Predicting the 2026 FIFA World Cup Winner through Machine Learning Analysis

Hacettepe University
VBM 683 - Machine Learning

Prepared By:
Mert Caliskan

Introduction

- FIFA World Cup Overview
  - One of the most prominent international football tournaments
  - Played every four years since 1930
- Recap of the 2022 FIFA World Cup
  - Location: Qatar
  - Final: Argentina vs. France
  - Champion: Argentina
- Upcoming 2026 FIFA World Cup
  - 48 teams
  - No information about groups and qualified teams yet
  - 16 cities among North America countries:
    - Canada
    - Mexico
    - USA

Objectives

The primary objectives are to:

- Utilize machine learning techniques to analyse historical data of international games and develop predictive models for determining the winner of 2026 FIFA World Cup.
- Conduct a comprehensive evaluation of the developed prediction models to identify and select the best-performing model, including the assessment of performance metrics.

Paul, the Octopus

From Tentacles to Terabytes

- Predicted 12 out of 14 games correctly. Accuracy = ~85.7%.

Data Collection

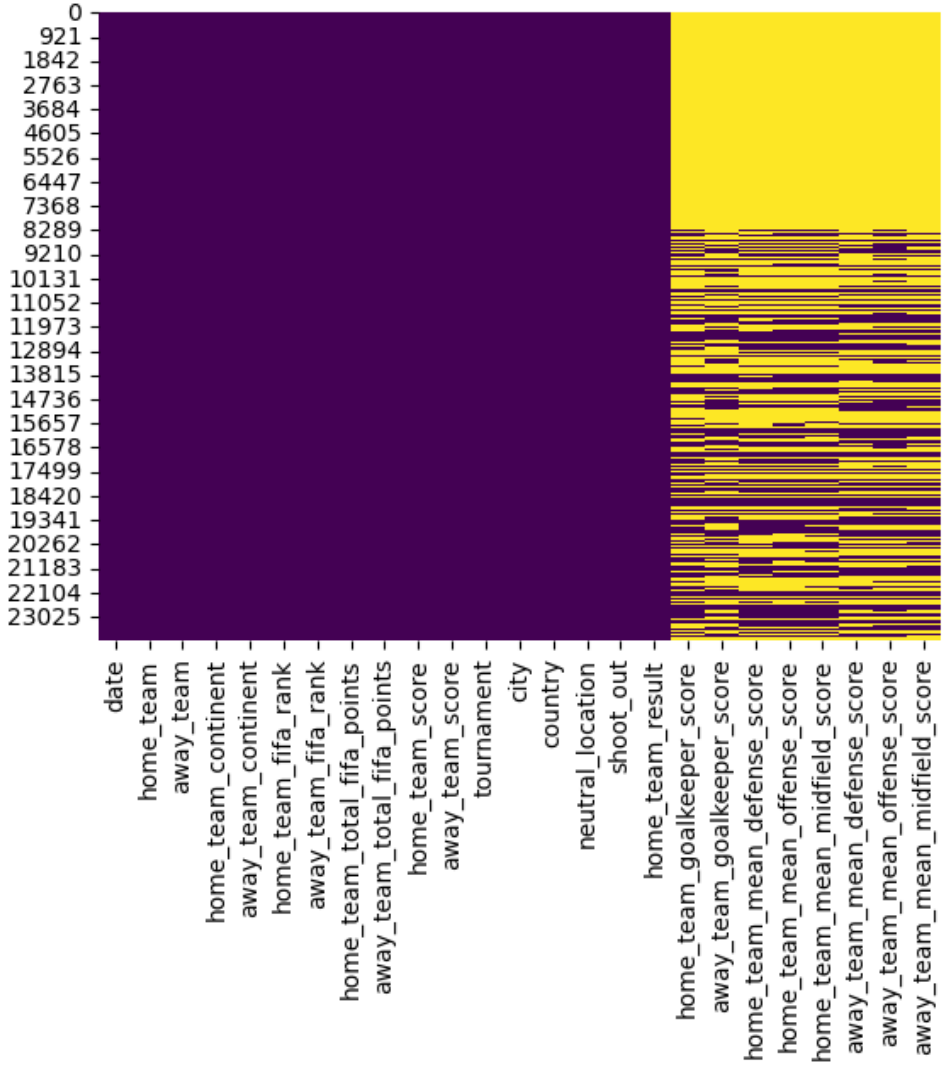
Data Source

- Retrieved from kaggle
  (https://www.kaggle.com/datasets/brenda89/fifa-world-cup-2022/data)
- International games from August 1993 to June 2022
- 23921 games (rows) x 25 features (columns) including:
  - Date
  - Teams and Their Continents
  - Fifa Ranks
  - Total FIFA Points
  - Goals Scored
  - Tournament
  - Location
  - Penalty Shootouts
  - Result
  - Average Player Position Strength

Data Collection

Data Preprocessing

- Checked for Duplicate Data
- Transformed Data Types
- Checked Missing Values
  - Heatmap Visualization



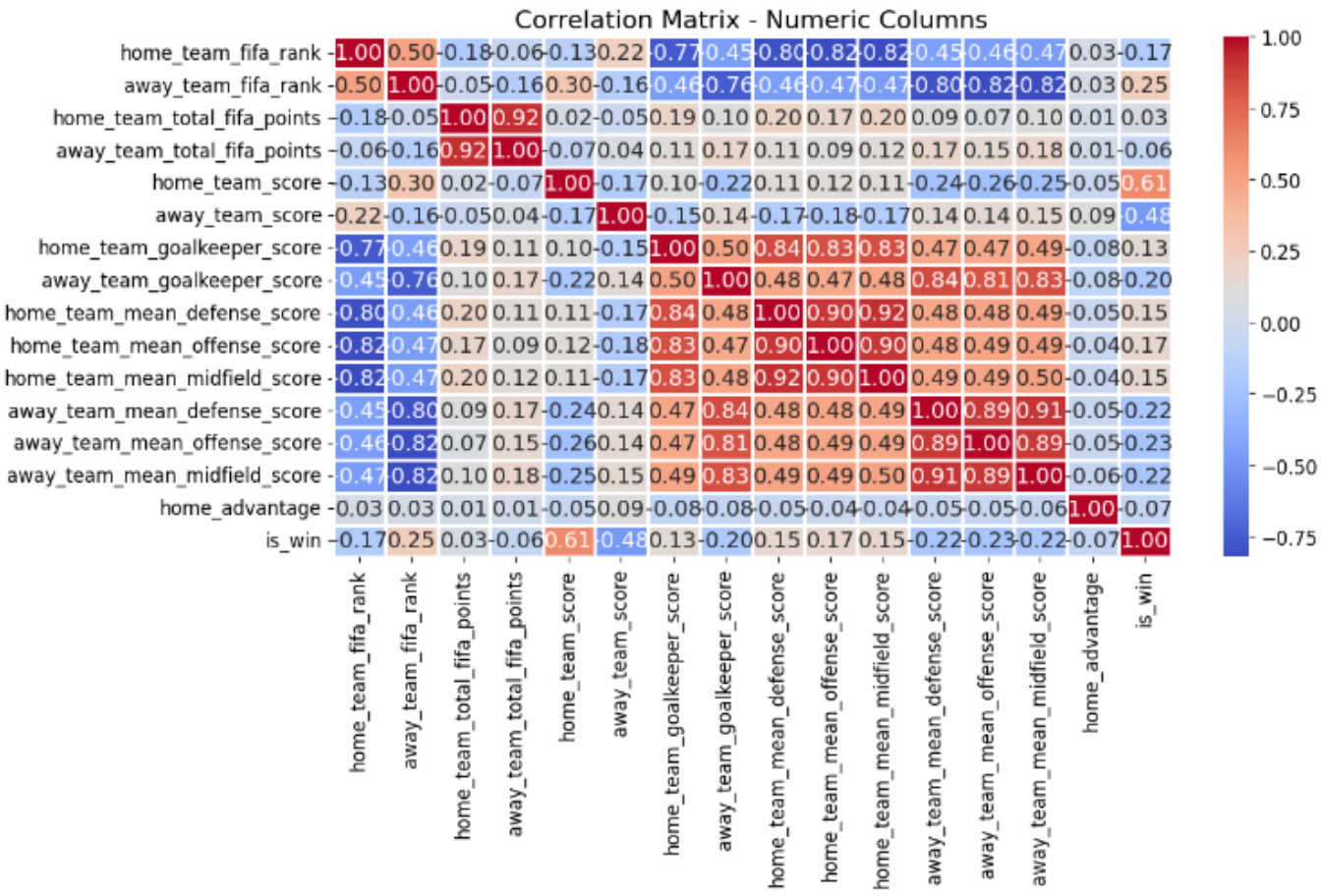
Heatmap Visualization of the Dataset

Data Collection

Feature Engineering

- New Features Added
  - Home Advantage (Binary)
  - Is Win? (Binary)

- Correlation Matrix Examined

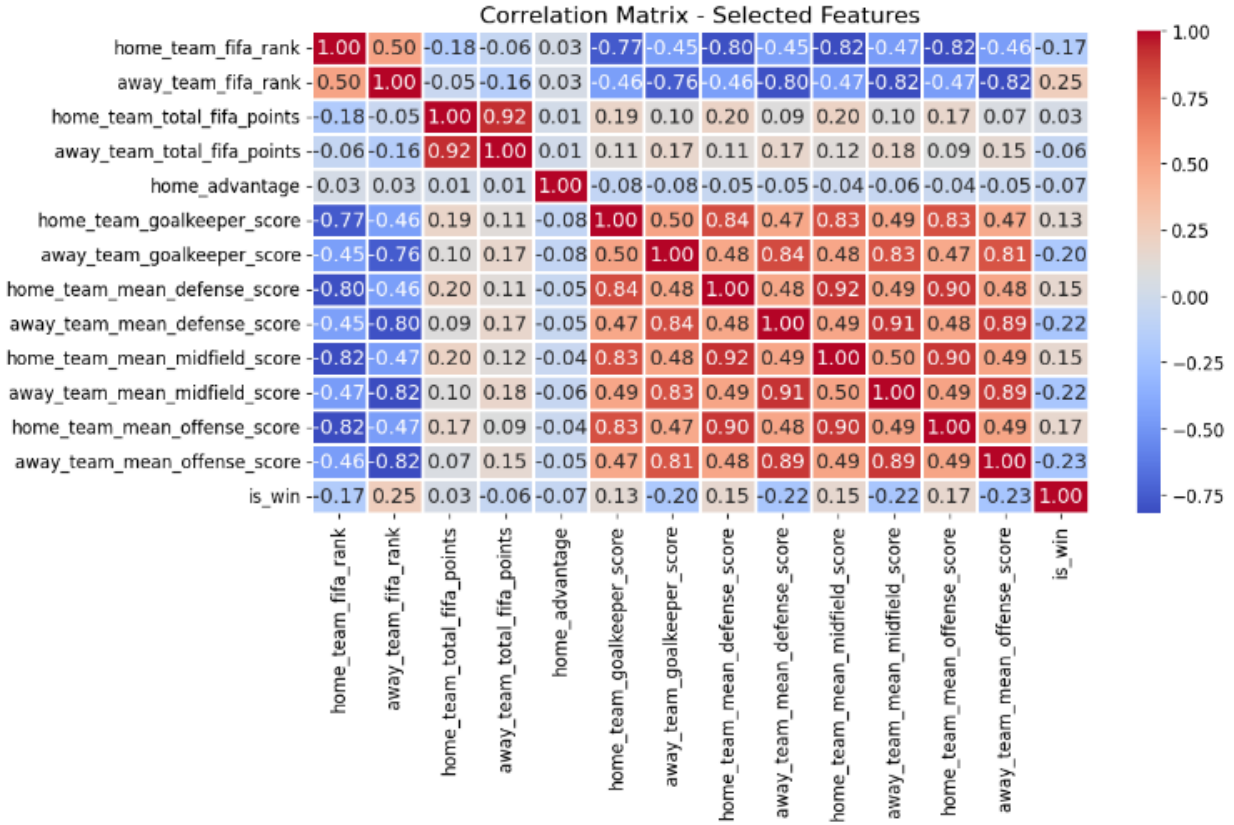


Correlation Matrix of Numerical Features

Data Collection

Feature Engineering

- Following Features were Selected:
  - Home and Away Team FIFA Ranks
  - Home and Away Team Total FIFA Points
  - Home Advantage
  - Home and Away Team Position Scores

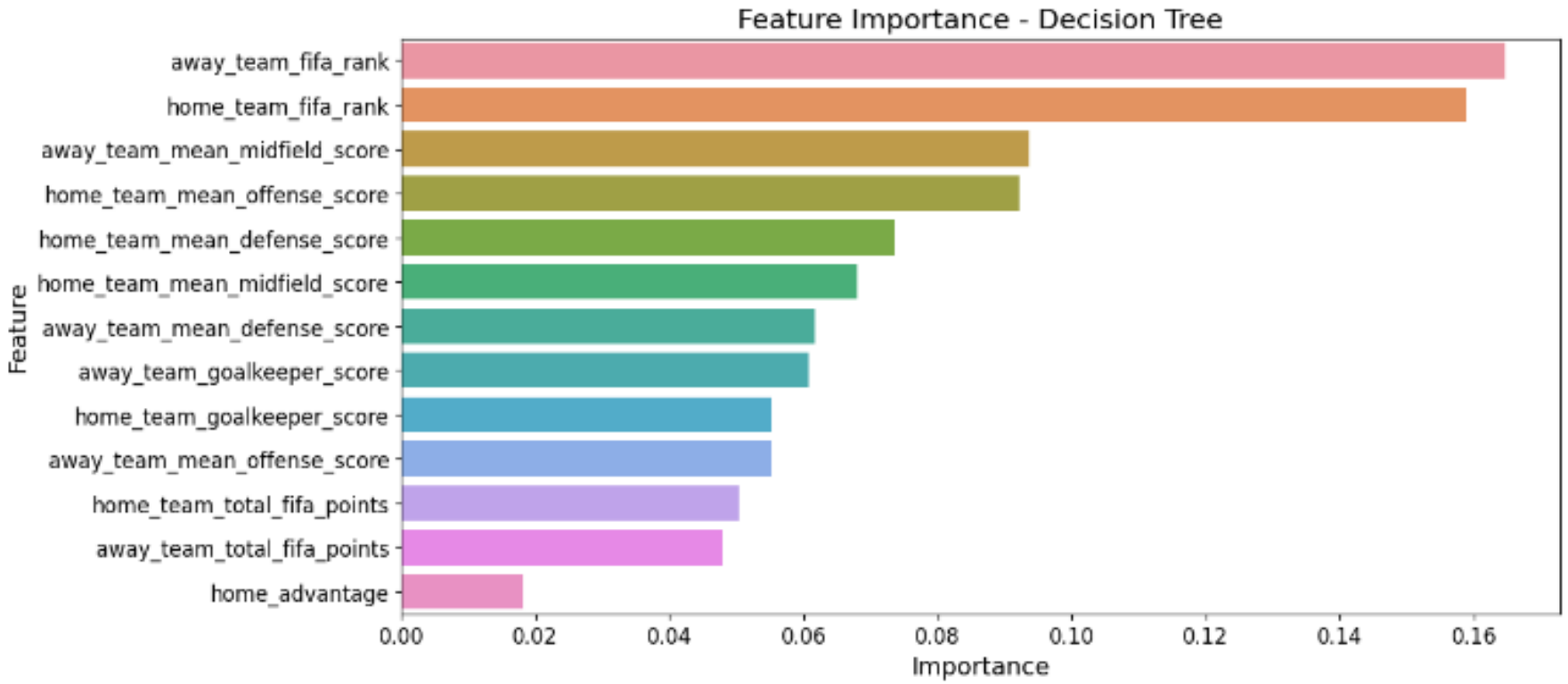


Correlation Matrix of Selected Features

Data Collection

Feature Engineering

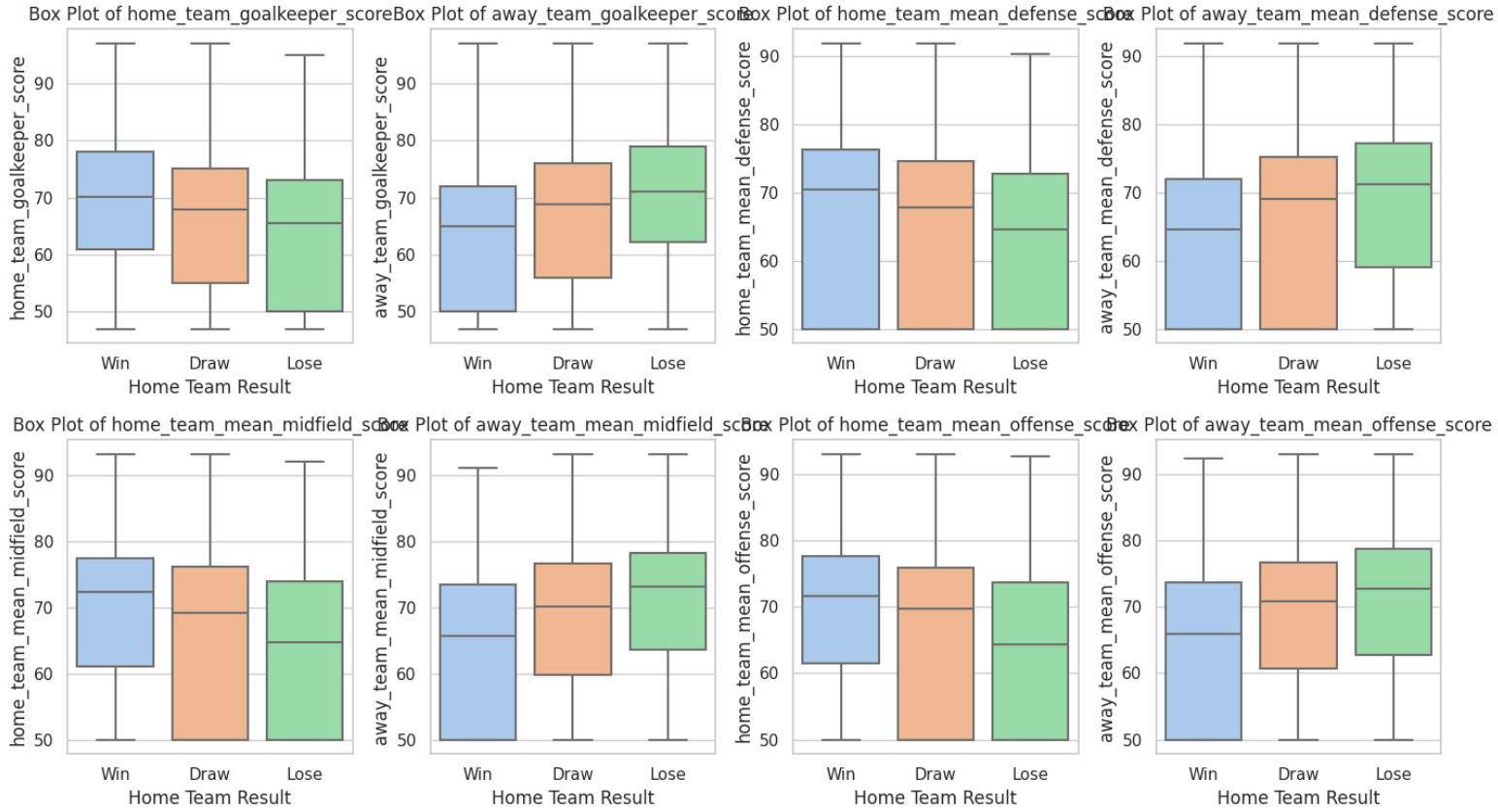
- Feature Importance for Decision Tree



Data Collection

Feature Engineering

- Boxplots
  - Each Player Position Score vs. Home Team Result

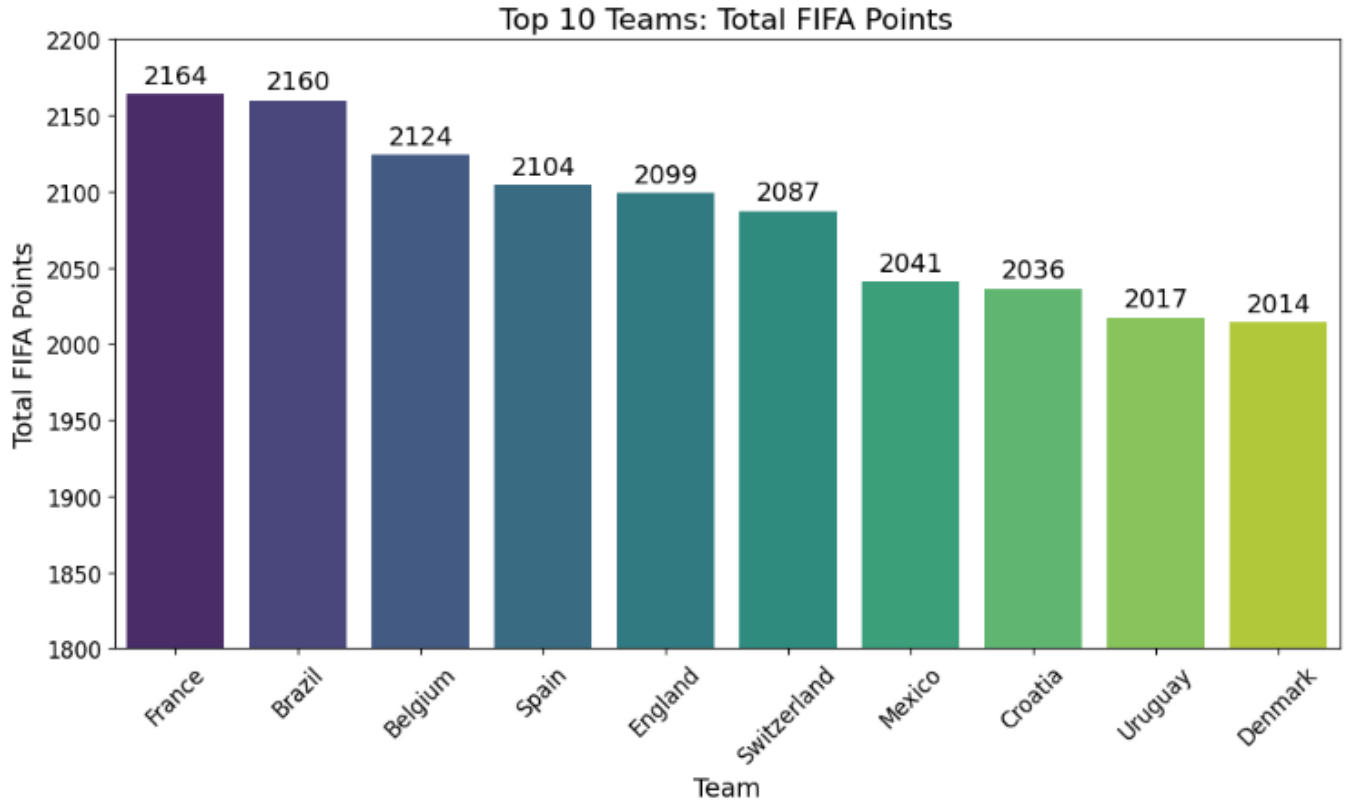


Exploratory Data Analysis

- Top 10 Teams with Highest Total FIFA Ranks
- Top 10 Teams with Highest Player Position Scores
- Top 10 Teams with Highest Win Rates
- Distribution of Home Team Results

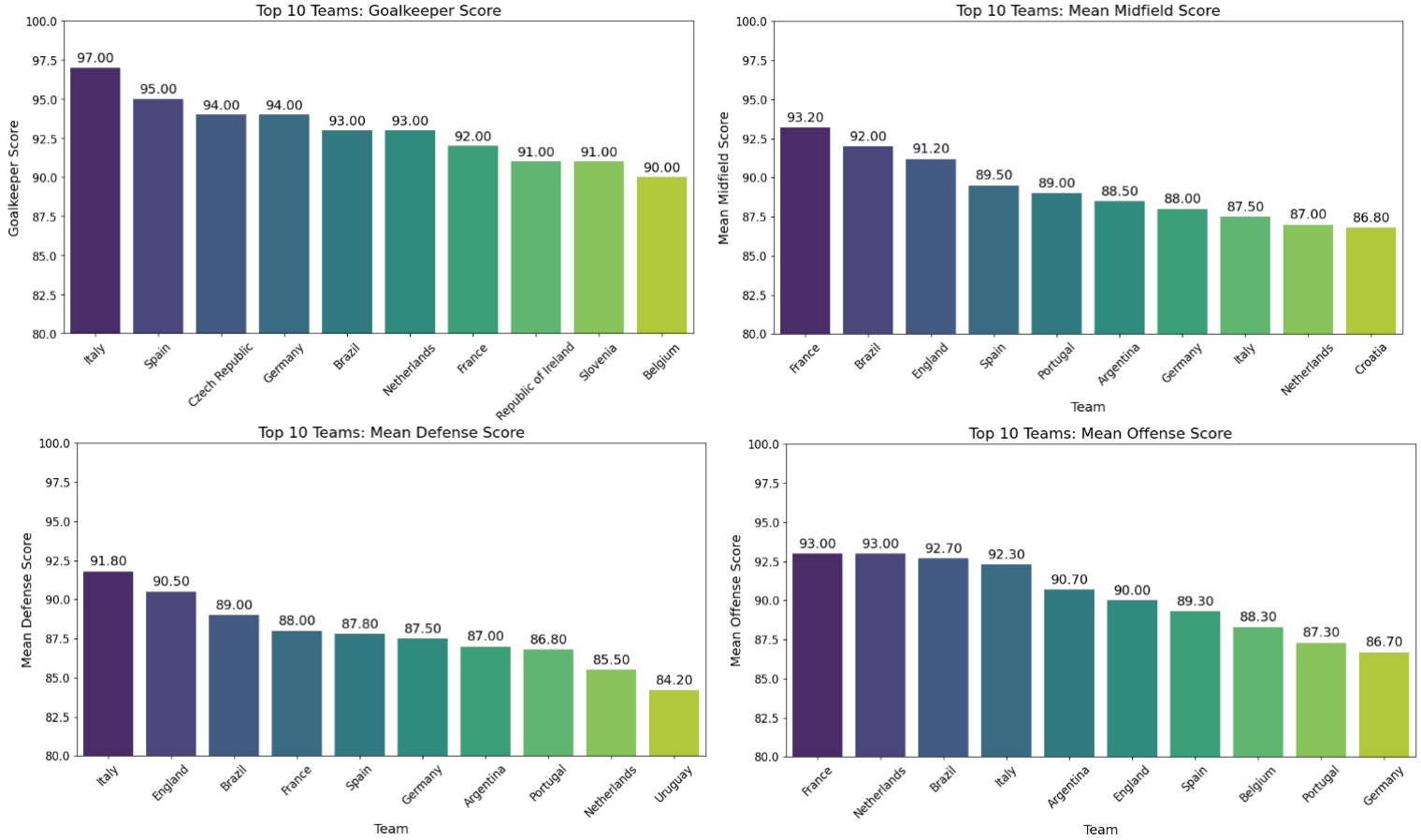
Exploratory Data Analysis

- Top 10 Teams with Highest Total FIFA Ranks



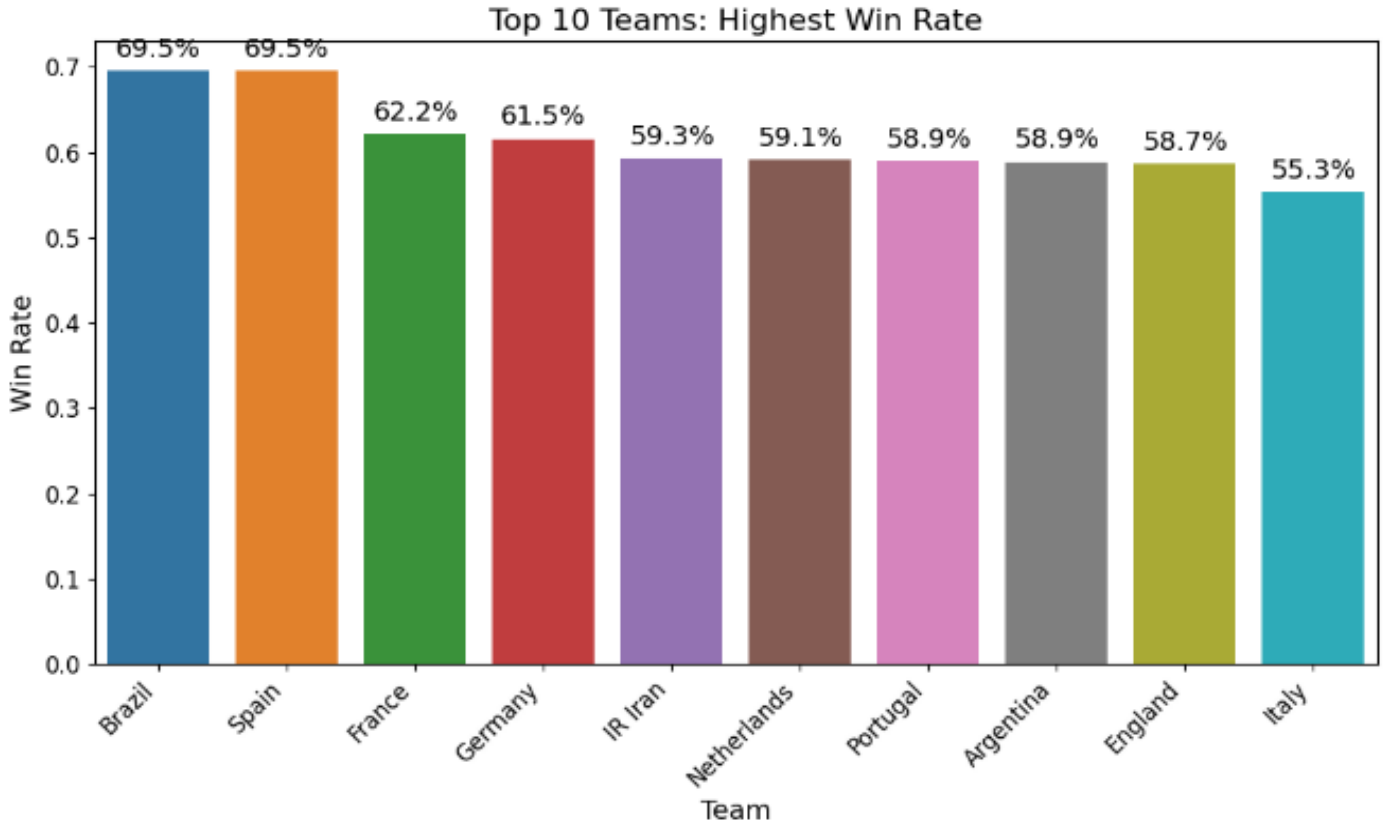
Exploratory Data Analysis

- Top 10 Teams with Highest Player Position Scores



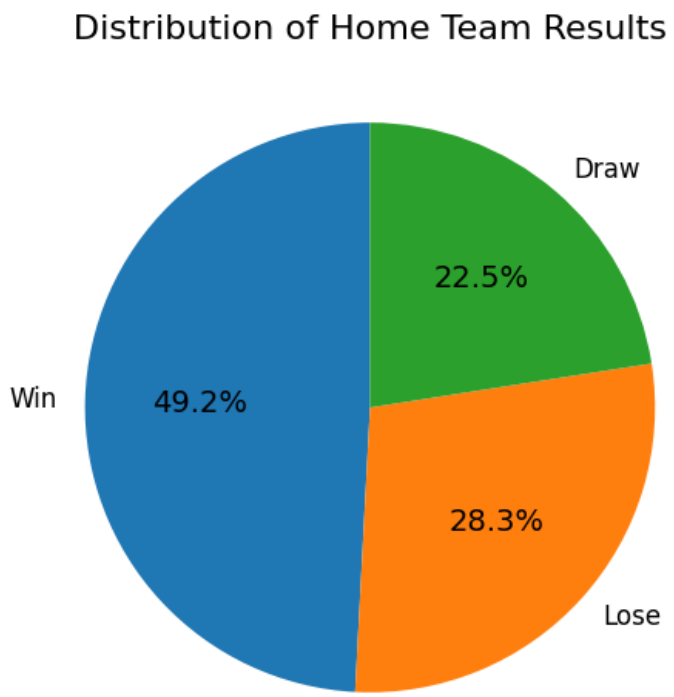
Exploratory Data Analysis

- Top 10 Teams with Highest Win Rates



Exploratory Data Analysis

- Distribution of Home Team Results



Prediction Model

Methodology

- Classification Models (sklearn)
  - Decision Tree
  sklearn.tree.DecisionTreeClassifier
  - Neural Network
  sklearn.neural_network.MLPClassifier
  - Bayes Classifier
  sklearn.naive_bayes.GaussianNB
  - Support Vector Machines (SVM)
  sklearn.svm.SVC
  - Deep Learning
  sklearn.neural_network.MLPClassifier

- Dataset Training & Testing (train_test_split)
  - 80% for training
  - 20% for testing

Prediction Model

Methodology

- Performance Metrics
  - Confusion Matrix
  - Accuracy, Precision, Recall, F-Measure
  - Precision vs. Recall Curve
  - Receiver Operating Characteristic (ROC) Curve

Prediction Model

Results & Discussion

- Confusion Matrix

TABLE I. CONFUSION MATRIX – DECISION TREE MODEL

		Predicted	
		Win	*Not Win*
Actual	*Win*	1422	950
	Not Win	896	1517

TABLE II. CONFUSION MATRIX – NEURAL NETWORK MODEL

		Predicted	
		Win	*Not Win*
Actual	*Win*	1927	445
	Not Win	1117	1296

TABLE I. CONFUSION MATRIX

		Predicted Class	
		Class = YES	*Class = NO*
Actual Class	*Class = YES*	TP	FN
	Class = NO	FP	TN

TABLE III. CONFUSION MATRIX – BAYES CLASSIFIER MODEL

		Predicted	
		Win	*Not Win*
Actual	*Win*	1622	750
	Not Win	767	1646

TABLE IV. CONFUSION MATRIX – SUPPORT VECTOR MACHINES MODEL

		Predicted	
		Win	*Not Win*
Actual	*Win*	1620	752
	Not Win	729	1684

TABLE V. CONFUSION MATRIX – DEEP LEARNING MODEL

		Predicted	
		Win	*Not Win*
Actual	*Win*	864	1508
	Not Win	245	2168

Prediction Model

Results & Discussion

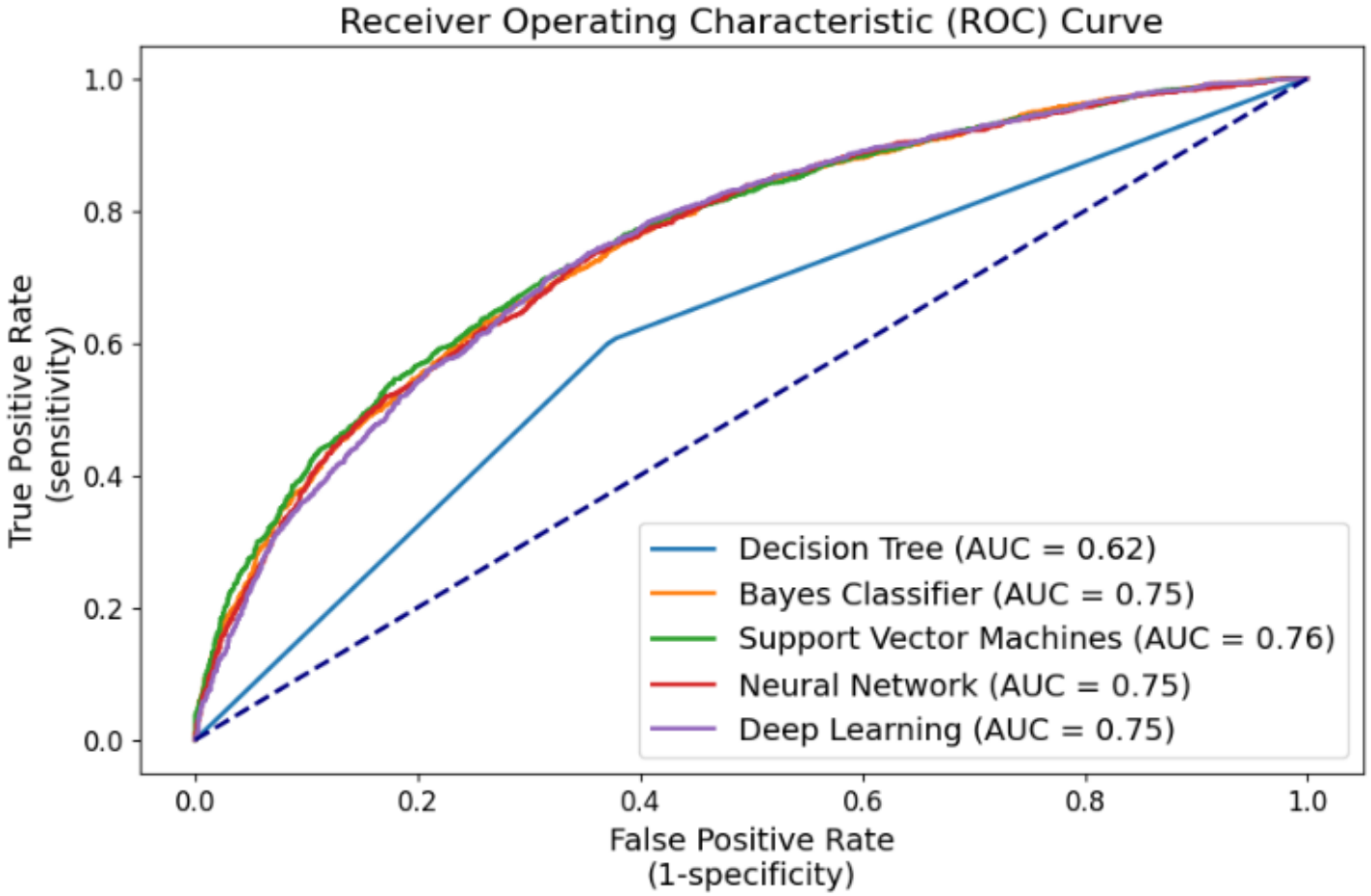
- Accuracy, Precision, Recall, F-Measure

Model	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Decision Tree	60.86	60.69	59.70	60.19
Neural Network	66.96	70.03	58.31	63.63
Bayes Classifier	68.30	67.89	68.38	68.14
Support Vector Machines	69.05	68.97	68.30	68.63
Deep Learning	69.13	70.92	63.95	67.26

Prediction Model

Results & Discussion

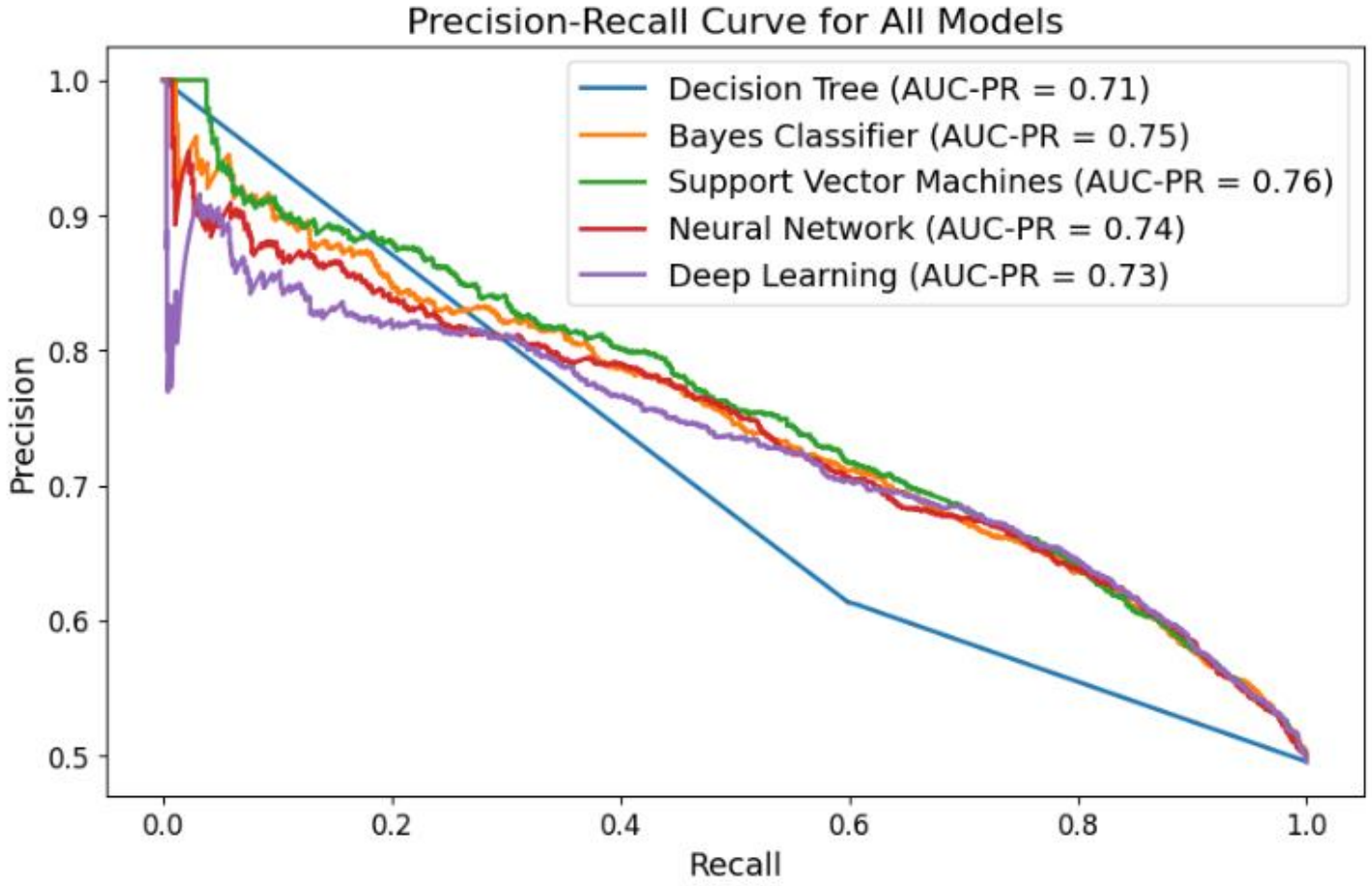
- Receiver Operating Characteristic (ROC) Curve



Prediction Model

Results & Discussion

- Precision vs. Recall



Prediction Model

Results & Discussion

- Support Vector Machines (SVM) model
  - Accuracy = 69.05%
  - Precision = 68.97%
  - Recall = 68.30%
  - F-Measure = 68.63%
  - AUC = 76%
  - AUC-PR = 75%

  Overall – Balanced Performance!

Simulation

Methodology

- Prediction model created based on SVM
- Assumption: Top 48 teams based on their total FIFA points qualified for the 2026 FIFA World Cup.


France

Turkiye

Belgium

Uruguay

Spain

Poland

Brazil

Italy

England

Switzerland

Colombia

Peru

Mexico

Croatia

Denmark

Sweden

Japan

Morocco

Germany

Argentina

Russia

Portugal

USA

Serbia

Netherlands

Wales

Senegal

Chile

IR Iran

Ukraine

Austria

Iraq

Tunisia

Korea R.

Algeria

Czech R.

Nigeria

Costa Rica

Romania

Egypt

Canada

Northern Ireland

Australia

Slovakia

Paraguay

Cameroon

Venezuela

Rep. of Ireland

Simulation

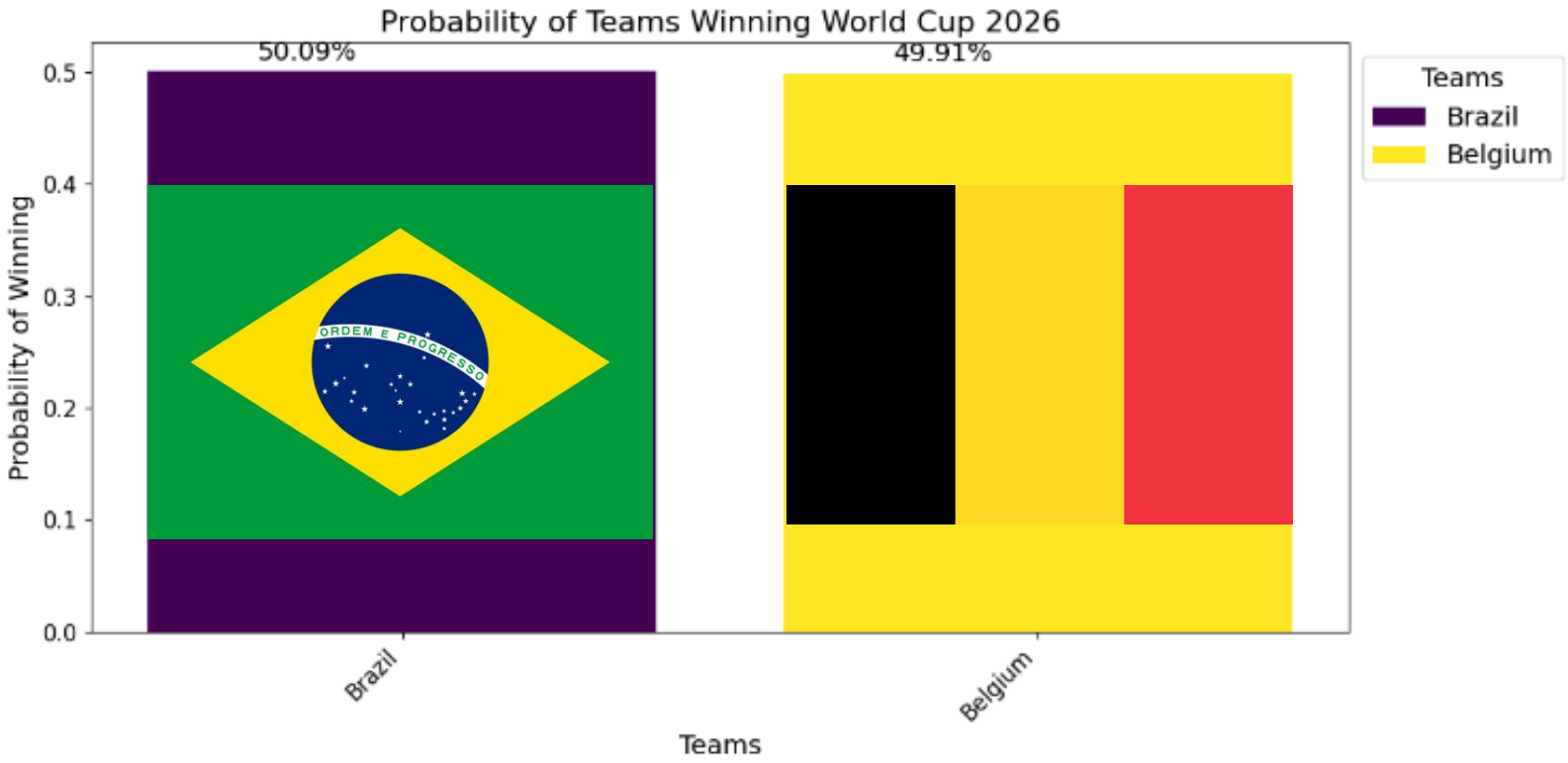
Methodology

- Number of iterations = 10,000
- Random shuffling of teams at each stage to consider uncertainty
- Stages:
  - Group Stage
  - Round of 32
  - Round of 16
  - Quarterfinals
  - Semifinals
  - Final

Simulation

Results & Discussion

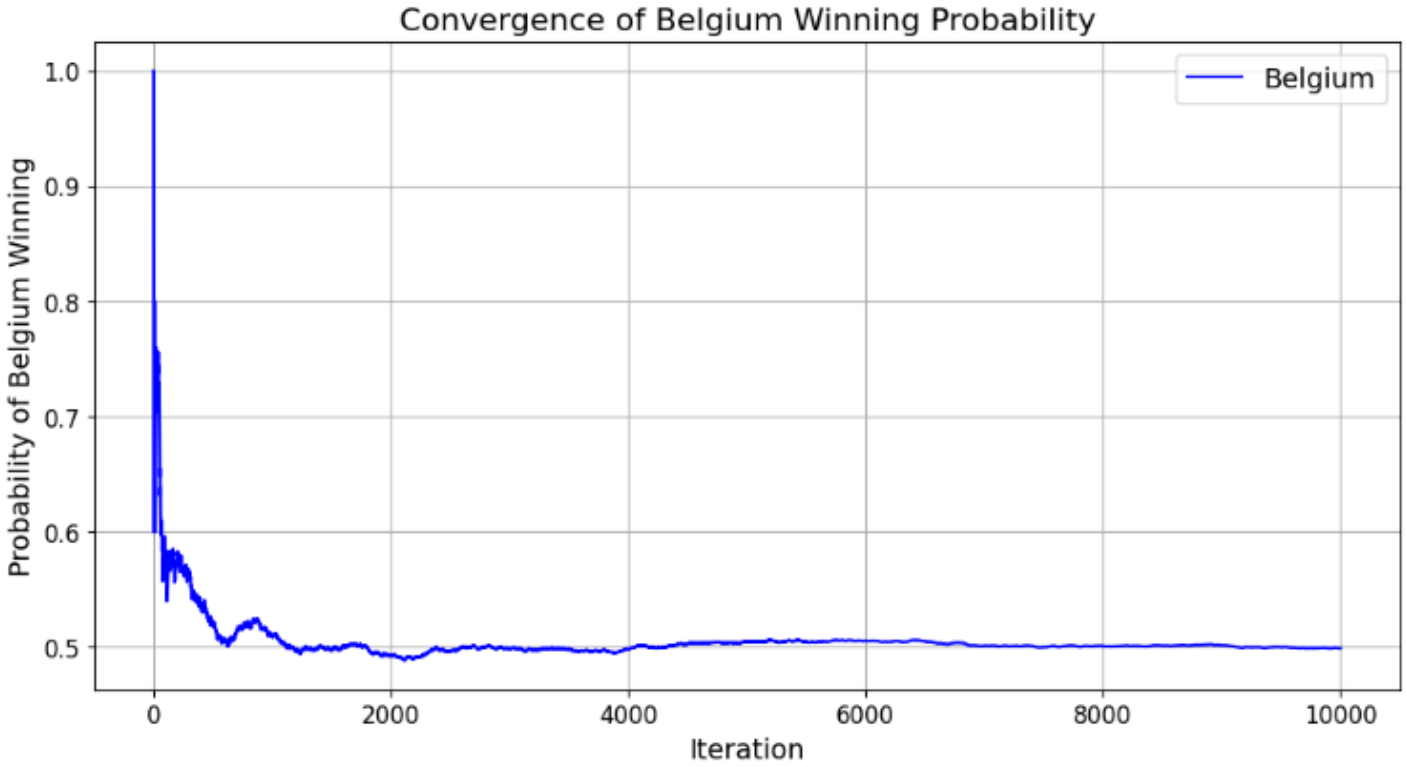
- Brazil: 50.09%
- Belgium: 49.91%



Simulation

Results & Discussion

- Convergence Graph



Conclusion

- Objective:
  - To predict the winner of 2026 FIFA World Cup according to historical data using machine learning techniques
- Steps Implemented:
  - Data Collection and Preprocessing
  - Exploratory Data Analysis
  - Evaluation of Prediction Models
    - SVM Selected for Balanced Performance
  - Simulation of the Tournament
    - Random Shuffling with 10,000 iterations
    - Competitive Results Between Brazil (50.09%) and Belgium (49.91%)

Future Work

- Addressing Missing Data
- Enhanced Data Exploration and Feature Engineering
- Exploring other Machine Learning Classification Models
- Hyperparameter Tuning
- Inclusion of Year Information
- Considering Games that Resulted in Draw
- Excluding Friendly Games from the Dataset
- Revisiting the Simulation After the Qualified Teams and Groups are Officially Published

THANK YOU!

#This is not an investment advice!