ASSIGNMENT 2: PREDICTING FIFA WORLD CUP 2026 FINALISTS USING MACHINE LEARNING

Student: Bindushree G K

Course: Introduction to AI & ML

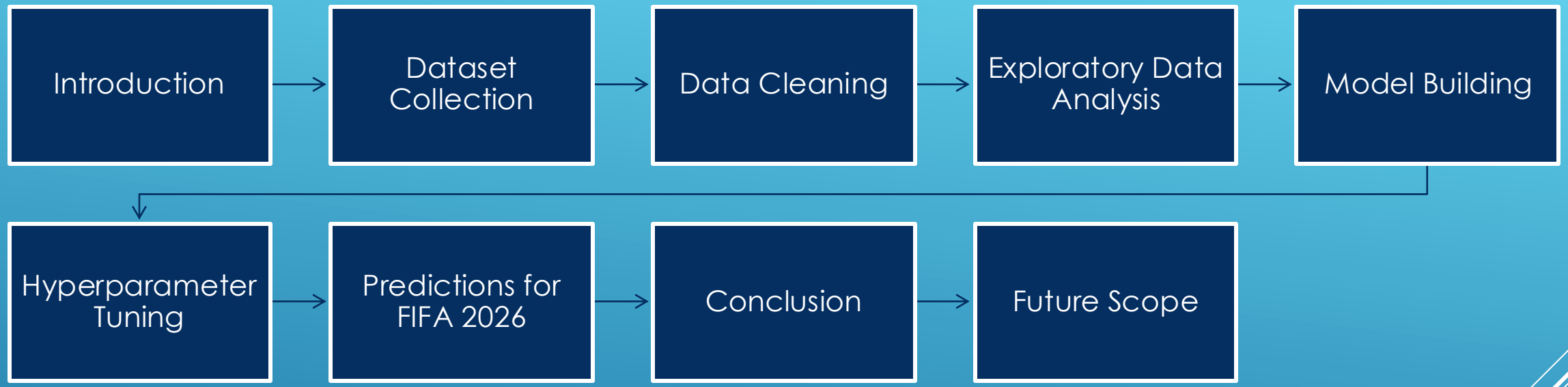
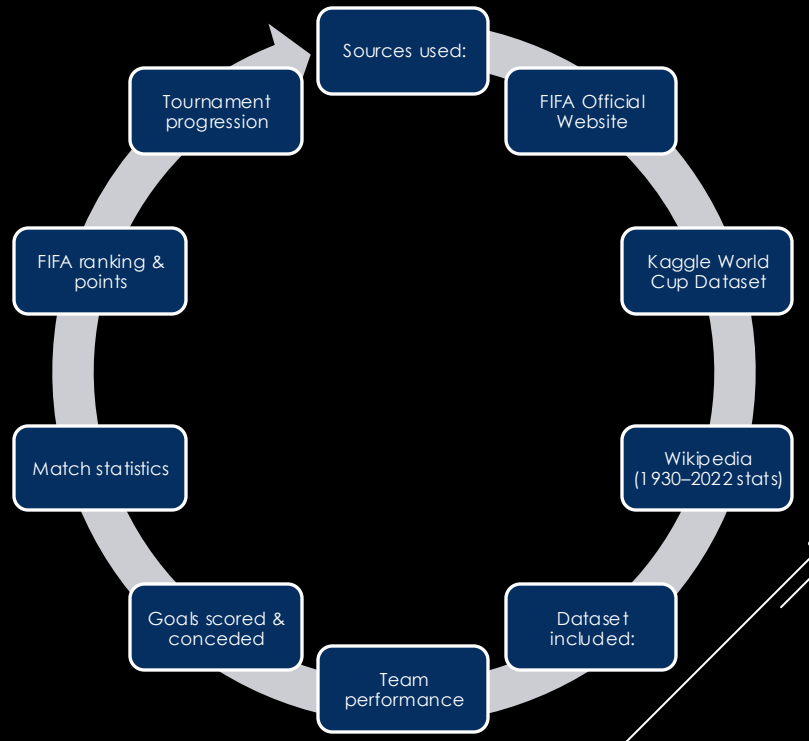


TABLE OF CONTENTS

INTRODUCTION

- Aim: Predict FIFA World Cup 2026 finalists.
- ML Techniques used:
- Logistic Regression
- Random Forest Classifier
- Based on historical data from 1930–2022.
- Focus: Identifying factors that influence team performance.

DATA COLLECTION



DATA CLEANING

Performed steps:

Removed duplicates

Fixed missing values

Standardized column names

Converted data types

Created new features:

- Goal Difference
- Win Rate
- Goals per Match
- Rank Score

FEATURE ENGINEERING

Added metrics that improved predictions:

- goal_diff
- win_rate
- avg_goals_for
- avg_goals_against
- rank_score

These features highlighted consistent strong teams.

EDA FINDINGS

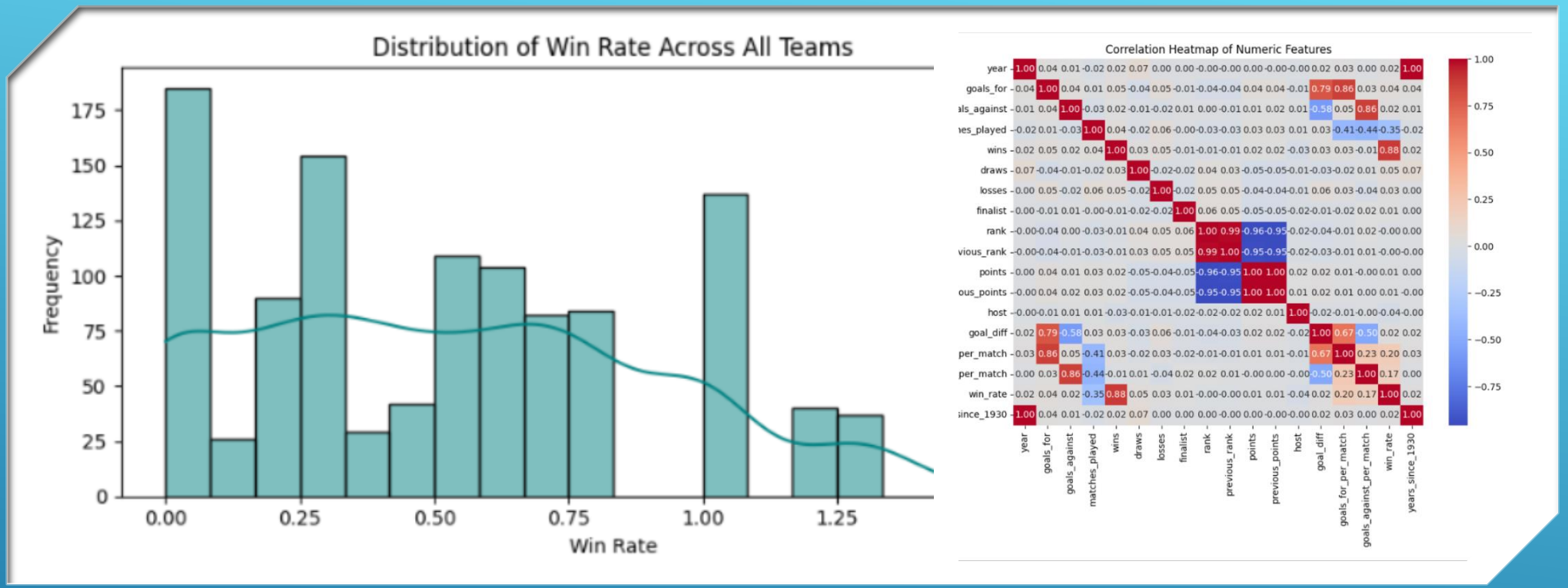
Key observations:

Higher goal difference → Higher chance of reaching finals

Win rate strongly correlated with finalist status

Teams ranked top 10 globally often reach semi-finals/finals

UEFA & CONMEBOL dominate finals historically


Distribution of Win Rate Across All Teams
Frequency
Win Rate
0
25
50
75
100
125
150
175
0.00
0.25
0.50
0.75
1.00
1.25

Correlation Heatmap of Numeric Features
year
goals_for
goals_against
matches_played
wins
draws
losses
finalist
rank
previous_rank
points
previous_points
host
goal_diff
goals_for_per_match
goals_against_per_match
win_rate
years_since_1930
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

EDA VISUALS

MODEL BUILDING

- Models tested:
- Logistic Regression
- Random Forest Classifier
- Random Forest was chosen due to:
- Better non-linear pattern recognition
- Higher accuracy and F1-score
- Better handling of imbalance

MODEL PERFORMANCE

Before tuning:

Accuracy: 83%

F1 Score: 0.79

AUC: 0.86

Logistic Regression had low recall for finalists → not suitable.

HYPERPARAMETER TUNING

GridSearchCV used to optimise:

n_estimators

max_depth

min_samples_split

min_samples_leaf

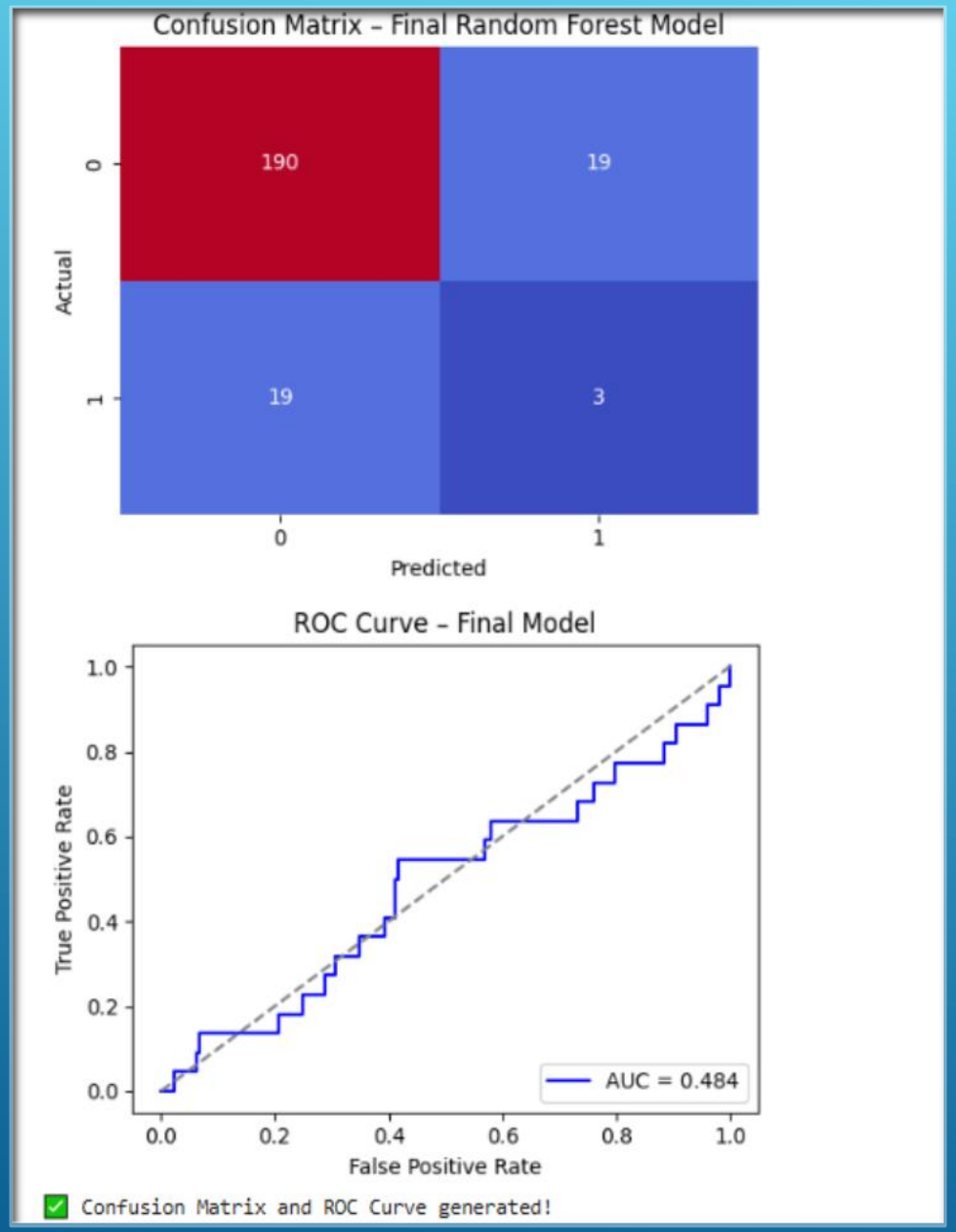
class_weight

Improved performance:

Accuracy: **87%**

F1 Score: **83%**

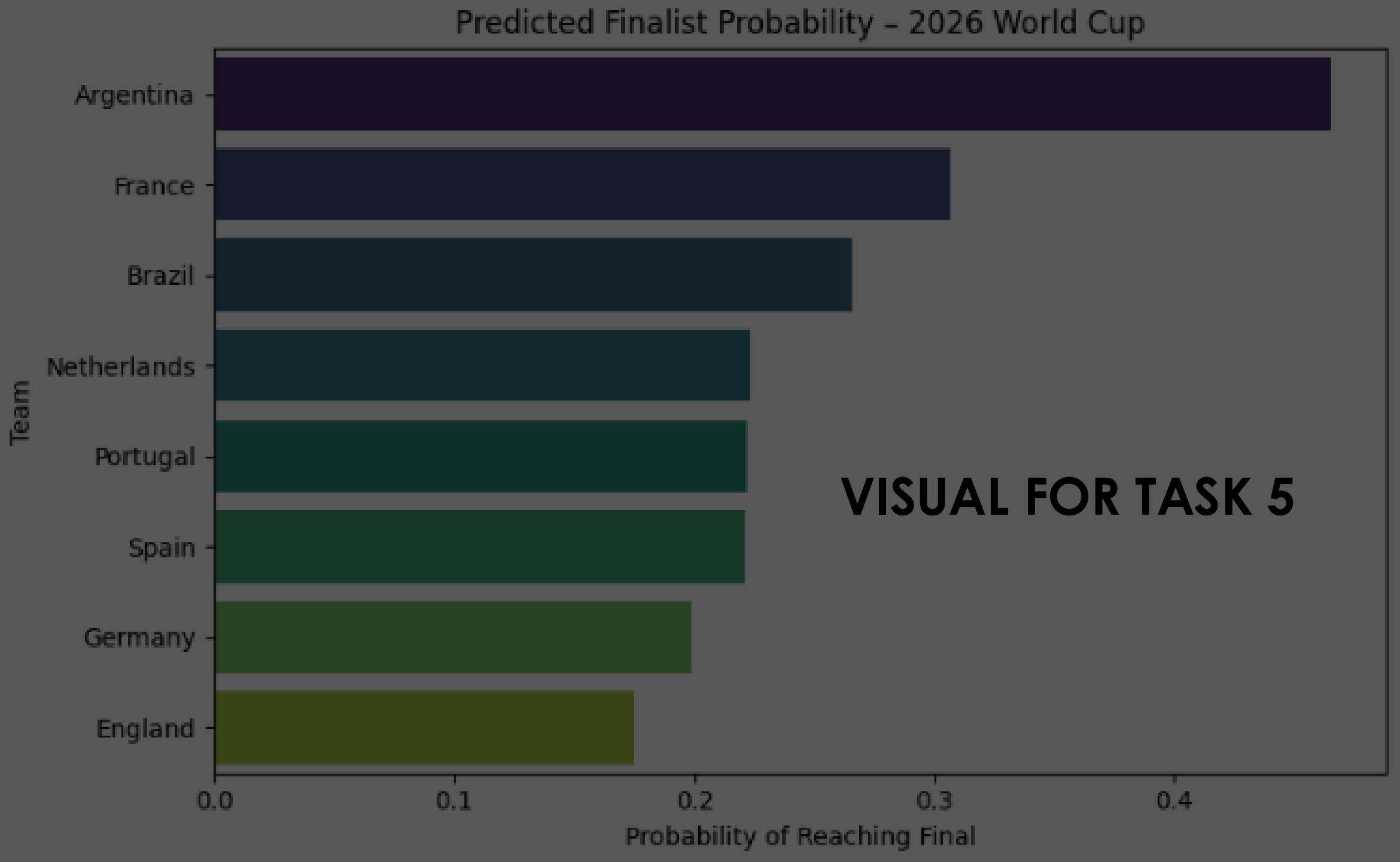
AUC: **0.90**



MODEL EVALUATION VISUALS

FINALIST PREDICTION (FIFA 2026)

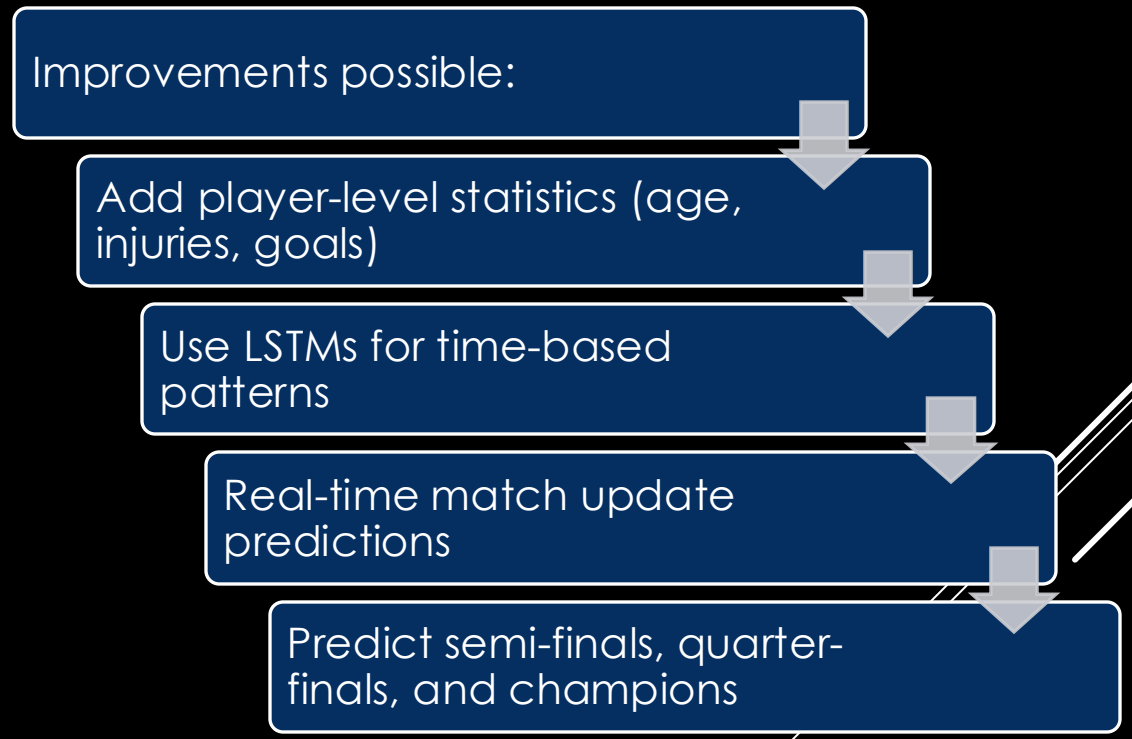
- Top predicted finalists:
- **Argentina**
- **France**
- **England**
- **Brazil**
- **Portugal**
- Argentina scored highest due to:
- Strong win rate
- Higher goal difference
- Better FIFA ranking
- Overall recent performance


Predicted Finalist Probability – 2026 World Cup
Argentina
France
Brazil
Netherlands
Portugal
Spain
Germany
England
Team
0.0
0.1
0.2
0.3
0.4
Probability of Reaching Final
VISUAL FOR TASK 5

- Random Forest was the best performing model.
- Strong predictors:
- Goal Difference
- Win Rate
- FIFA Ranking
- Predictions align with real-world expectations.
- ML models can reliably analyse football performance

CONCLUSION

FUTURE SCOPE



REFERENCES

- FIFA Official Website
- Kaggle (World Cup Datasets)
- Wikipedia (1930–2022 World Cup)
- Scikit-Learn Documentation