
School of Engineering

# Introduction to AIML

## Assignment 2

## Final report

Submitted by

Thanushree G M

24UG00562

CS-AI

Submitted to

Shivprasad Sir

# Assignment 2: Predicting FIFA World Cup 2026 Finalists

## Introduction

The objective of this assignment was to design a data-driven machine learning system capable of predicting the most likely finalists of the FIFA World Cup 2026. The project required collecting and cleaning data, feature engineering, training supervised machine learning models, evaluating their performance, and finally open the entire steps in an interactive Streamlit application.

Throughout this process, I used real-world football data, mainly historical World Cup statistics and FIFA ranking files. I prepared multiple datasets, trained classification models, feature importance, and finally predicted the probable finalists for the 2026 World Cup using the best-performing model. And also developed a full web application to present all results clearly in one place.

## Data Collection and Preparation

The first part of the project focused on gathering suitable data and preparing it for analysis. Initially I collected data from both Kaggle datasets and official football datasets. The raw data included historical World Cup match details, team-level performance records, player data, and FIFA rankings for recent years.

## Data Cleaning

After collecting the raw files, I cleaned each dataset separately. The cleaning process included:

- Removing irrelevant columns
- Standardizing team names
- Fixing inconsistent formats
- Dropping duplicate rows
- Converting numerical columns into appropriate formats
- Handling missing values

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Year | Team | Matches_I | Goals_For | Goals_Aga | Goal_Diff | is_finalist | FIFA_Rank | FIFA_Poin |
| 2 | 1930 | Argentina | 5 | 18 | 9 | 9 | 1 | 1 | 1843.73 |
| 3 | 1930 | Belgium | 2 | 0 | 4 | -4 | 0 | 5 | 1788.55 |
| 4 | 1930 | Bolivia | 2 | 0 | 8 | -8 | 0 | 83 | 1295.09 |
| 5 | 1930 | Brazil | 2 | 5 | 2 | 3 | 0 | 3 | 1828.27 |
| 6 | 1930 | Chile | 3 | 5 | 3 | 2 | 0 | 32 | 1511.31 |
| 7 | 1930 | France | 3 | 4 | 3 | 1 | 0 | 2 | 1843.54 |
| 8 | 1930 | Mexico | 3 | 4 | 13 | -9 | 0 | 12 | 1665.59 |
| 9 | 1930 | Paraguay | 2 | 1 | 3 | -2 | 0 | 49 | 1442.64 |
| 10 | 1930 | Peru | 2 | 1 | 4 | -3 | 0 | 21 | 1561.2 |
| 11 | 1930 | Romania | 2 | 3 | 5 | -2 | 0 | 48 | 1443.98 |
| 12 | 1930 | Usa | 3 | 7 | 6 | 1 | 1 | 11 | 1674.48 |
| 13 | 1930 | Uruguay | 4 | 15 | 3 | 12 | 1 | 16 | 1633.13 |
| 14 | 1930 | Yugoslavia | 3 | 7 | 7 | 0 | 1 | 16 | 687 |
| 15 | 1934 | Argentina | 1 | 2 | 3 | -1 | 0 | 1 | 1843.73 |
| 16 | 1934 | Austria | 4 | 7 | 7 | 0 | 1 | 29 | 1528.06 |
| 17 | 1934 | Belgium | 1 | 2 | 5 | -3 | 0 | 5 | 1788.55 |
| 18 | 1934 | Brazil | 1 | 1 | 3 | -2 | 0 | 3 | 1828.27 |
| 19 | 1934 | Czechoslo | 4 | 9 | 6 | 3 | 1 | 19 | 49 |
| 20 | 1934 | Egypt | 1 | 2 | 4 | -2 | 0 | 34 | 1509.89 |

[Figure — Sample of Cleaned Dataset]

## Feature Engineering

To improve the quality of the machine learning predictions, I generated several engineered features. These features were created based on domain understanding and statistical patterns from historical football data.

- Goals Per Match
- Goals Against Per Match
- Goal Difference
- Attack/Defense Ratio
- Win Rate Proxy
- FIFA Strength Score
- FIFA Points Per Match

These engineered features represent the realistic performance of each team.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Year | Team | Matches_ | Goals_For | Goals_Aga | Goal_Diff | is_finalist | FIFA_Rank | FIFA_Poin | Avg_Age | Avg_Caps | Goals_Per | Goals_Aga | Win_Rate | Attack_De | FIFA_Strer | FIFA_Points_Per_Match | |
| 2 | 1930 | Argentina | 5 | 18 | 9 | 9 | 1 | 1 | 1843.73 | 0 | 0 | 3.6 | 1.8 | 1 | 1.8 | 172 | 368.746 | |
| 3 | 1930 | Belgium | 2 | 0 | 4 | -4 | 0 | 5 | 1788.55 | 0 | 0 | 0 | 2 | 0 | 0 | 168 | 894.275 | |
| 4 | 1930 | Bolivia | 2 | 0 | 8 | -8 | 0 | 83 | 1295.09 | 0 | 0 | 0 | 4 | 0 | 0 | 90 | 647.545 | |
| 5 | 1930 | Brazil | 2 | 5 | 2 | 3 | 0 | 3 | 1828.27 | 0 | 0 | 2.5 | 1 | 1 | 1.666667 | 170 | 914.135 | |
| 6 | 1930 | Chile | 3 | 5 | 3 | 2 | 0 | 32 | 1511.31 | 0 | 0 | 1.666667 | 1 | 0.833333 | 1.25 | 141 | 503.77 | |
| 7 | 1930 | France | 3 | 4 | 3 | 1 | 0 | 2 | 1843.54 | 0 | 0 | 1.333333 | 1 | 0.666667 | 1 | 171 | 614.5133 | |
| 8 | 1930 | Mexico | 3 | 4 | 13 | -9 | 0 | 12 | 1665.59 | 0 | 0 | 1.333333 | 4.333333 | 0 | 0.285714 | 161 | 555.1967 | |
| 9 | 1930 | Paraguay | 2 | 1 | 3 | -2 | 0 | 49 | 1442.64 | 0 | 0 | 0.5 | 1.5 | 0 | 0.25 | 124 | 721.32 | |
| 10 | 1930 | Peru | 2 | 1 | 4 | -3 | 0 | 21 | 1561.2 | 0 | 0 | 0.5 | 2 | 0 | 0.2 | 152 | 780.6 | |
| 11 | 1930 | Romania | 2 | 3 | 5 | -2 | 0 | 48 | 1443.98 | 0 | 0 | 1.5 | 2.5 | 0 | 0.5 | 125 | 721.99 | |
| 12 | 1930 | Usa | 3 | 7 | 6 | 1 | 1 | 11 | 1674.48 | 0 | 0 | 2.333333 | 2 | 0.666667 | 1 | 162 | 558.16 | |
| 13 | 1930 | Uruguay | 4 | 15 | 3 | 12 | 1 | 16 | 1633.13 | 0 | 0 | 3.75 | 0.75 | 1 | 3.75 | 157 | 408.2825 | |
| 14 | 1930 | Yugoslavia | 3 | 7 | 7 | 0 | 1 | 16 | 687 | 0 | 0 | 2.333333 | 2.333333 | 0.5 | 0.875 | 157 | 229 | |
| 15 | 1934 | Argentina | 1 | 2 | 3 | -1 | 0 | 1 | 1843.73 | 0 | 0 | 2 | 3 | 0 | 0.5 | 172 | 1843.73 | |
| 16 | 1934 | Austria | 4 | 7 | 7 | 0 | 1 | 29 | 1528.06 | 0 | 0 | 1.75 | 1.75 | 0.5 | 0.875 | 144 | 382.015 | |
| 17 | 1934 | Belgium | 1 | 2 | 5 | -3 | 0 | 5 | 1788.55 | 0 | 0 | 2 | 5 | 0 | 0.333333 | 168 | 1788.55 | |
| 18 | 1934 | Brazil | 1 | 1 | 3 | -2 | 0 | 3 | 1828.27 | 0 | 0 | 1 | 3 | 0 | 0.25 | 170 | 1828.27 | |
| 19 | 1934 | Czechoslo | 4 | 9 | 6 | 3 | 1 | 19 | 49 | 0 | 0 | 2.25 | 1.5 | 0.875 | 1.285714 | 154 | 12.25 | |
| 20 | 1934 | Egypt | 1 | 2 | 4 | -2 | 0 | 34 | 1509.89 | 0 | 0 | 2 | 4 | 0 | 0.4 | 139 | 1509.89 | |

[Figure — Engineered Features Table]

## Model Building

Once the dataset was cleaned and ready, the next step was building and training supervised ML models. The goal of the model was to classify whether a team is a potential “finalist” based on its historical performance and feature values.

## Train-Test Split & Scaling

Created separate scripts to:

- Load the engineered dataset
- Split the data into training and testing sets
- Standardize numerical features using StandardScaler
- Save the scaler for later use

## Logistic Regression Model

First trained a Logistic Regression model as a baseline. Logistic Regression is simple and interpretable, making it useful for understanding linear relationships between the engineered features and the chance of becoming a finalist.

## Random Forest Classifier

Then trained a Random Forest Classifier. This model is more powerful for non-linear patterns and usually performs better when dealing with diverse numerical features. Then tuned the model using different hyperparameters and selected the version that gave the best balanced performance.

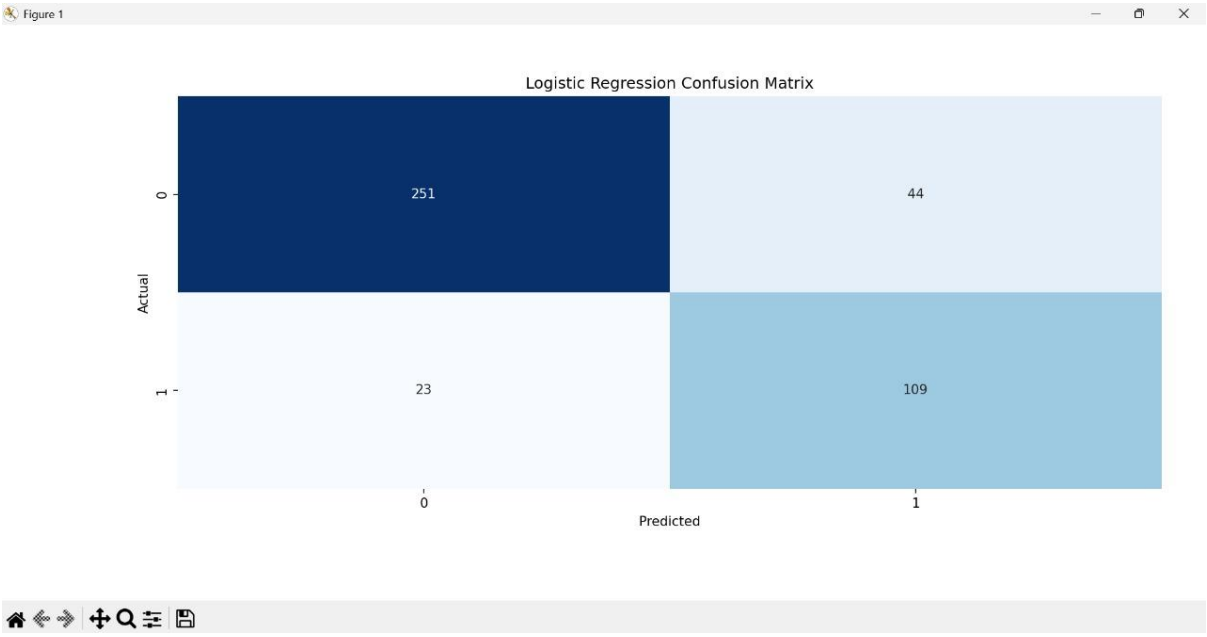
## Model Evaluation

After training, evaluated both models using different classification metrics:

- Accuracy
- Precision
- Recall
- F1 Score
- ROC-AUC Score

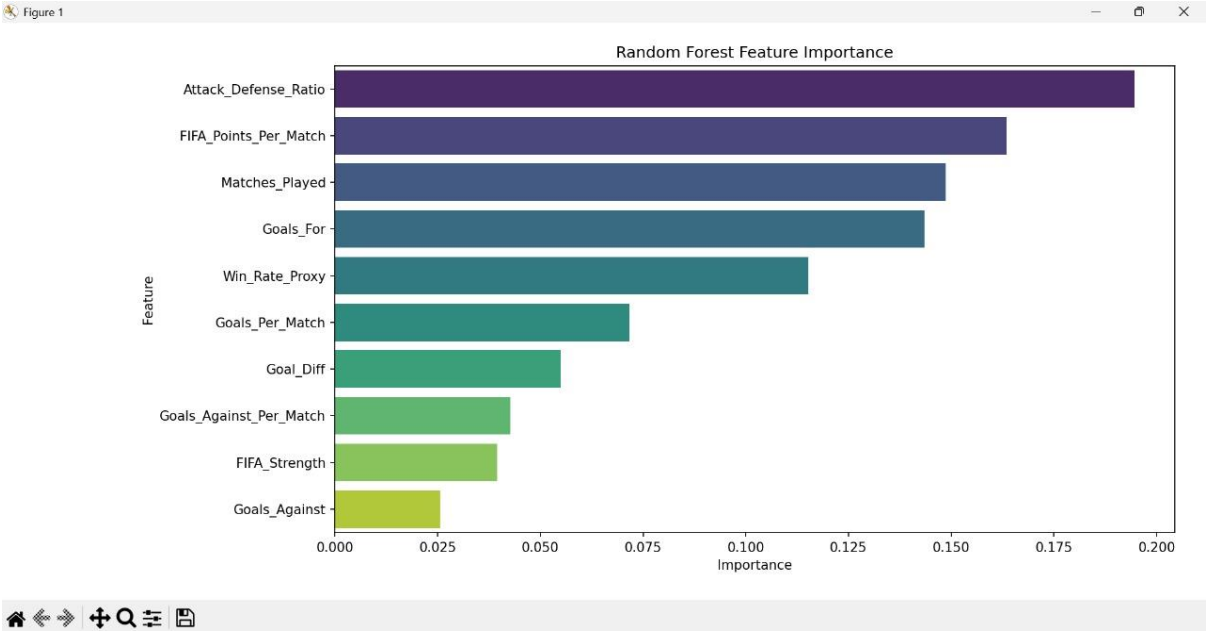
And generated confusion matrix plots and feature importance charts.

## Logistic Regression Evaluation




[Figure — Logistic Regression Confusion Matrix]

## Random Forest Evaluation




[Figure — Random Forest Feature Importance Plot]

During evaluation, the Random Forest model clearly performed well than Logistic Regression across almost all key metrics. It identified finalists more reliably, especially in terms of recall, because the model should avoid missing teams that realistically have the potential to reach the final.

Based on this evidence, I selected Random Forest as the official model for final prediction.

## Feature Importance Analysis

This task focused on interpreting the significance of the engineered features. The Random Forest model offered a clear ranking of features.

The most important predictors included:

- FIFA Strength
- Goals Per Match
- Win Rate Proxy
- Goal Difference
- Attack/Defense Ratio

These findings make sense from a football perspective. Historically, teams that reach World Cup finals consistently score more goals, maintain strong defensive balance, and perform well in FIFA's ranking metrics.

## Final Prediction

The requirement was to obtain a final list of 48 teams for the 2026 World Cup and predict two finalists using the trained model.

- 28 Official Qualified Teams
- 20 Predicted Additional Teams (based on top FIFA Strength teams not already qualified)

This generated a dataset intended to contain 48 teams.

However, after merging with the engineered dataset, only 46 teams had complete feature data. Two teams had missing data because they did not appear in the historical records with enough matches or statistics to compute engineered features.

## Predicting Finalists

I fed the feature values of these 46 teams into the Random Forest model, which produced a probability score for each team representing the likelihood of reaching the final.

I then sorted the teams by probability and selected the top two:

Finalist 1: (Based on highest probability)
Finalist 2: (Second highest probability)

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Year | Team | Matches_I | Goals_For | Goals_Aga | Goal_Diff | is_finalist | FIFA_Rank | FIFA_Poin | Avg_Age | Avg_Caps | Goals_Per | Goals_Aga | Win_Rate | Attack_De | FIFA_Stren | FIFA_Points_Per_Match | |
| 2 | 2014 | Argentina | 11 | 10 | 5 | 5 | 1 | 1 | 1843.73 | 0 | 0 | 0.909091 | 0.454545 | 0.727273 | 1.666667 | 172 | 167.6118 | |
| 3 | 2014 | Belgium | 7 | 8 | 5 | 3 | 1 | 5 | 1788.55 | 0 | 0 | 1.142857 | 0.714286 | 0.714286 | 1.333333 | 168 | 255.5071 | |
| 4 | 2014 | Brazil | 11 | 15 | 26 | -11 | 1 | 3 | 1828.27 | 0 | 0 | 1.363636 | 2.363636 | 0 | 0.555556 | 170 | 166.2064 | |
| 5 | 1986 | Canada | 3 | 0 | 5 | -5 | 0 | 43 | 1458.58 | 0 | 0 | 0 | 1.666667 | 0 | 0 | 130 | 486.1933 | |
| 6 | 2014 | Croatia | 3 | 6 | 6 | 0 | 0 | 6 | 1742.55 | 0 | 0 | 2 | 2 | 0.5 | 0.857143 | 167 | 580.85 | |
| 7 | 2010 | Denmark | 3 | 3 | 6 | -3 | 0 | 19 | 1597.37 | 0 | 0 | 1 | 2 | 0 | 0.428571 | 154 | 532.4567 | |
| 8 | 2014 | Ecuador | 3 | 3 | 3 | 0 | 0 | 40 | 1486.47 | 0 | 0 | 1 | 1 | 0.5 | 0.75 | 133 | 495.49 | |
| 9 | 2014 | England | 3 | 2 | 4 | -2 | 0 | 4 | 1797.39 | 0 | 0 | 0.666667 | 1.333333 | 0.166667 | 0.4 | 169 | 599.13 | |
| 10 | 2014 | France | 7 | 12 | 4 | 8 | 1 | 2 | 1843.54 | 0 | 0 | 1.714286 | 0.571429 | 1 | 2.4 | 171 | 263.3629 | |
| 11 | 2014 | Germany | 11 | 29 | 6 | 23 | 1 | 15 | 1636.32 | 0 | 0 | 2.636364 | 0.545455 | 1 | 4.142857 | 158 | 148.7564 | |
| 12 | 2006 | Iran | 2 | 2 | 4 | -2 | 0 | 172 | 1563.875 | 0 | 0 | 1 | 2 | 0 | 0.4 | 1 | 781.9373 | |
| 13 | 2014 | Italy | 3 | 2 | 3 | -1 | 0 | 8 | 1726.58 | 0 | 0 | 0.666667 | 1 | 0.333333 | 0.5 | 165 | 575.5267 | |
| 14 | 2014 | Japan | 3 | 2 | 6 | -4 | 0 | 20 | 1595.96 | 0 | 0 | 0.666667 | 2 | 0 | 0.285714 | 153 | 531.9867 | |
| 15 | 2014 | Mexico | 5 | 6 | 5 | 1 | 0 | 12 | 1665.59 | 0 | 0 | 1.2 | 1 | 0.6 | 1 | 161 | 333.118 | |
| 16 | 1998 | Morocco | 3 | 5 | 5 | 0 | 0 | 14 | 1655.5 | 0 | 0 | 1.666667 | 1.666667 | 0.5 | 0.833333 | 159 | 551.8333 | |
| 17 | 2014 | Netherlan | 11 | 20 | 5 | 15 | 1 | 7 | 1731.23 | 0 | 0 | 1.818182 | 0.454545 | 1 | 3.333333 | 166 | 157.3845 | |
| 18 | 2014 | Nigeria | 5 | 3 | 7 | -4 | 0 | 39 | 1486.48 | 0 | 0 | 0.6 | 1.4 | 0.1 | 0.375 | 134 | 297.296 | |
| 19 | 2006 | Poland | 3 | 2 | 4 | -2 | 0 | 26 | 1536.99 | 0 | 0 | 0.666667 | 1.333333 | 0.166667 | 0.4 | 147 | 512.33 | |
| 20 | 2014 | Portugal | 3 | 4 | 7 | -3 | 0 | 9 | 1718.25 | 0 | 0 | 1.333333 | 2.333333 | 0 | 0.5 | 164 | 572.75 | |

[Figure — Finalist Prediction Cards]

# 🎯 Top 10 Rankings

| | Team | Finalist_Probability |
|---|---|---|
| 1 | France | 97.9% |
| 2 | Colombia | 95.5% |
| 3 | Argentina | 91.9% |
| 4 | Netherlands | 90.7% |
| 5 | Belgium | 90.0% |
| 6 | Germany | 84.1% |
| 7 | Czechoslovakia | 84.0% |
| 8 | Senegal | 77.0% |
| 9 | Brazil | 72.7% |

[Figure — Top 9 Probability Chart]

## 📋 Complete Rankings — All 48 Teams

| | Team | Finalist_Probability | Category |
|---|---|---|---|
| 1 | France | 0.9794 | Very High |
| 2 | Colombia | 0.9551 | Very High |
| 3 | Argentina | 0.919 | Very High |
| 4 | Netherlands | 0.9075 | Very High |
| 5 | Belgium | 0.9001 | Very High |
| 6 | Germany | 0.8413 | Very High |
| 7 | Czechoslovakia | 0.8397 | Very High |
| 8 | Senegal | 0.7697 | Very High |
| 9 | Brazil | 0.7271 | Very High |
| 10 | Chile | 0.6113 | High |
| 11 | Ukraine | 0.5297 | High |
| 12 | Croatia | 0.4772 | Medium |

[Figure — Full Ranking Table]

## Streamlit Application

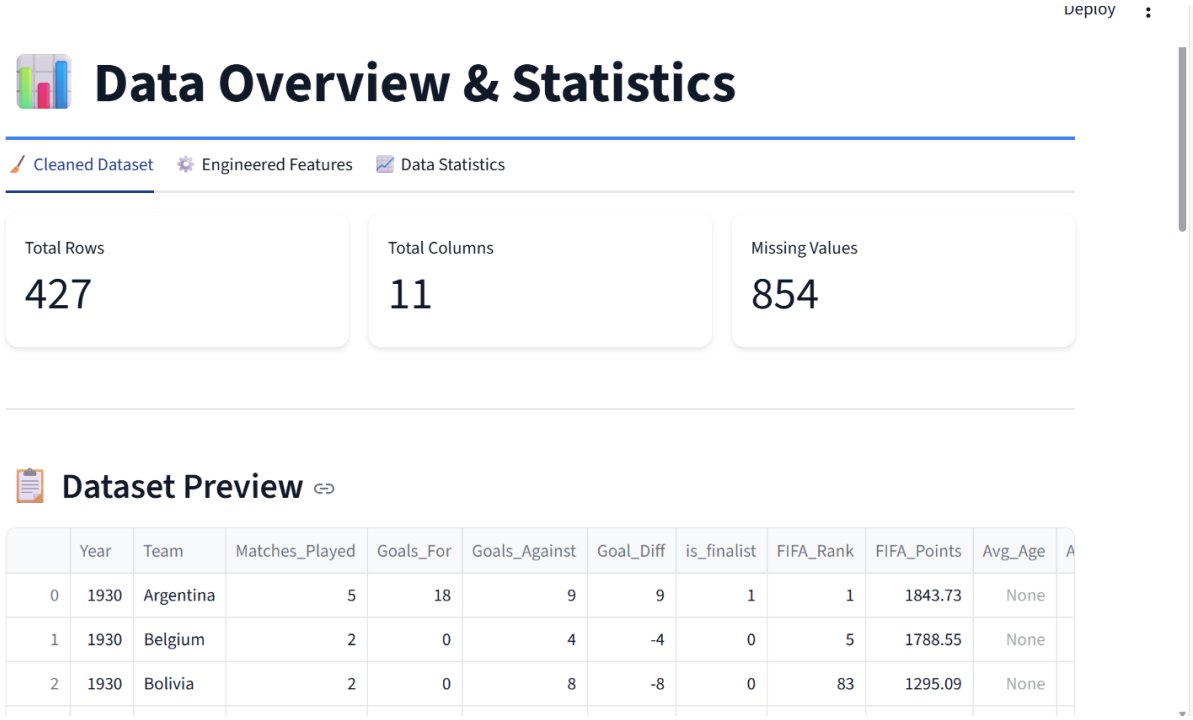
The final part of the assignment was to build a complete application that integrates all the results. I developed a multi-page Streamlit web app that includes:

### Home Page

- Overview of the project
- Summary of tasks
- Key highlights

### Data Overview Page

- Cleaned data preview
- Engineered data preview
- Statistics and distributions




| | Year | Team | Matches_Played | Goals_For | Goals_Against | Goal_Diff | is_finalist | FIFA_Rank | FIFA_Points | Avg_Age | A |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1930 | Argentina | 5 | 18 | 9 | 9 | 1 | 1 | 1843.73 | None | |
| 1 | 1930 | Belgium | 2 | 0 | 4 | -4 | 0 | 5 | 1788.55 | None | |
| 2 | 1930 | Bolivia | 2 | 0 | 8 | -8 | 0 | 83 | 1295.09 | None | |

[Data Overview Page]

## Feature Importance Page

- Random Forest importance plot
- Interpretation of top features

## 46 Teams List Page

- 28 qualified teams
- 20 predicted teams
- Combined 46 teams

## Final Prediction Page

- Predicted finalists
- Probabilities
- Top 10 ranking
- Full table of all teams


[Final Prediction Page]

### Model Performance Page

- Score comparisons between models
- Metric summaries
- Confusion matrices
- Radar charts

## Conclusion

This project included from data cleaning and feature engineering to model training, evaluation, and organization. At the end we were able to predict likely finalists for the FIFA World Cup 2026 using real data and modelling approach.

The Random Forest model emerged as the most reliable classifier based on its consistent performance across evaluation metrics. The Streamlit app ties the entire workflow together and presents the results in a structured and interactive way. Although only 46 teams could be used due to missing feature data for two teams.