SVKM’s NMIMS
Mukesh Patel School of Technology Management & Engineering
A.Y. 2022 - 23
Course: Machine Learning

Project Report

Program	B.Tech (Artificial Intelligence)	
Semester	IV	
Name of the Project:	Football Prediction using Random Forest Classifier	
Details of Project Members		
Batch	Roll No.	Name
B1	I007	Bhavya Bavishi
B1	I012	Nehaal Choudhary
B1	I020	Aum Ghag
Date of Submission: 10/04/2023		

Contribution of each project Members:

Roll No.	Name:	Contribution
I007	Bhavya Bavishi	Applied Machine Learning
I012	Nehaal Choudhary	Data Analysis
I020	Aum Ghag	Webscrapping

Note:

1. Create a readme file if you have multiple files
2. All files must be properly named (N004_MLProject)
3. Submit all relevant files of your work

   Report, ipynb, pdf, dataset, any other files)
4. **Plagiarism is highly discouraged (Your report will be checked for plagiarism)**

Project Report

Football Prediction using Random Forest Classifier

by

Bhavya Bavishi, Roll number: I007

Nehaal Choudhary, Roll number: I012

Aum Ghag, Roll number: I020

Course: Machine Learning

AY: 2022-23

Table of Contents

I. Project Idea and applications

This project demonstrates the prediction of Football matches of the English Premier League. This project is trained on all the data from 2000-2023. This project can be used to predict the upcoming football games with the required parameters.

II.Dataset details

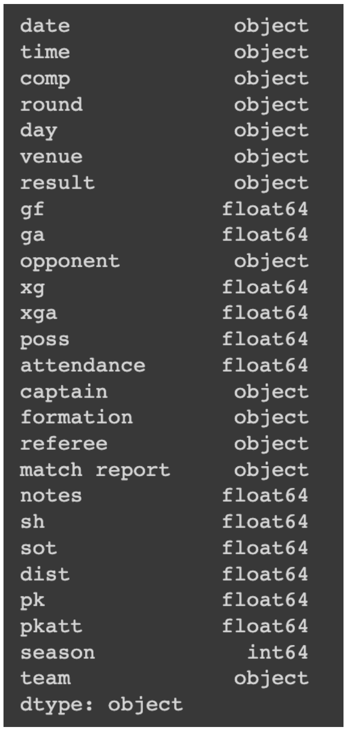
Describe the following:

1. How did you acquire the dataset
2. Features and Meaning of the features of the dataset
3. Size of the dataset
4. Any other important factors

1. We have used webscrapping to create the dataset from scratch. We have used the link given below and scrapped the data from there and created a pandas dataframe.

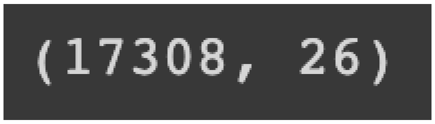
https://fbref.com/en/comps/9/Premier-League-Stats

2.



```
date             object
time             object
comp             object
round            object
day              object
venue            object
result           object
gf              float64
ga              float64
opponent         object
xg              float64
xga             float64
poss            float64
attendance      float64
captain          object
formation        object
referee          object
match report     object
notes           float64
sh              float64
sot             float64
dist            float64
pk              float64
pkatt           float64
season            int64
team             object
dtype: object
```

3.



```
(17308, 26)
```

III. Preprocessing and Visualization

Describe the following:

1. Preprocessing steps with proper justification
2. Visualization – Tools, and inferences

1. Preprocessing includes the cleaning of data which was more difficult than usual since we created the dataset from webscrapping. Firstly, we dropped all the rows which were not required. We then changed the datatype of numeric rows from object to int64 or float64. Then we used label encoding to classify W as 1 and L as 0.

2. Visualization is an important tool in machine learning that helps in understanding and interpreting the data, models, and results. Visualization tools enable the representation of complex data in an intuitive and visual form, making it easier to identify patterns, trends, and anomalies.

IV.Model Creation

Describe the following:

1. Machine learning techniques used with proper justification
2. Train test size

1. Random Forest is a machine learning algorithm that is widely used for classification, regression, and feature selection tasks. It is a type of ensemble learning method that combines multiple decision trees to create a forest. Each decision tree is created using a random subset of the features and training data. The main advantage of the Random Forest algorithm is that it reduces overfitting and increases accuracy by combining multiple decision trees. It also provides important feature selection information by ranking the importance of each feature based on the amount of information it contributes to the classification.

2. In machine learning, the dataset is typically split into two subsets: a training set and a test set. The training set is used to train the machine learning model, while the test set is used to evaluate its

performance on unseen data. The train size included all the football seasons from 2000-2022 and the test data included the 2023 season.

V. Model Evaluation

Describe the following:

1. Evaluation metric chosen and why?

1. Random forest classifier can be used to forecast football game results. The method constructs numerous decision trees, each of which predicts the result of the match using a portion of the features that are available. The forecasts of each individual decision tree are then combined to get the final prediction.

Because it can handle a lot of features and capture complicated relationships between them, the random forest classifier is a well-liked option for forecasting football games. Additionally, compared to certain other machine learning algorithms, it is less prone to overfitting, which might be crucial when working with scant data.

.

VI. Learning from the Project

Include learning from the project:

- How this project helped you?
  Improved data analysis skills
  Increased knowledge of machine learning algorithms
  Improved problem-solving skills
  Enhanced domain knowledge
  Practical Experience
- What new aspects did you learn?
- we learned and implemented web scraping ,its various methods and its application to models.

- we worked with it using beautiful soup algorithm .
- advanced data analysis for bigger size samples which would be more applicable in actual models was learned and prcatised by the team .
- the concept and implementation of rolling average was the key learning in the classification of the data which gave us a better precision score and improved accuracy
- we also learned how modifying the rolling average would give us different accuracies and results ,thus getting the right value to get the best results.

VII. Challenges Faced

The webscrapping part was the most challenging part of the project. We had to use different links to create different tables and then combine those tables into a single dataframe. Then we converted the dataframe into a csv file and used the same to create our model which gave us a precision score 0.6484

VIII. Conclusion

In conclusion, a football prediction project involves using data analysis and machine learning techniques to predict the outcome of football matches. The project requires defining the problem clearly, selecting appropriate features, cleaning and processing data, selecting the appropriate machine learning algorithm, and evaluating the performance of the model.

Through this project, one can develop skills in data analysis, machine learning, problem-solving, and domain knowledge. The project can provide practical experience in developing a machine learning model from start to finish, and the knowledge gained can be applied to a variety of industries and applications beyond football.

Overall, a football prediction project can be a challenging but rewarding undertaking that can enhance one's skills and knowledge in the field of data science and machine learning.

IX. References

"Predicting Football Match Outcomes using Machine Learning Techniques" by João Luís Cardoso, et al. (2018) - This paper describes a machine learning-based approach for predicting football match outcomes using data from the 2016-2017 English Premier League season.

"Football Match Result Prediction Using Machine Learning Techniques: A Comparative Study" by Sahan Bulathwela and Kasun De Zoysa (2020) - This paper compares the performance of different machine learning algorithms for predicting football match outcomes using data from the 2018 FIFA World Cup.

"Football Analytics: Science, Data and Advanced Analytics" by Ian Graham - This book provides an introduction to football analytics and the use of data and advanced analytics to gain insights into player performance, team strategy, and match outcomes.