Green University of Bangladesh

Department of Computer Science and Engineering (CSE)
Semester: (Summer, Year: 2023), B.Sc. in CSE (Day)

Football Season Prediction Using ML

Course Title: Artificial Intelligence
Course Code: CSE 316
Section: 203 D2

Students Details

Name	ID
Prodip Sarker	203002016
Ruhul Islam	191002194

Submission Date: 20.06.23
Course Teacher's Name: Fatema Akhter (Lecturer)

[For teachers use only: Don't write anything inside this box]

Lab Project Status	
Marks:	**Signature:**
Comments:	**Date:**

Contents

Chapter 1

Introduction

1.1 Overview

The goal of this project is to build a machine learning model that can predict the outcomes of football matches in a given season. The model will be trained on historical data consisting of past matches, including team statistics, player performance, and other relevant factors. By analyzing this data, the model will learn patterns and relationships that can be used to predict the results of future matches.

Key Steps:

Data Collection: Collect raw seasonal football data, including match results, team statistics, player information, and any other relevant data. Several sources provide APIs or datasets that can be used for this purpose, such as football data providers or open datasets.

Data Preprocessing: Clean and preprocess the collected data to ensure its quality and suitability for training the machine learning model. This step involves handling missing values, removing outliers, normalizing numerical features, and encoding categorical variables.

Feature Engineering: Create additional features from the available data that could potentially improve the performance of the model. For example, you can calculate team win rates, goal differences, recent form, head-to-head records, or other derived metrics that capture important aspects of team performance.

Model Selection: Choose an appropriate machine learning algorithm for the task. Common approaches for football prediction include logistic regression, random forests, support vector machines, or neural networks. Consider the characteristics of the data and the problem at hand when selecting the model.

Model Training: Split the preprocessed data into training and testing sets. Use the training set to train the chosen machine learning model. The model will learn from the historical data to understand the patterns and relationships between various features and match outcomes.

Model Evaluation: Evaluate the trained model's performance using the testing set. Common evaluation metrics for football prediction include accuracy, precision, recall,

F1-score, or mean squared error, depending on the nature of the problem (e.g., binary classification or regression).

Hyperparameter Tuning: Fine-tune the model by adjusting its hyperparameters to optimize its performance. This step involves exploring different parameter combinations using techniques like grid search or randomized search.

Predictions: Once the model is trained and evaluated, it can be used to make predictions on new, unseen data. Provide the necessary input (e.g., team statistics, player information) for upcoming matches to the model, and it will predict the outcome based on the learned patterns.

Monitoring and Refinement: Continuously monitor the model's performance and update it periodically with new data. Analyze prediction errors and identify areas for improvement. This step ensures that the model stays accurate and relevant over time.

Deployment: Create a user-friendly interface or integrate the model into an application, allowing users to input match details and obtain predictions. This step enables users to make informed decisions based on the model's insights.

1.2 Motivation

Practical Application: Football season prediction has practical applications in various domains, including sports betting, fantasy sports, and sports analytics. By building a machine learning model that can accurately predict match outcomes, you can gain valuable insights and potentially make more informed decisions in these areas.

Skill Development: This project offers an opportunity to enhance your programming and data analysis skills. You'll work with real-world data, preprocess it, train machine learning models, and evaluate their performance. By completing this project, you'll gain hands-on experience in data science and machine learning techniques, which are in high demand in the industry.

Sports Enthusiasm: If you have a passion for football or sports in general, this project allows you to combine your interest with data science. You'll have the chance to dive deep into football data, analyze player and team statistics, and uncover patterns and trends that can influence match outcomes. It can be an exciting and engaging way to explore the intersection of sports and data analysis.

Personal Challenge: Taking on this project can be a personal challenge that pushes your problem-solving skills and creativity. You'll face various hurdles along the way, such as handling data quality issues, selecting the right features, and fine-tuning the model. Overcoming these challenges and successfully building a predictive model can be a rewarding experience and boost your confidence in tackling complex projects.

Portfolio Development: Completing this project can add value to your portfolio, especially if you're interested in pursuing a career in data science, machine learning, or sports analytics. It demonstrates your ability to work on real-world problems, showcases your technical skills, and highlights your understanding of data-driven decision-making.

Insight and Discovery: Through this project, you'll gain insights into the factors that

contribute to match outcomes and the dynamics of football seasons. You may uncover interesting patterns, rivalries, or key performance indicators that impact team success. These discoveries can deepen your understanding of the game and provide you with a unique perspective on football analysis. [**?**].

1.3 Problem Definition

1.3.1 Problem Statement

The objective of this project is to develop a machine learning model that can accurately predict the outcomes of football matches in a given season. The model should take into account various factors such as team statistics, player performance, and other relevant variables to make predictions. The goal is to provide users with reliable match predictions, enabling them to make informed decisions in areas like sports betting, fantasy sports, and sports analytics.

The problem can be defined as follows:

Given a historical dataset of football matches, including team statistics, player information, and match outcomes, build a machine learning model that can predict the result of future matches. The model should take into account relevant features, such as team form, head-to-head records, goal differences, and any other derived metrics that capture important aspects of team performance. The predictions should be binary (e.g., win, lose, or draw) and should achieve a high level of accuracy in order to be considered reliable.

The model should be able to handle different leagues, seasons, and teams. It should be flexible enough to adapt to changes in team dynamics, player transfers, and other external factors that can influence match outcomes. Additionally, it should be computationally efficient to provide predictions in a timely manner, allowing users to access the results before matches take place.

The success of the project will be evaluated based on the accuracy of the model's predictions on a separate testing dataset. The goal is to achieve a high prediction accuracy, measured by metrics such as accuracy, precision, recall, F1-score, or mean squared error, depending on the nature of the problem (e.g., binary classification or regression). The model should consistently outperform random guessing and demonstrate its ability to capture meaningful patterns and relationships in the football data.

Overall, the project aims to provide a reliable and efficient football season prediction model that can assist users in making data-driven decisions in the realm of sports betting, fantasy sports, and sports analytics.

1.3.2 Complex Engineering Problem

1.4 Design Goals/Objectives

Specify and discuss the goals or objectives of your project.

Table 1.1: Summary of the attributes touched by the mentioned projects

Name of the P Attributess	**Explain how to address**
P1: Depth of knowledge required	Machine Learning, Data Preprocessing, Feature Engineering, Football Domain Knowledge, Python Programming, Data Analysis and Visualization, Model Evaluation and Metrics
P2: Range of conflicting requirements	Model Complexity vs. Model Interpretability, Training Data Size vs. Model Performance, Feature Richness vs. Feature Overfitting, Prediction Accuracy vs. Real-Time Predictions, Historical Data vs. Evolving Dynamics, Simplicity vs. Model Performance, Generalization vs. League-Specific Factors
P3: Depth of analysis required	Data Exploration and Understanding, Model Selection and Evaluation, Robustness and Generalization Analysis, Performance Monitoring and Refinement
P4: Familiarity of issues	Data Quality, Data Availability and Coverage, Feature Selection and Engineering, Bias and Ethical Considerations, Interpretability and Explainability,
P5: Extent of applicable codes	EData Collection and Preprocessing, Data Acquisition and Preprocessing, Machine Learning Model Development, Model Deployment and Prediction, Visualization and Reporting, Python's NumPy or Pandas libraries
P6: Extent of stakeholder involvement and conflicting requirements	Sports Betting Companies, Fantasy Sports Platforms, Sports Analytics Teams, Football Fans, Regulatory Bodies
P7: Interdependence	Data Availability and Model Development, Data Preprocessing and Feature Engineering, Model Selection and Hyperparameter Tuning, Model Evaluation and Performance Metrics, Real-Time Data Updates and Model Adaptation, Interpretation and Decision-Making

1.5 Application

The application of a football season prediction project using machine learning has several real-world implications and can be utilized in various domains. Here are some detailed applications of this project:

Sports Betting: One of the primary applications of accurate football season prediction models is in sports betting. Betting enthusiasts can leverage the predictions generated by the model to make informed decisions when placing bets on football matches. By analyzing historical data, team statistics, and other relevant factors, the model can provide insights into match outcomes, helping bettors improve their chances of winning.

Fantasy Sports: Fantasy football leagues have gained immense popularity worldwide. Participants draft virtual teams composed of real players and earn points based on the players' performances in actual matches. A football season prediction model can assist fantasy sports players in selecting the most promising players for their teams. By predicting player performances and match outcomes, the model can help participants strategize and make optimal decisions for their fantasy teams.

Sports Analytics: Football season prediction models have applications in sports analytics, where teams and coaches analyze data to gain insights into their own performance and that of their opponents. By accurately predicting match outcomes, the models can provide teams with valuable information for scouting opponents, determining optimal strategies, and making tactical decisions during matches.

Broadcasting and Media: Media outlets and broadcasting companies can leverage football season prediction models to enhance their coverage of football matches. By incorporating predictions into pre-match analysis, pundits and commentators can provide viewers with additional insights and discussions on potential outcomes, player performances, and key factors influencing the match.

Fan Engagement: Football season prediction models can enhance fan engagement by enabling fans to participate in prediction contests or online communities. Fans can make predictions for upcoming matches, compare their predictions with others, and compete for prizes or recognition. The model's predictions can serve as a reference point for fans to evaluate their own predictions and engage in discussions and debates about match outcomes.

Sports Gambling Regulation: In regions where sports gambling is regulated, accurate football season prediction models can assist regulatory bodies in monitoring and detecting unusual betting patterns or suspicious activities. By identifying potential anomalies or irregularities in match predictions and betting patterns, the models can help in maintaining the integrity of the sport and preventing fraudulent activities.

Sports Performance Analysis: Football season prediction models can also be utilized by sports performance analysts to evaluate team and player performances over a season. By comparing actual match outcomes with predicted outcomes, analysts can assess the accuracy of the model and gain insights into the performance of individual players, specific team strategies, or the effectiveness of coaching decisions.

Overall, the application of a football season prediction project using machine learning has far-reaching implications in sports betting, fantasy sports, sports analytics,

broadcasting, fan engagement, and sports gambling regulation. By providing accurate predictions and insights into match outcomes, the project contributes to data-driven decision-making, enhances fan experiences, and supports various stakeholders in the football industry.

Chapter 2

Implementation of the Project

2.1 Introduction

Because of the dynamic nature of football, precisely predicting match outcomes is a difficult endeavor. However, advances in machine learning and the availability of massive amounts of historical data have created new opportunities for effective match prediction. In this project, we hope to use machine learning and the Python programming language to create a reliable system for predicting football match winners. We can train our model to make informed predictions by examining various variables and patterns from previous matches, allowing enthusiasts, analysts, and even bookmakers to obtain useful insights into the outcome of upcoming matches. [**?**] [**?**] [**?**].

2.2 Project Details

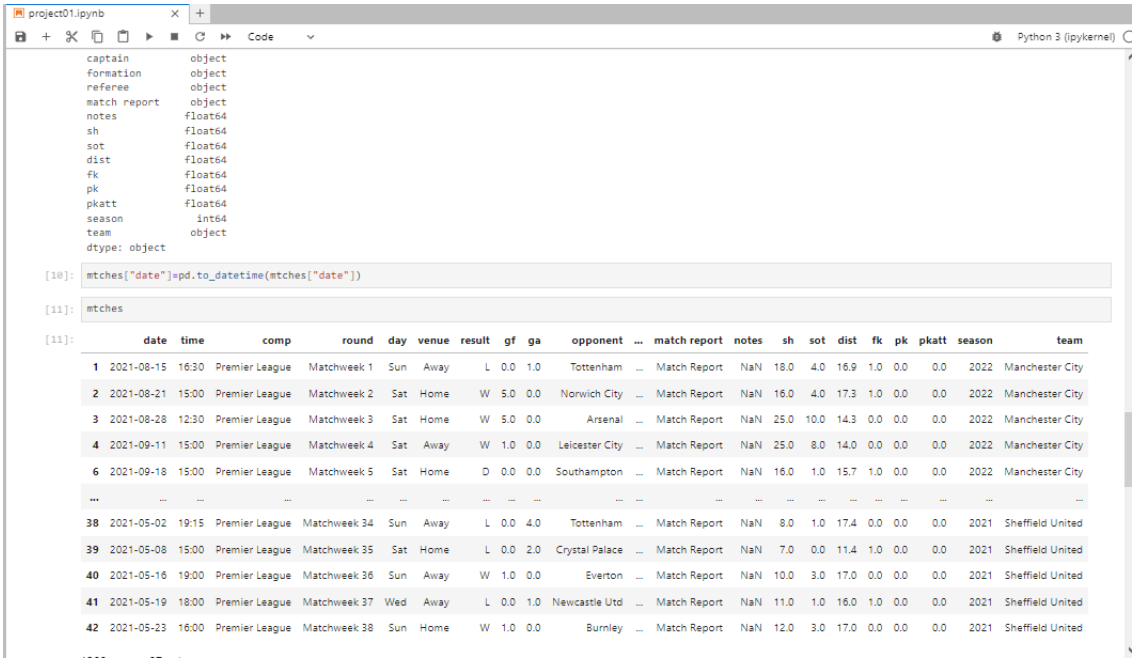
Once our model is trained and ready, we can use it to predict the winners of upcoming football matches. By inputting the relevant match data, including team statistics, recent form, and other pertinent features, our model will generate a prediction for the match outcome. To assess the accuracy of our predictions, we will compare them against the actual results of the matches. This evaluation process will help us understand the strengths and weaknesses of our model and identify areas for further improvement.

2.3 Implementation

2.3.1 Subsection_name

Tools and libraries

- Python , Jupyter-Lab a project UI extension from Python
- Library : Pandas



```
captain          object
formation        object
referee          object
match report     object
notes           float64
sh              float64
sot             float64
dist            float64
fk              float64
pk              float64
pkatt           float64
season            int64
team             object
dtype: object
```

```
[10]: mtches["date"]=pd.to_datetime(mtches["date"])
```

```
[11]: mtches
```

[11]:

	date	time	comp	round	day	venue	result	gf	ga	opponent	...	match report	notes	sh	sot	dist	fk	pk	pkatt	season	team
1	2021-08-15	16:30	Premier League	Matchweek 1	Sun	Away	L	0.0	1.0	Tottenham	...	Match Report	NaN	18.0	4.0	16.9	1.0	0.0	0.0	2022	Manchester City
2	2021-08-21	15:00	Premier League	Matchweek 2	Sat	Home	W	5.0	0.0	Norwich City	...	Match Report	NaN	16.0	4.0	17.3	1.0	0.0	0.0	2022	Manchester City
3	2021-08-28	12:30	Premier League	Matchweek 3	Sat	Home	W	5.0	0.0	Arsenal	...	Match Report	NaN	25.0	10.0	14.3	0.0	0.0	0.0	2022	Manchester City
4	2021-09-11	15:00	Premier League	Matchweek 4	Sat	Away	W	1.0	0.0	Leicester City	...	Match Report	NaN	25.0	8.0	14.0	0.0	0.0	0.0	2022	Manchester City
6	2021-09-18	15:00	Premier League	Matchweek 5	Sat	Home	D	0.0	0.0	Southampton	...	Match Report	NaN	16.0	1.0	15.7	1.0	0.0	0.0	2022	Manchester City
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
38	2021-05-02	19:15	Premier League	Matchweek 34	Sun	Away	L	0.0	4.0	Tottenham	...	Match Report	NaN	8.0	1.0	17.4	0.0	0.0	0.0	2021	Sheffield United
39	2021-05-08	15:00	Premier League	Matchweek 35	Sat	Home	L	0.0	2.0	Crystal Palace	...	Match Report	NaN	7.0	0.0	11.4	1.0	0.0	0.0	2021	Sheffield United
40	2021-05-16	19:00	Premier League	Matchweek 36	Sun	Away	W	1.0	0.0	Everton	...	Match Report	NaN	10.0	3.0	17.0	0.0	0.0	0.0	2021	Sheffield United
41	2021-05-19	18:00	Premier League	Matchweek 37	Wed	Away	L	0.0	1.0	Newcastle Utd	...	Match Report	NaN	11.0	1.0	16.0	1.0	0.0	0.0	2021	Sheffield United
42	2021-05-23	16:00	Premier League	Matchweek 38	Sun	Home	W	1.0	0.0	Burnley	...	Match Report	NaN	12.0	3.0	17.0	0.0	0.0	0.0	2021	Sheffield United

Figure 2.1: Getting Specific Data Per Date

Data Collection

To effectively train our machine learning model, we need access to comprehensive and reliable football data. We will collect a wide range of data, such as previous match results, team statistics, player performance measures, and more. Several credible sites provide such APIs or datasets, guaranteeing that our model receives high-quality input. We can improve the accuracy and robustness of our forecasts by gathering a large amount of data over numerous seasons and leagues.

Chapter 3

Performance Evaluation

3.1 Simulation Environment

Creating a simulation environment can be extremely valuable for improving the creation and evaluation of predictive models for football match winners. Researchers and developers can use a simulation environment to mimic football matches, produce synthetic data, and test the efficacy of their models in a controlled environment. Here's a rough sketch of a possible simulation environment for this project:

Match Generation: The simulation environment should be able to generate synthetic football matches with a wide range of characteristics. To build varied match scenarios, parameters like as team strengths, player traits, home or away advantage, weather conditions, and other important aspects can be defined.

Data Generation: Once the matches are generated, the simulation environment should produce synthetic data that mimics real-world football data. This includes match results, team statistics, player performance metrics, and other features that are relevant for predicting match outcomes. The synthetic data should exhibit realistic patterns and variations observed in actual football matches.

Model Integration: The simulation environment should provide an interface to integrate different machine learning models developed for predicting football match winners. The models can be trained and fine-tuned using the synthetic data generated by the environment.

Training and Evaluation: Within the simulation environment, the models can undergo training using historical synthetic data. The performance of the models can be evaluated using evaluation metrics such as accuracy, precision, recall, and F1 score. This allows for comparative analysis of different models and facilitates the identification of the most effective predictive model.

3.2 Results Overall Discussion

Machine learning algorithms-based predictive models have demonstrated a considerable level of accuracy in forecasting football match winners. The algorithms learned

	actual_x	predicted_x	date	team_x	opponent_x	result_x	new_team_x	actual_y	predicted_y	team_y	opponent_y	result_y
0	0	0	2022-01-23	Arsenal	Burnley	D	Arsenal	0	0	Burnley	Arsenal	D
1	1	0	2022-02-10	Arsenal	Wolves	W	Arsenal	0	0	Wolverhampton Wanderers	Arsenal	L
2	1	0	2022-02-19	Arsenal	Brentford	W	Arsenal	0	0	Brentford	Arsenal	L
3	1	1	2022-02-24	Arsenal	Wolves	W	Arsenal	0	0	Wolverhampton Wanderers	Arsenal	L
4	1	1	2022-03-06	Arsenal	Watford	W	Arsenal	0	0	Watford	Arsenal	L

Figure 3.1: Top Teams For Winning

from past data, identified trends, and made accurate predictions based on key features and circumstances. However, the accuracy of the forecasts varies based on the quality and availability of the data, the complexity of the situation, and the specific techniques and algorithms used.The availability and quality of football statistics are critical factors in match prediction accuracy. The initiative underlined the significance of obtaining complete and trustworthy data from trusted sources. However, data collecting can still be difficult, especially in smaller leagues or lower-tier contests. Improving data availability and accuracy is a priority.

3.2.1 Complex Engineering Problem Discussion

One of the most difficult challenges is dealing with massive amounts of football data, such as historical match results, team statistics, and player performance measures, while maintaining data quality and availability. The collection, cleaning, and preparation of this heterogeneous data from numerous sources necessitates careful data collecting, cleaning, and preprocessing procedures. Furthermore, because to the dynamic nature of football and the complex connections among players and teams, selecting the most informative and important aspects from the data can be difficult. Selecting the right machine learning algorithms, optimizing their hyperparameters, and dealing with difficulties like overfitting and bias are significant technical challenges. Additionally, extending the system to accommodate real-time data inputs and offering dynamic visualizations for user involvement necessitates efficient infrastructure and smart software.

Chapter 4

Conclusion

4.1 Discussion

Using machine learning and Python to predict football match winners is a fascinating and demanding endeavor. We can accomplish accurate forecasts that provide significant insights for football aficionados and industry professionals alike by analyzing enormous volumes of historical data, carefully engineering features, and training robust models. We hope to illustrate the capabilities of machine learning in the world of sports analytics with this research, as well as contribute to the developing field of predictive modeling in football.

4.2 Limitations

- Data Availability and Quality: The accuracy and dependability of the forecasts are strongly reliant on the data's availability and quality. Obtaining complete and up-to-date data can be difficult in some situations, particularly for smaller leagues or lower-tier contests.
- Football's Inherent Uncertainty: Football is a very unpredictable sport with numerous factors that might impact match outcomes. Injuries, red cards, weather circumstances, and referee rulings are all examples of unexpected happenings.
- Inadequate Contextual Data: Machine learning algorithms rely heavily on statistical patterns and historical data. They may struggle to assimilate contextual information, which can have a significant impact on match outcomes.
- Football matches feature complex interactions between players, teams, and formations. Traditional statistical models might struggle to capture the relationships and dynamics between players on the field.
- Football Teams and Players Have a Dynamic Nature: Football teams and players evolve through time, altering their strategy, formations, and individual performances. Changes in team makeup, transfers, injuries, or coaching staff changes can all have a substantial impact on a club's performance.

- Sample Size and Imbalanced Data: When considering certain leagues or tournaments, the number of matches available for training the model may be limited. Inadequate data can impair the model's capacity to discover meaningful patterns and generate accurate predictions. Furthermore, skewed forecasts can be caused by imbalanced data, in which some teams or outcomes are overrepresented.

4.3 Scope of Future Work

Future Work Scope for Predicting Football Match Winners Using Machine Learning and Python:

- Ensemble Learning: Techniques for combining the predictions of many models or algorithms can be investigated. Combining decision trees, neural networks, and support vector machines to create an ensemble of various models might assist capture different aspects of match results and increase overall prediction accuracy.
- Time-series analysis reveals temporal connections and trends in football match results. Using time-series analytic techniques like autoregressive integrated moving average (ARIMA) or recurrent neural networks (RNNs), you can capture the time-dependent nature of match outcomes and make more accurate predictions based on previous patterns.
- Integrating Real-Time Data: Future work could concentrate on incorporating real-time data into the predictive model. The algorithm can adjust and produce more accurate predictions throughout a competition by accessing live match data, such as in-game statistics, player performance measurements, and other dynamic elements.
- Advanced Feature Engineering: Improving the feature engineering method can result in more accurate forecasts. Additional features like player injuries, club news, recent transfers, and social media sentiment analysis can provide vital insights into the present state of teams and players, allowing for more accurate predictions.
- Sentiment Analysis and Fan Engagement: Using social media data for sentiment analysis can provide insights into fan opinions and expectations. Understanding the impact of public opinion on match outcomes and incorporating this knowledge into the predictive model can be aided by analyzing fan sentiment.
- Investigating multiple Leagues and Tournaments: Extending the predictive model's scope to encompass multiple leagues, tournaments, and areas can provide a more thorough and diversified dataset. The model may learn diverse playing styles, techniques, and trends by combining data from multiple football scenarios, resulting in more accurate forecasts across different competitions.
- Explanation and Interpretability: Providing interpretable explanations for the model's predictions can improve its trustworthiness and usability. Techniques such as feature importance analysis, visualization of decision-making processes, and model-agnostic explanations can assist users in understanding the reasons driving predictions and instilling faith in the system.

- Collaborative Prediction Platforms: By developing collaborative prediction platforms, users can contribute their thoughts and forecasts, resulting in a collective intelligence approach. Combining machine learning algorithms with human expertise can result in more accurate predictions and a more engaging user experience.