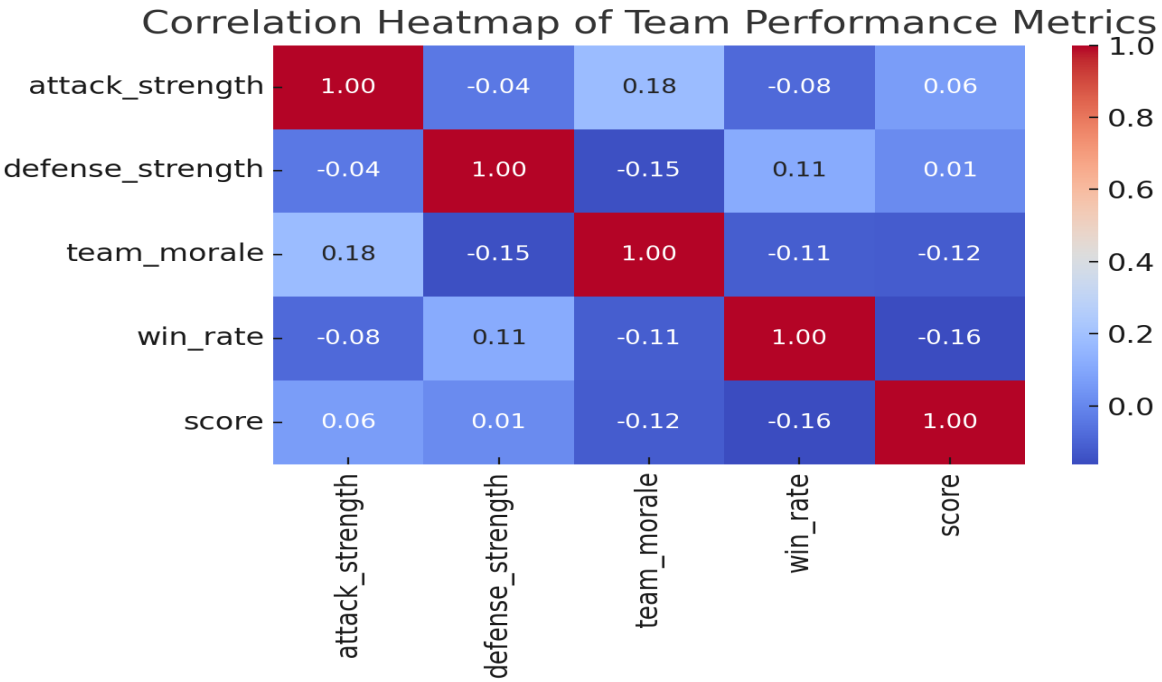
CLI vs Flask app

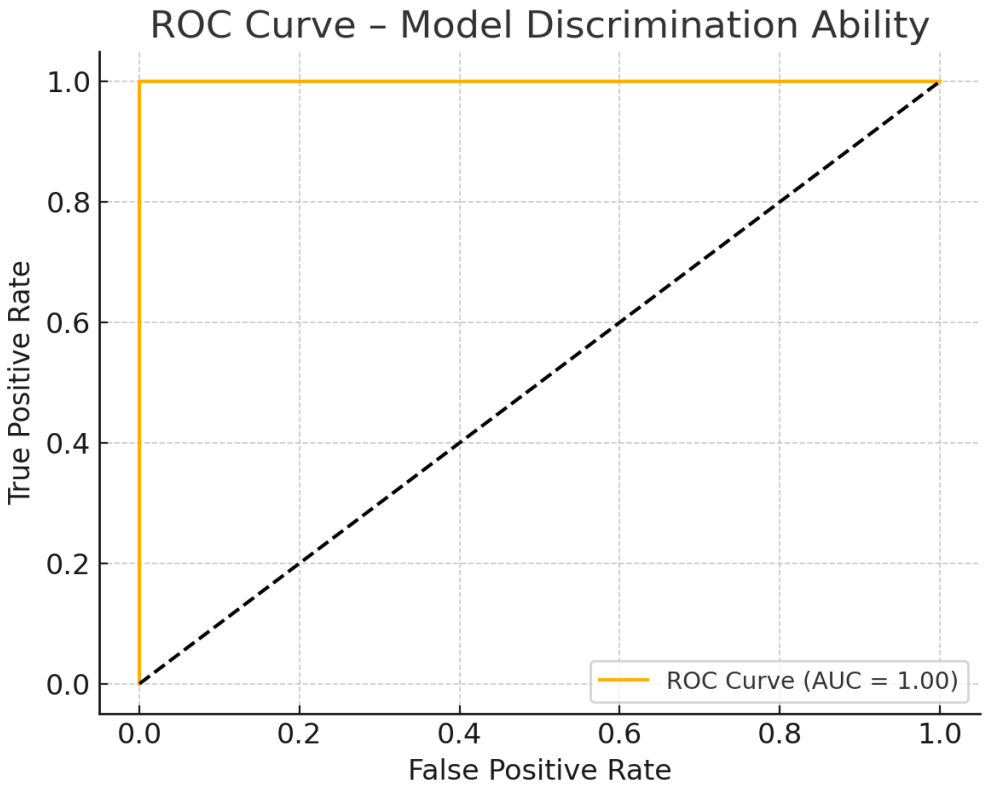
- **CLI app** — reproducible, scriptable pipeline: load → preprocess → train → eval → save outputs.
- **Flask UI** — user-friendly demo: upload data, run predictions, view/download results & charts.
- **Why both** — CLI for reproducibility/automation; Flask for presentation & non-technical users.

GRAPHS


Distribution of Predicted Scores (All Teams)
Frequency
Score
0
1
2
3
4
5
6
7
0.0
0.2
0.4
0.6
0.8


Top 10 Predicted Teams – FIFA 2026
Predicted Score
Team
0.0
0.2
0.4
0.6
0.8
Spain
Brazil
Argentina
France
Netherlands
Germany
England
Italy
Portugal
Uruguay


Correlation Heatmap of Team Performance Metrics
attack_strength
defense_strength
team_morale
win_rate
score
1.00 -0.04 0.18 -0.08 0.06
-0.04 1.00 -0.15 0.11 0.01
0.18 -0.15 1.00 -0.11 -0.12
-0.08 0.11 -0.11 1.00 -0.16
0.06 0.01 -0.12 -0.16 1.00
attack_strength
defense_strength
team_morale
win_rate
score
1.0
0.8
0.6
0.4
0.2
0.0


ROC Curve – Model Discrimination Ability
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve (AUC = 1.00)