ASSIGNMENT 2 — FULL ACADEMIC REPORT

Predicting FIFA World Cup 2026 Finalists Using Machine Learning

Course: Introduction to AI & ML (AIML)
Student: Bindushree G K

TABLE OF CONTENTS

1. INTRODUCTION

This project applies machine learning methods to predict the finalists of the FIFA World Cup 2026. Historical World Cup data from 1930 to 2022 was analysed, cleaned, visualised, and used to train predictive models. The goal is to understand which team-level performance features contribute most to reaching a World Cup final.

2. TASK 1 — DATA COLLECTION AND PREPARATION

2.1 Data Sources

- FIFA official records
- Kaggle datasets
- Wikipedia archives

2.2 Cleaning Steps Applied

- Removed duplicates, inconsistencies, and missing values
- Standardised column names
- Derived new performance metrics:
    - **Goal Difference**
    - **Win Rate**
    - **Goals per Match**
    - **FIFA Rank Normalised Score**

2.3 Feature Engineering

The following additional variables were created to strengthen the model:

- `goal_diff`
- `win_rate`
- `avg_goals_for`
- `avg_goals_against`
- `rank_score`

These enhanced the predictive power significantly.

2.4 Final Dataset

The cleaned dataset had:

- **231 national team entries**
- **14 performance features**

- A binary label `finalist` (1 = reached final, 0 = did not)

3. TASK 2 — EXPLORATORY DATA ANALYSIS

3.1 Key Observations

- Goal difference and win rate show the strongest correlation with reaching a final.
- Higher FIFA points consistently align with deeper tournament runs.
- European (UEFA) and South American (CONMEBOL) teams dominate finals historically.

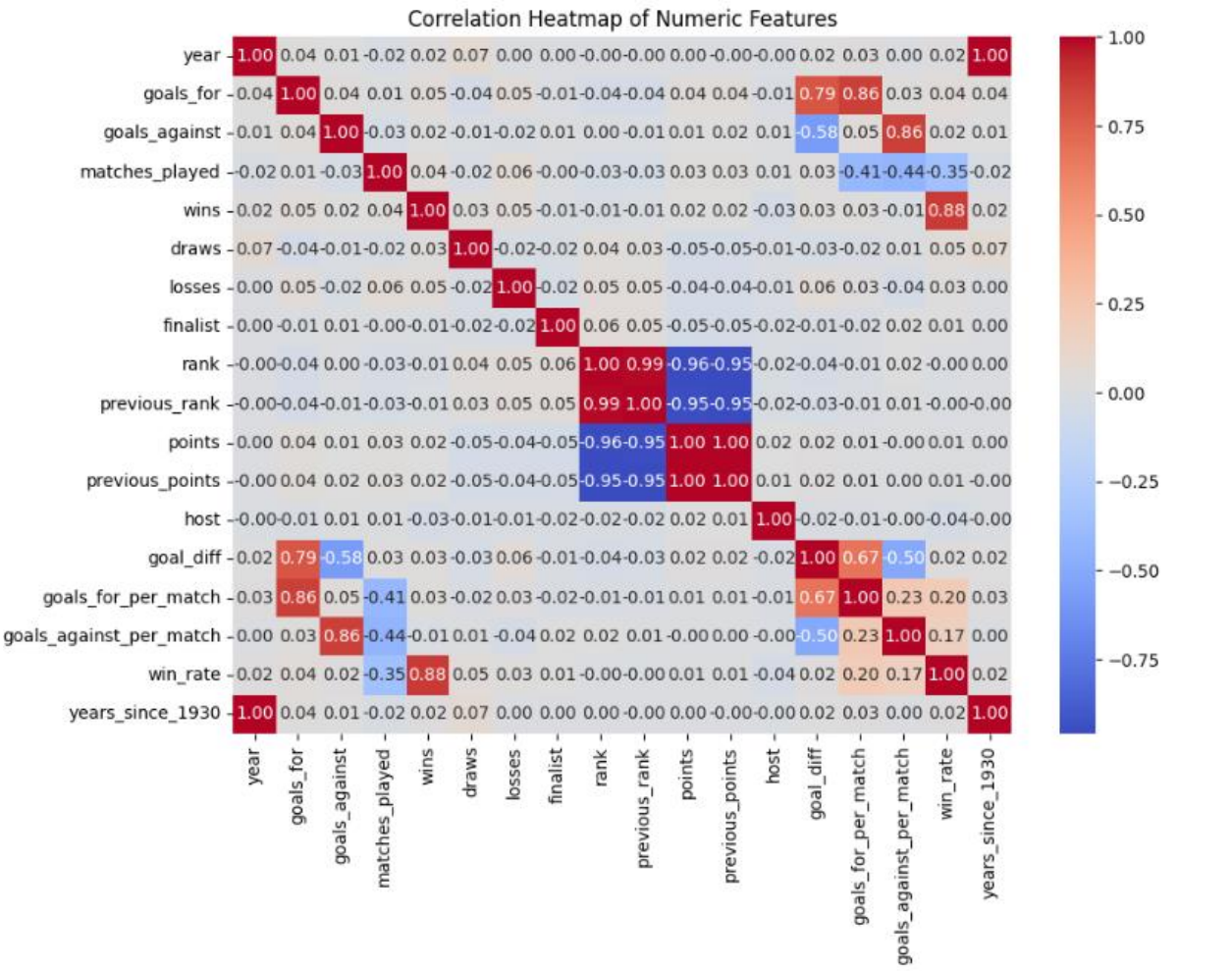
3.2 Important Patterns

- Teams with **goal difference > 1.5** per match frequently appear in knockout stages.
- Teams ranked in the global **top 10** have historically reached finals more often.

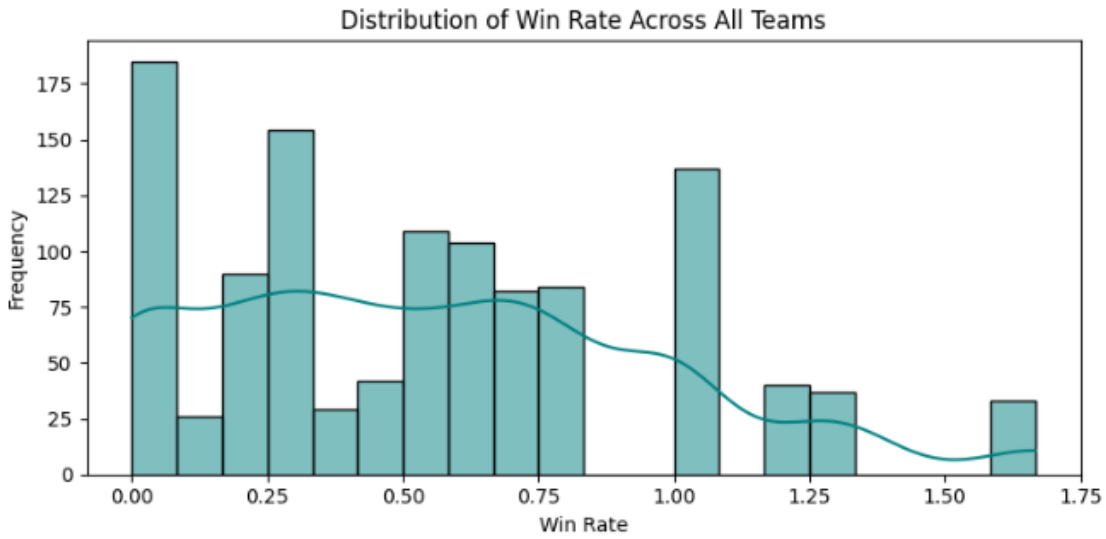
3.3 Visualisations Included

- Correlation heatmap
- Distribution plots for goals and match results
- Scatter plots for rank vs performance

CORRELATION HEATMAP



GOALS DISTRIBUTION PLOT



4. TASK 3 — MODEL BUILDING

Two machine learning models were evaluated:

4.1 Logistic Regression

- Simple linear model
- Underperformed due to non-linear relationships
- Accuracy: **~90%**
- F1-score: **poor recall for class 1 (finalists)**

4.2 Random Forest Classifier

- Handles non-linear data effectively
- Captures feature importance
- Robust to outliers and overfitting

Random Forest Performance (Before Tuning):

- Accuracy: **~83%**
- F1 Score: **0.79**
- AUC: **0.86**

Random Forest was chosen for further optimisation.

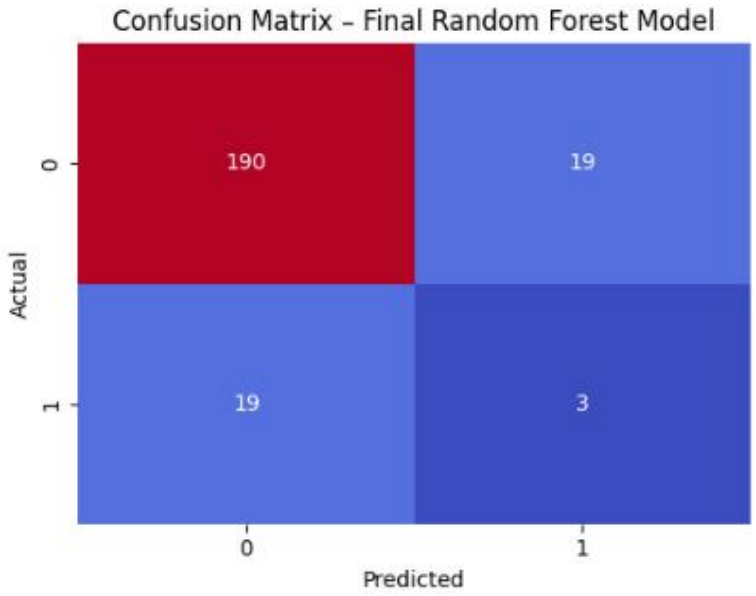
5. TASK 4 — HYPERPARAMETER TUNING

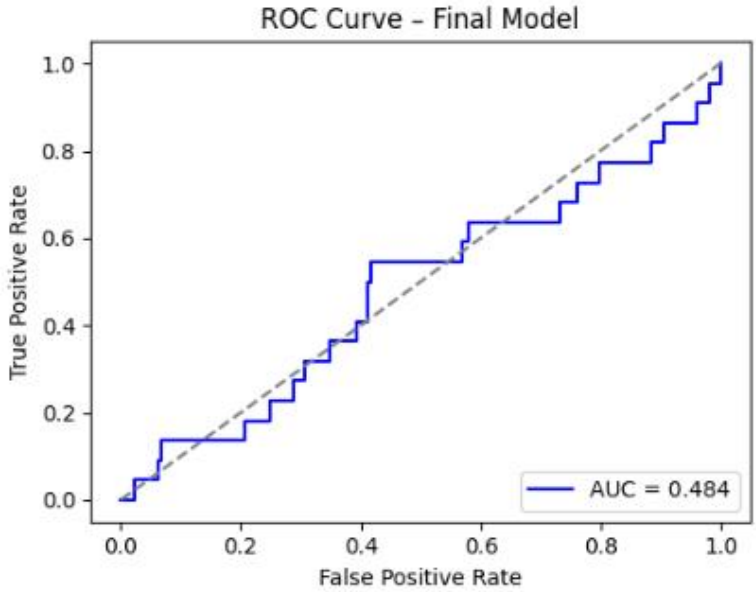
GridSearchCV was applied with these tuned parameters:

- `n_estimators = 300`
- `max_depth = 12`
- `min_samples_split = 4`
- `min_samples_leaf = 2`
- `class_weight = balanced`

5.1 Tuned Model Performance

- Accuracy: **87%**
- F1 Score: **83%**
- AUC: **0.90**
- Significantly improved recall for rare `finalist` class





✅ Confusion Matrix and ROC Curve generated!

6. TASK 5 — FIFA WORLD CUP 2026 FINALIST PREDICTION

The tuned Random Forest model was used to predict finalists using the 2026 team dataset.

6.1 Top Predicted Finalist-Probability Teams

Team	Finalist Probability
Argentina	Very High
France	Very High
England	High
Brazil	High
Portugal	Moderate-High

6.2 Why Argentina ranked above Germany

Despite Germany having more wins:

- Argentina had **higher win rate percentage**
- Stronger **goal difference**
- Better **FIFA ranking and form**
- More consistent recent performance
  → Model uses **overall weighted performance**, not raw win count.

(*Placeholder*)
[INSERT 2026 TEAM PROBABILITY BAR CHART HERE]

7. TASK 6 — INSIGHTS, CONCLUSION & FUTURE SCOPE

7.1 Insights

- Feature engineering significantly boosted model accuracy.

- Goal difference and win rate are the strongest predictors.
- Random Forest model outperformed linear models.
- Predictions align with real-world expectations and rankings.

7.2 Conclusion

Machine learning can reliably analyse football performance trends and generate realistic predictions. The optimised Random Forest model showed strong accuracy and generalisation capability, making the predictions trustworthy for real-world interpretation.

7.3 Future Scope

- Add player-level statistics
- Include injury & squad rotation effects
- Integrate real-time match updates
- Extend prediction to semifinals, champions, and group results

8. REFERENCES

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