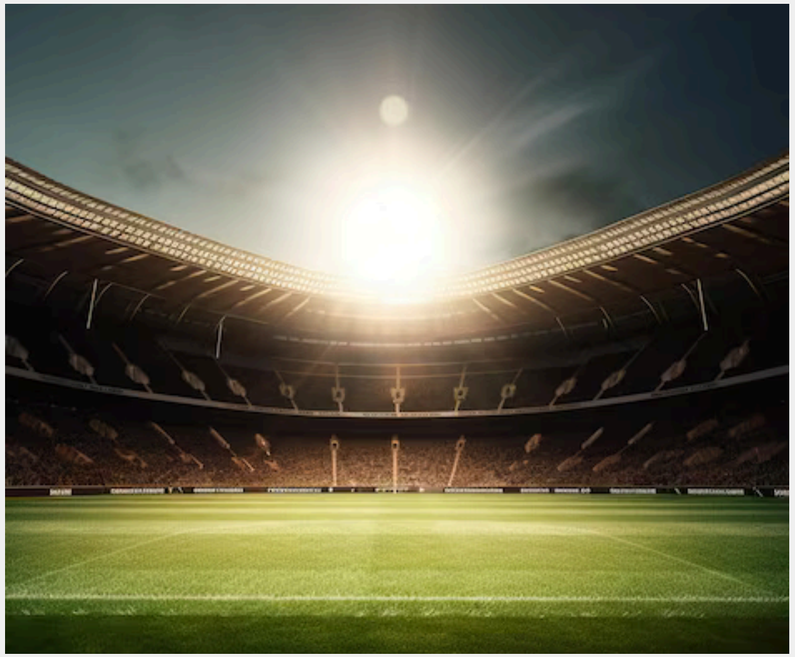
# FIFA 2026 World Cup Winner Predictor & Simulator

## A Deep Dive into Data-Driven Football Projections.

# About the Model

A FIFA machine learning (ML) model is a type of predictive model that uses machine learning algorithms to analyze historical football data and make predictions about upcoming matches or tournaments organized by FIFA World cup 2026. This type of model leverages the power of artificial intelligence and data analysis to provide insights into various aspects of the game, such as match outcomes, player performances, team strategies, and more.



# LIBRARIES USED

01 numpy

02 pandas

03 seaborn

04 matplotlib

05 sklearn

# DATASET INFO

## international_matches.csv

```
df = pd.read_csv("international_matches.csv")
df.columns
```

```
Index(['date', 'home_team', 'away_team', 'home_team_continent',
       'away_team_continent', 'home_team_fifa_rank', 'away_team_fifa_rank',
       'home_team_total_fifa_points', 'away_team_total_fifa_points',
       'home_team_score', 'away_team_score', 'tournament', 'city', 'country',
       'neutral_location', 'shoot_out', 'home_team_result',
       'home_team_goalkeeper_score', 'away_team_goalkeeper_score',
       'home_team_mean_defense_score', 'home_team_mean_offense_score',
       'home_team_mean_midfield_score', 'away_team_mean_defense_score',
       'away_team_mean_offense_score', 'away_team_mean_midfield_score'],
      dtype='object')
```

## Groupes - V3.csv

```
df = pd.read_csv("Groupes - V3.csv")
df.columns
```

```
Index(['team', 'groups', 'First match against', 'Second match against',
       'Third match against'],
      dtype='object')
```

# DATA PREPROCESSING

1. Eliminating irrelevant columns
2. Null values replaced with 0

# Machine Learning Model



## Data Splitting

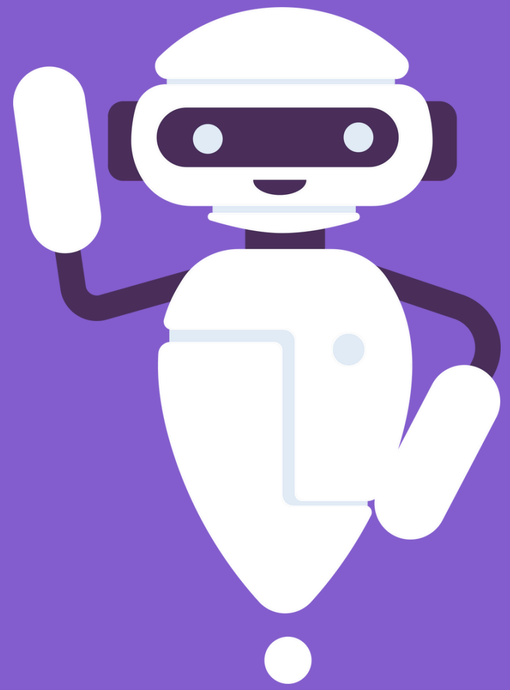
- Splitting of dataset into training and testing subsets using the train_test_split function.
- X contains the features we're using for prediction (rank difference and point difference), and y is the target variable (whether the home team won).

## Logistic Regression with Polynomial Features

- Use of a logistic regression model to predict whether the home team won or not.

## Model Fitting

- Training the model on the training data (X_train, y_train) using the .fit() method.



# Machine Learning Model

```
#X, y = df.loc[:,["team_diff","avg_rank_diff","rank_diff", "point_diff"]], df["home_won"]
X, y = df.loc[:,["rank_diff", "point_diff"]], df["home_won"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1, random_state=42)

logreg = linear_model.LogisticRegression(C = 1e-5)    #C = 1e-5 is the regularization strength parameter.
features = PolynomialFeatures(degree = 2)
model = Pipeline([("polynomial_features", features),("logistic_regression", logreg)])
model = model.fit(X_train, y_train)
```

# MATCH SIMULATOR

The Model is made using real-world FIFA Scenario. It mimics how matches are done in the real world

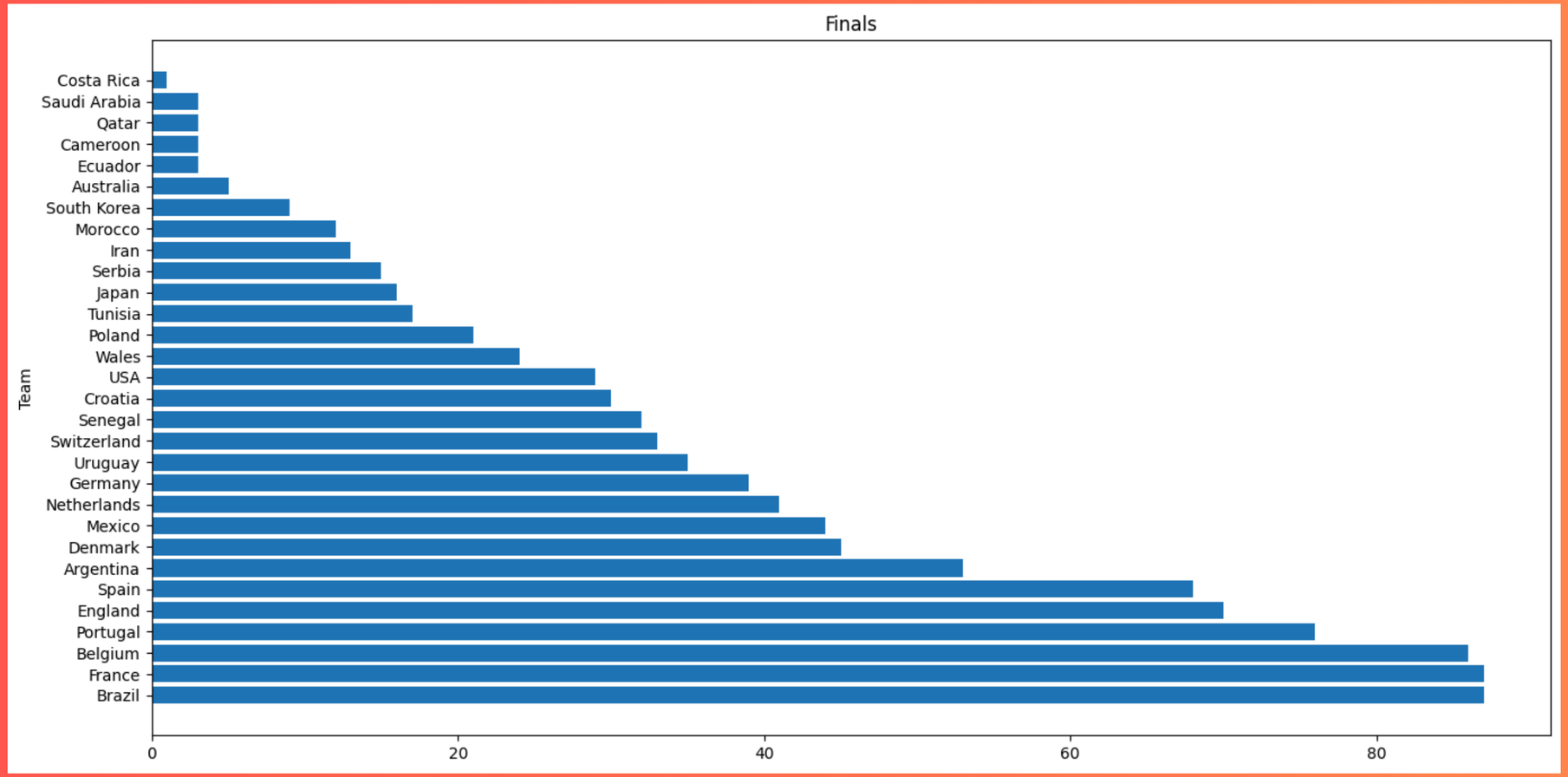
**01 ROUND OF 16**

**02 QUARTER FINALS**

**03 SEMI FINALS**

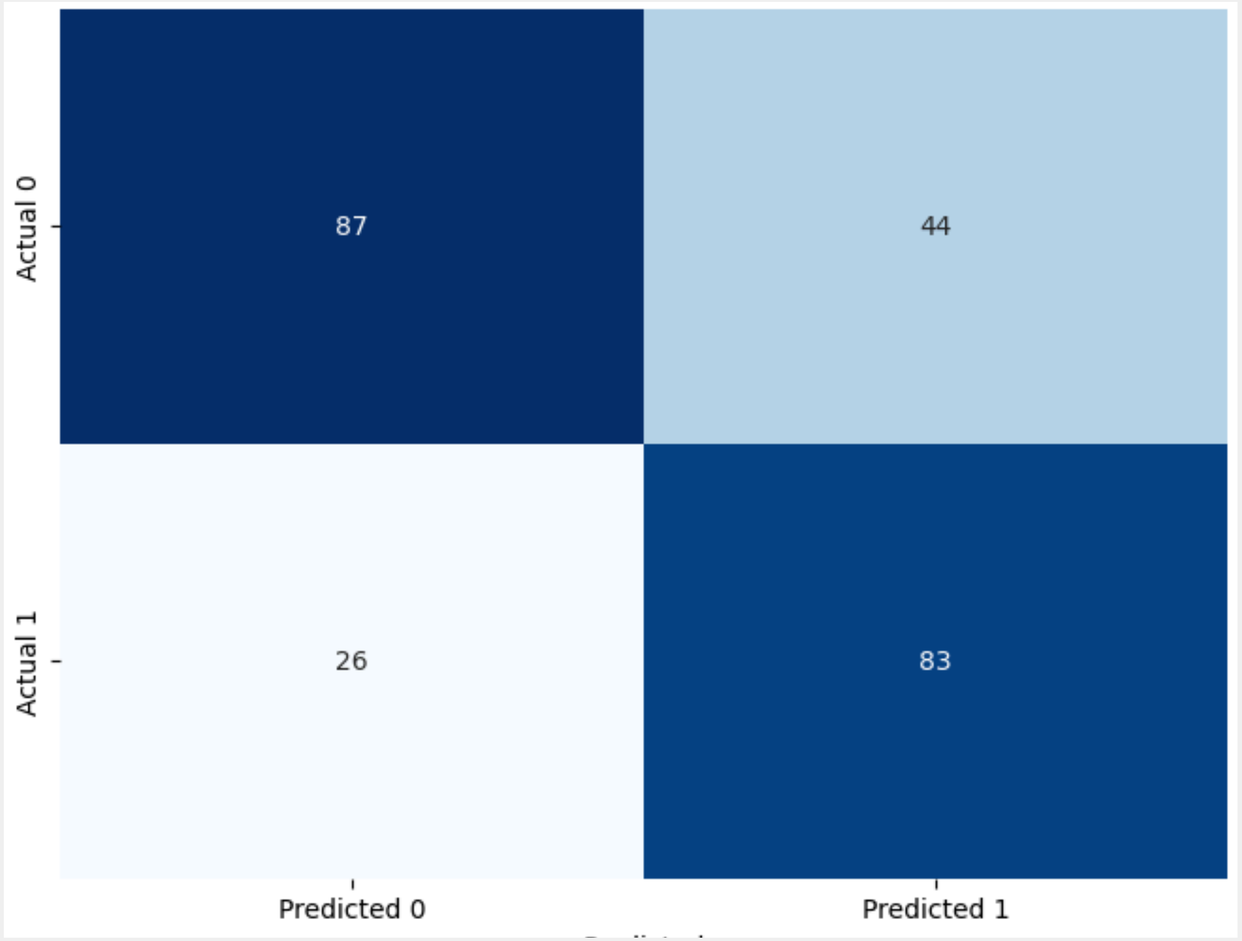
**04 FINALS**

# RESULTS


Finals
Team
Costa Rica
Saudi Arabia
Qatar
Cameroon
Ecuador
Australia
South Korea
Morocco
Iran
Serbia
Japan
Tunisia
Poland
Wales
USA
Croatia
Senegal
Switzerland
Uruguay
Germany
Netherlands
Mexico
Denmark
Argentina
Spain
England
Portugal
Belgium
France
Brazil
0
20
40
60
80

# Model Evaluation

## Confusion Matrix:



# Model Evaluation

## Performance Metrics:

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

MODEL'S ACCURACY = 0.71

$$Recall = \frac{TP}{TP + FN}$$

MODEL'S RECALL = 0.76

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

MODEL'S F1 SCORE = 0.70

$$Precision = \frac{TP}{TP + FP}$$

MODEL'S PRECISION = 0.65

# EXPLORATORY DATA ANALYSIS

# HOME TEAM ADVANTAGE

Does the home team have an advantage over the away team?

# OTHER STATISTICS

Highest Defense Score

Highest Attack Score

Highest Midfield Score

Most all-over wins

# TOP 15 TEAMS WITH THE HIGHEST WINNING PERCENTAGE