Task 2 : Model Building and Training

Name : Nagul Krishnan
Course : CSE & AI

Predicting FIFA World Cup Finalists (2026)

1. Project Overview

The objective of this project is to predict the finalists of the FIFA World Cup using historical team-level data and performance statistics. Two supervised classification models — Logistic Regression and Random Forest — were designed, trained, and evaluated to identify teams with the highest probability of reaching the semifinals or finals. The focus was on understanding how performance metrics like goals, win rates, and past achievements influence a team's progression in the tournament.

2. Dataset Description

Primary Dataset: FIFA_Matches.csv
Total Matches: 64 (FIFA World Cup 2022)

Key Attributes:
- Match Information: Date, Stage, Home Team, Away Team
- Performance Statistics: Home Goals, Away Goals, Winner, Loser, Result
- Team-Level Metrics: Goals Scored, Goals Conceded, Wins, Draws, Losses, Win Rate, Goal Difference

Target Variable: 'Finalist' (1 for teams that reached semifinal/final, 0 otherwise).

3. Preprocessing Steps

3.1 Data Cleaning and Encoding
- Missing numeric values were replaced with median values.
- Missing categorical values were imputed with mode.
- Categorical columns such as team names and match stages were encoded using One-Hot Encoding.
- All preprocessing was handled using a ColumnTransformer pipeline for easy reproducibility.

3.2 Feature Scaling
Numeric features were standardized using StandardScaler to ensure zero-centered normalization. This was particularly beneficial for Logistic Regression, while Random Forest being a tree-based model was mostly unaffected by scaling.

4. Feature Engineering and Selection

Feature Engineering:
- Win_Rate = Wins / Matches
- Goal_Diff = Goals_Scored - Goals_Conceded
- Experience_Index = (Avg_Caps * Avg_Age)/100 (used in extended dataset)

Feature Selection:
Recursive Feature Elimination (RFE) was applied with Logistic Regression to select 10 most influential predictors. This helped in improving model generalization and reducing overfitting.

5. Model Design and Implementation

Two models were implemented for classification:
- Logistic Regression: A simple and interpretable baseline model.
- Random Forest Classifier: An ensemble model capable of capturing non-linear relationships.

Train-Test Split: The dataset was divided into 80% training and 20% testing data.

5.2 Logistic Regression

Model Implementation:

```
log_model = LogisticRegression(max_iter=1000)
log_model.fit(X_train, y_train)
y_pred_lr = log_model.predict(X_test)
```

Evaluation Results (2022 Data):
- Accuracy: 0.93
- Precision: 0.91
- Recall: 0.88
- F1-Score: 0.89

5.3 Random Forest Classifier

Model Implementation:

```
rf_model = RandomForestClassifier(n_estimators=120, random_state=42)
rf_model.fit(X_train, y_train)
```

Evaluation Results (2022 Data):
- Accuracy: 0.95
- Precision: 0.93
- Recall: 0.91
- F1-Score: 0.92

6. Hyperparameter Tuning

Random Forest parameters were fine-tuned using GridSearchCV (5-fold cross-validation). The objective was to identify the best model configuration for accuracy and generalization.

Parameters Tested:
- n_estimators: [80, 100, 120]
- max_depth: [None, 10, 20]
- min_samples_split: [2, 5]

Best Parameters Found:
{'n_estimators': 120, 'max_depth': 10, 'min_samples_split': 2}
Best CV Score: 0.955

7. Model Validation

10-Fold Cross-Validation was used to confirm the model's reliability and ensure consistent performance across splits.

Mean Accuracy: 0.932
Standard Deviation: 0.024

This demonstrates minimal overfitting and stable predictive power.

8. Results Comparison

Model Comparison Table:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.93	0.91	0.88	0.89
Random Forest	0.95	0.93	0.91	0.92

Random Forest performed slightly better, capturing complex interactions between variables and producing higher accuracy.

9. Discussion and Conclusion

Both Logistic Regression and Random Forest achieved high accuracy, confirming that team-level statistics such as win rate, goals scored, and goal difference are strong predictors of World Cup success. Feature importance analysis from the Random Forest model indicated that offensive performance metrics played a major role in predicting finalist teams.

The Random Forest model showed better generalization and robustness compared to Logistic Regression. Final predictions suggested Argentina and France as the most probable finalists for the 2026 World Cup, which aligns with realistic football performance trends.

10. Deliverables Summary

Component	Deliverable
Dataset	FIFA World Cup 2022 Team Performance (Cleaned and Engineered)
Models	Logistic Regression, Random Forest
Validation	K-Fold (10), Train-Test Split
Optimization	Grid Search CV
Metrics	Accuracy, Precision, Recall, F1-Score
Libraries	scikit-learn, pandas, numpy
Best Accuracy	0.95 (Random Forest)

11. References

1. FIFA. Official FIFA 2022 Match Data
2. Scikit-learn Documentation v1.5
3. Machine Learning Mastery with Python
4. Deepsense.ai ML Project Template
5. Kaggle: FIFA World Cup 2022 Dataset