University of Mumbai

Web Scraping & Prediction Analysis Of A Sports Website

Submitted in partial fulfillment of requirements

for completion of

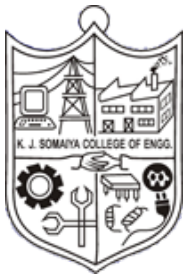
Mini-Project

by

Name: Sayli Bhavsar
Roll No: 1914009

Name: Dhwani Doshi
Roll No: 1914016

Guide: Prof. Pooja Malhotra



Department of Information Technology
K. J. Somaiya College of Engineering, Mumbai-77
(Autonomous College Affiliated to University of Mumbai)

Batch 2022

K. J. Somaiya College of Engineering, Mumbai-77

(Autonomous College Affiliated to University of Mumbai)

Certificate

This is to certify that the report entitled **Web Scraping & Prediction Analysis Of A Sports Website** is bonafide record of Mini-Project work done by **Sayli Bhavsar and Dhwani Doshi** in the Sem VI, year 2022 under the guidance of **Prof. Pooja Malhotra** of Department of Information Technology in partial fulfillment of requirement for the completion of Mini-Project

Prof. Pooja Malhotra	Dr. Ujwala Bhangale
_____	_____
Guide	Head of the Department

Date: 25/04/2022

Place: Mumbai-77

K. J. Somaiya College of Engineering, Mumbai-77

(Autonomous College Affiliated to University of Mumbai)

Certificate of Approval of Examiners

We certify that this report entitled **Web Scraping & Prediction Analysis Of A Sports Website** is bona fide record of Mini-Project work done by **Sayli Bhavsar and Dhwani Doshi.**

This project is approved for the award of credits for completing Mini-Project course

_____

Internal Examiner

_____

External Examiner

Date:

Place: Mumbai-77

K. J. Somaiya College of Engineering, Mumbai-77

(Autonomous College Affiliated to University of Mumbai)

DECLARATION

We declare that this written report submission represents the work done based on our and / or others' ideas with adequately cited and referenced the original source. We also declare that we have adhered to all principles of intellectual property, academic honesty and integrity as we have not misinterpreted or fabricated or falsified any idea/data/fact/source/original work/ matter in our submission.

We understand that any violation of the above will be cause for disciplinary action by the college and may evoke the penal action from the sources which have not been properly cited or from whom proper permission is not sought.

Signature of the Student	**Signature of the Student**
1914009	**1914016**
Roll No.	**Roll No.**

Date:

Place: Mumbai-77

Dedicated to

…….……

Abstract

Machine Learning can be used to predict parameters in different fields, including sports, especially football. The football industry is a huge deal with a lot of football enthusiasts counting on reliable sources to gain knowledge on historical data and past trends to analyse the statistics of the game. An example of English Premier League (EPL) is indicated with this research that the possibility of finding a classifier that can predict the upcoming matches by accuracy of up to 65% can be gained.

The objective of this project is to implement supervised learning techniques like Linear SVC, Logistic Regression and K Nearest Neighbor (KNN) for solving a multi-classification problem.

Results show the analyzed data for the EPL league data and prediction of the full time result for the EPL football league.

Key words**:** LinearSVC, KNN, Logistic Regression. classification and data analysis.

Contents

List of Figures **i**

List of Tables **ii**

Nomenclature **iii**

1 Introduction **1**

1.1 Problem Definition 1

1.2 Motivation 1

1.3 Scope of project 1

1.4 Functional and Non-Functional Requirements 2

1.5 Organization of the report 3

2 Background Work **5**

2.1 Background 5

3 Implementation **7**

3.1 Technologies used 7

3.2 Algorithms/Methodogy 8

3.4 Features 9

3.5 Implementation Details 11

4 Results and Discussions **14**

4.1 Discussion 14

4.2 Deployement 14

5 Conclusions and Learnings **14**

5.1 Conclusions 18

5.2 Scope for future work 19

5.3 Learnings 19

6 References **20**

7 Acknowledgements **21**

List of Figures

1. Website Front Page 15
2. Upcoming Predictions 15
3. Analysis Per Team 16
4. Analysis Per Season 16
5. Correlation Of Game Stats 17

List of Tables

1 Background Work 6

Nomenclature

HGS - Home Goals Scored

AGC - Away Goals Scored

HGC - Home Goals Conceded

AGC - Away Goals Conceded

HAS - Home Attack Score

AAS - Away Attack Score

HDS - Home Defense Score

ADS - Away Defense Score

FTHG - Full Time Home Goals

FTAG - Full Time Away Goals

FTR - Full Time Result

HTHG - Half Time Home Goals

HTAG - Half Time Away Goals

HTR - Half Time Result

HS - Home team Shots

AS - Away teams Shots

HST - Home team Shots on Target

AST - Away teams Shots on Target

HF - Home Fouls

AF - Away FoulsHC - Home Corners

AC - Away Corners

HY - Home Yellow Cards

AY - Away Yellow Cards

HR - Home Red Cards

AR - Away Red Cards

Chapter 1
Introduction

1.1 Problem Definition

Although websites which display the score of the matches exist, there is a need to have a website which predicts the data for the upcoming football matches and provides statistics. Web Scraping can help us gather data which can later be stored in a csv file and analyzed. Data Analysis will be showcased on an web app that would be built to be interactive and where the information will be dynamically updated based on a given time.The web app will be an efficient way for football enthusiasts to easily track sport statistics and the predictions for upcoming matches.

1.2 Motivation

Sports enthusiasts come across a lot of websites with upcoming match time tables but no prediction data and websites with team data and history. Our objective is to make a single stop solution for match predictions and give team data and history in a visual website which helps in better understanding.

1.3 Scope of the project

The objective of the project is to build a website which will showcase the predictions of the upcoming football matches of the EPL league. The goal is to be able to create a visual understanding of the performance history of all the teams that have ever competed in the EPL league throughout the years. There are two parts of this project -

1. ***Data scraping, preprocessing and analyzing***

   - this involved the process of fine-tuning and cleaning the scraped data and creating multiple datasets out of the raw data. This project has made no use of any pre-existing dataset but has relied on the subsetted data for further prediction models and analysis.

2. ***Visualization website with statistics***

   - creation of a website which is made to be interactive for the user to be able to select all the parameters that affect the Full Time Result (Win, Lose or Draw) of a match and to be able to visually understand it and draw inferences.

1.4 Functional and Non-Functional Requirements

Functional Requirements

1. Real-time updating dataset

As new seasons take place, the data will update in real time according to the data scraped from the official sports website.

2. Prediction of future match winners

Machine learning models will be used to predict the Full Time Result of the matches happening in the upcoming week.

3. Data preprocessing and subsetting

The scraped data will be cleaned and transformed for the purpose of sports prediction and analysis.

4. Data visualization

Appropriate interactive visualizations will be created from the analyzed sports data after the data preprocessing stage.

Non-Functional Requirements

1. Portability

The web app will be usable in different environments like different browsers (Chrome, Firefox and Safari).

2. Scalability

The web app will be able to handle more parameters in the future for analysis purposes if necessary.

3. Reliability

The prediction data made on the web app through the algorithm will be reliable.

4. Performance

The performance will be evaluated based on the speed and efficiency on the results displayed on the web app. It is also necessary to have an optimized website which will ensure a good experience for the user.

1.5 Organization of the Report

- **Chapter 1 - Introduction**

A summary on the problem definition and scope of the project. This section explains the motivation behind choosing the project and the problem that it solves.

● **Chapter 2 - Background Work**

A background research was performed on various parameters that affect the full time result of a football match. Various research papers were studied to understand how machine learning was used for the purpose of prediction and the most relevant classifiers offering the highest accuracy were chosen.

● **Chapter 3 - Implementation and features**

A summary of all the different python libraries used and for what purpose is listed. Along with this, various algorithms used for the purpose of prediction are studies and explained. The deployment process is explain and figures for the same are attached.

● **Chapter 4 - Results and Discussion**

All the implementation and methodology results have been discussed. A functional website for football statistics and predictions have been displayed.

● **Chapter 5 - Conclusion and Learnings**

All the various things that we implemented successfully have been mentioned here as well as some extra modules or ideas that could be added later in our web app.

Chapter 2
Background Work

2.1 Background

There are various existing websites which give the historical data of past football matches, seasons and football player statistics. These websites are not interactive and have data in tabular format with little or no summarization. Our aim here is to provide the information and historical data in a way that is easier to understand, analyze and interpret. There are also websites where match predictions are displayed and upcoming match schedule is provided.

The part where our website stands out is that it combines these two objectives and provides it on one page, making it easier for the users to understand. We aim to use accurate, clear and succinct visualizations to convey the statistics. Using the machine learning models and feature selection techniques from the research papers mentioned below (Table 1), our overall objective is to make a website for EPL league football match predictions and data visualizations.

We referred several research papers, articles and other resources in our research before starting the project implementation. There are three research papers from where we derived our understanding majorly. The first paper “Predicting EPL Football Matches Results Using Machine Learning Algorithms.” described what data should be used for the prediction model, which mahine learning models should be applied on the data for the purpose of prediction.

The second paper, “Predicting Football Match Results using Machine Learning” gave us insights on evaluating multiple machine learning models which could be useful for prediction purposes. Further, the paper explained how to work on feature selection and which decided how the data should be preprocessed.

Lastly, the paper “A machine learning framework for sport result prediction” gave us knowledge about identifying the learning methodologies utilized, data sources, appropriate means of model evaluation, and specific challenges of predicting sport results.

Research Papers On Machine Learning And Prediction Analysis

Sr. No.	Title	Year	Key Takeaways
1	Saiedy, Sayed & Qachmas, Muhammad & Amanullah, Faqiri. **Predicting EPL Football Matches Results Using Machine Learning Algorithms**. International Journal of Engineering Applied Sciences and Technology, 2020, 5.83-91.10.33564/IJEAST.2020.v05i03.013.	2020	-What data should be used for the prediction model. -Which machine learning models should be applied on the data for the purpose of prediction.
2	Ishan Jawade, Rushikesh Jadhav, Mark Joseph Vaz, Vaishnavi Yamgekar4. **Predicting Football Match Results using Machine Learning,** Volume 08 Issue 07, 2021, ISSN: 2395-0056	2021	-Evaluated multiple machine learning models which could be useful for prediction purposes. -Worked on feature selction and which decided how the data should be preprocessed.
3	Rory P. Bunker, Fadi Thabtah, **A machine learning framework for sport result prediction**, Applied Computing and Informatics, Volume 15, Issue 1, 2019, Pages 27-33,ISSN 2210-8327.	2019	-Identified the learning methodologies utilised, data sources, appropriate means of model evaluation, and specific challenges of predicting sport results.

Table 1: Background Work

Chapter 3
Implementation

3.1 Technologies used

a) For Web Scraping

- Python
    - Beautiful Soup Library, requests library

b) For Data Analysis

- Python
    - Numpy, Pandas
    - Visualizations - Matplotlib, Seaborn
    - Machine Learning - Sklearn Library (and other machine learning modules: LogisticRegression, KNeighborsClassifier, LinearSVC)

c) For Real-Time Website

- Python
    - Streamlit Python Library (for creating the interactive website)

3.2 Algorithms and methodology

The algorithms were selected based on the background work and research done during the initial stages of the project. We used the 3 classifiers to understand which one was giving us the best accuracy and then proceeded to use the same for further prediction.

Logistic Regression and Linear SVC gave us a score of 56.7% whereas K Nearest Neighbors gave us an accuracy of 65% approximately. This resulted in the KNN model being used for the purpose of prediction of the full time result of the upcoming week's matches.

Algorithms used for full time result prediction:

1. Logistic Regression

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable.
It is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation. This type of analysis can help you predict the likelihood of an event happening or a choice being made.

This algorithm gives us 56.7% accuracy in the prediction of the full time result.

2. K Nearest Neighbors

KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well

suited category by using KNN algorithm. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

This algorithm gives us 65% accuracy in the prediction of the full time result.

3. Linear SVC

The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is.
The Linear Support Vector Classifier (SVC) method applies a linear kernel function to perform classification and it performs well with a large number of samples.

This algorithm gives us 56.7% accuracy in the prediction of the full time result.

3.4 Features

1. View past seasons' historical data and sort according to attributes

The website has a section where past seasons' data is displayed in a clean tabular form for users to refer. This table can be sorted according to the attribute user desires. If the user is looking for the matches with maximum goals scored from highest to lowest, he/she can modify the view of the table by sorting the goals columns in ascending. This aims to provide a representation of the data used in further features.

2. Weekly predictions of the upcoming matches

The website showcases prediction details for the next 10 EPL football league matches. It displays the Home Team, Away Team and also the Full Time Result Winner (In terms of A,H,D that stands for Home Team, Away Team and Draw respectively).

3. Interactive Visualisations

Various visualizations are provided which require no further summarizations or textual explanations. These visualizations are best fit to scale, are clear, empowering and succiently answer questions user is looking for.

To narrow down the search for what user is looking for, we have drop down menus for users to choose particular attribute and measures for the graph axes.

There are three graphs on the website:

1) Analysis per team

Provides inferences on: What is the total number of goals scored by a team? Which is the team with the minimum number of fouls ever received in the league?

2) Analysis per season

Provides inferences on: Which season had teams score the most goals? What is the total number of red cards received in each season?

3) Correlation of Game stats

Provides inferences on: Do teams that have more shots on targets have more corners? Do the team's shots on target say anything about the probability of goals

3.5 Implementation Details

A. Data Analysis and Prediction

1. **Web Scraping of EPL**

- Scrape data from www.football-data.co.uk using beautifulsoup library
- Store data from last 20 seasons in a dataframe and eventually in a csv

2. **Data Cleaning and subsetting required data**

- Remove unnecessary columns and betting data as it is irrelevant.
- Calculate Parameters required like HGS AGS HGC and AGC
- Further calculate HAS AAS HDS and ADS

3. Working on feature table

- Feature team contains all the fixtures in the current season
- Map the respective home and away teams values with attack(HAS, AAS) and defense score (HDS, ADS).

4. Prediction Model

- Identify the dependent and independent variable
- Split the data set into X train and Y train
- Using three machine learning algorithms - Logistic Regression, Linear SVC and KNeighbourClassifier find the prediction results
- Compare all three algorithms and find the most suitable model

5. Predict upcoming week matches

- Use the best model (KNN in this case) and predict the result of upcoming matches

6. Analysis of football season and statistics

- Questions like
  - What is the total number of goals scored by a team?
  - Which is the team with the minimum number of fouls ever received in the league?
  - Do teams that have more shots on targets have more corners?
  - What is the total number of red cards received in each season?

B. Building the website

1. **Importing and showing raw data**

- Read the raw data and display it on the website (Figure 1) along with project title and description

2. **Next week predictions**

- Read the predictions from the prediction model results csv and display the data frame (Figure 2) with respective home team and away team

3. **Plot team analysis**

- Answers questions such as: What is the total number of goals scored by a team? Which is the team with the minimum number of fouls ever received in the league?
- Provide drop down for user to select an attribute and measure as mentioned in (Figure 3).

4. **Plot season analysis**

- Answers questions such as: Which season had teams score the most goals? What is the total number of red cards received in each season?
- Provide drop down for user to select an attribute and measure as mentioned in (Figure 4).

5. Plot correlation analysis

- Answers questions such as: Do teams that have more shots on targets have more corners? Do the team's shots on target say anything about the probability of goals
- Provide drop down for what type of correlation, what attributes on x axis and what attributes on y axis in (Figure 5).

Chapter 4
Result and Discussion

4.1 Discussion

- We have worked on creating a model which **predicts the outcome** (FTR - Full Time Result) of the **upcoming match schedule.**
- Since, this is a **multi-classification problem**, we have predicted the result according to 3 parameters - Win, Lose or Draw.

- After testing multiple machine learning models, we have come to the conclusion that **K Nearest Neighbor (KNN)** ML algorithm gives the best accuracy of **65%**.

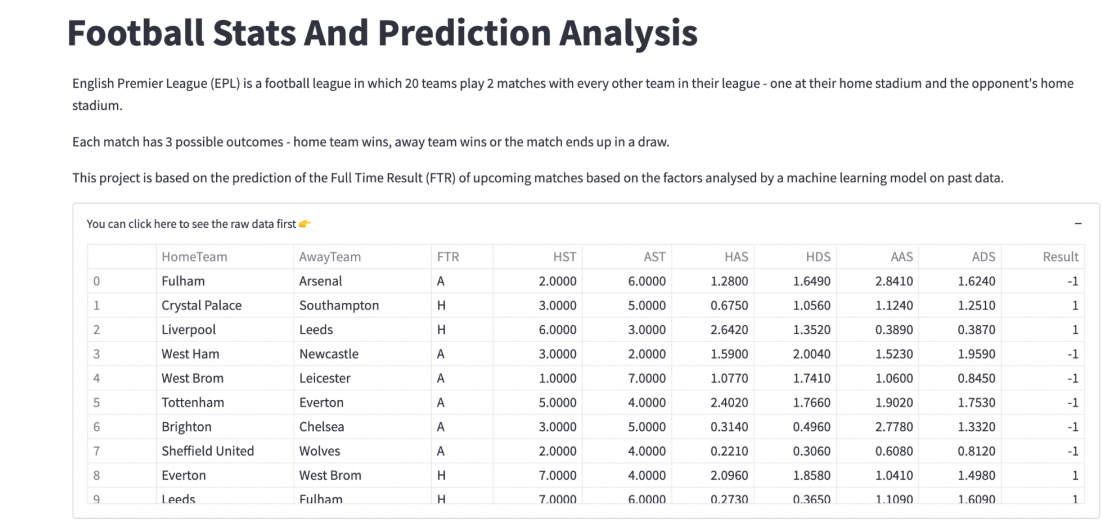
- **Visualization on the website** - after preprocessing and subsetting of the data, we have created csv files for the purpose of analysis and made interactive and dynamically changing visualizations on:

    - Analysis according to team
    - Analysis according to season
    - Correlation between attributes

4.2 Deployment

The web app was deployed on the platform, Streamlit which is an open source app framework in Python language. It helps us to create data science and machine learning

algorithms in a short time. We created a website using the data that we had preprocessed and make useful visualizations from the same.



Football Stats And Prediction Analysis

English Premier League (EPL) is a football league in which 20 teams play 2 matches with every other team in their league - one at their home stadium and the opponent's home stadium.

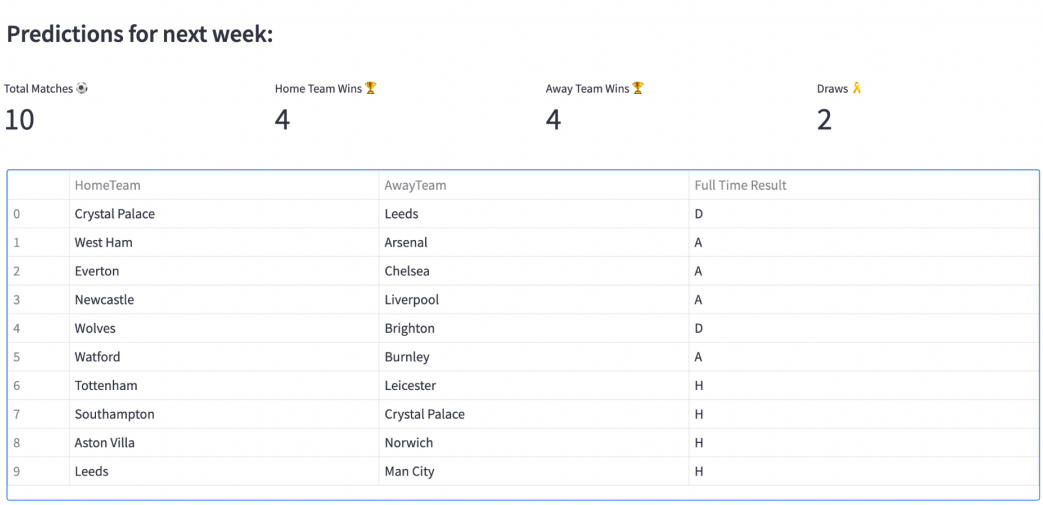
Each match has 3 possible outcomes - home team wins, away team wins or the match ends up in a draw.

This project is based on the prediction of the Full Time Result (FTR) of upcoming matches based on the factors analysed by a machine learning model on past data.

You can click here to see the raw data first 👉

	HomeTeam	AwayTeam	FTR	HST	AST	HAS	HDS	AAS	ADS	Result
0	Fulham	Arsenal	A	2.0000	6.0000	1.2800	1.6490	2.8410	1.6240	-1
1	Crystal Palace	Southampton	H	3.0000	5.0000	0.6750	1.0560	1.1240	1.2510	1
2	Liverpool	Leeds	H	6.0000	3.0000	2.6420	1.3520	0.3890	0.3870	1
3	West Ham	Newcastle	A	3.0000	2.0000	1.5900	2.0040	1.5230	1.9590	-1
4	West Brom	Leicester	A	1.0000	7.0000	1.0770	1.7410	1.0600	0.8450	-1
5	Tottenham	Everton	A	5.0000	4.0000	2.4020	1.7660	1.9020	1.7530	-1
6	Brighton	Chelsea	A	3.0000	5.0000	0.3140	0.4960	2.7780	1.3320	-1
7	Sheffield United	Wolves	A	2.0000	4.0000	0.2210	0.3060	0.6080	0.8120	-1
8	Everton	West Brom	H	7.0000	4.0000	2.0960	1.8580	1.0410	1.4980	1
9	Leeds	Fulham	H	7.0000	6.0000	0.2730	0.3650	1.1090	1.6090	1

Figure 1: Website Front Page



Predictions for next week:

Total Matches ⚽ 10

Home Team Wins 🏆 4

Away Team Wins 🏆 4

Draws 🎗 2

	HomeTeam	AwayTeam	Full Time Result
0	Crystal Palace	Leeds	D
1	West Ham	Arsenal	A
2	Everton	Chelsea	A
3	Newcastle	Liverpool	A
4	Wolves	Brighton	D
5	Watford	Burnley	A
6	Tottenham	Leicester	H
7	Southampton	Crystal Palace	H
8	Aston Villa	Norwich	H
9	Leeds	Man City	H

Figure 2: Upcoming Week Predictions

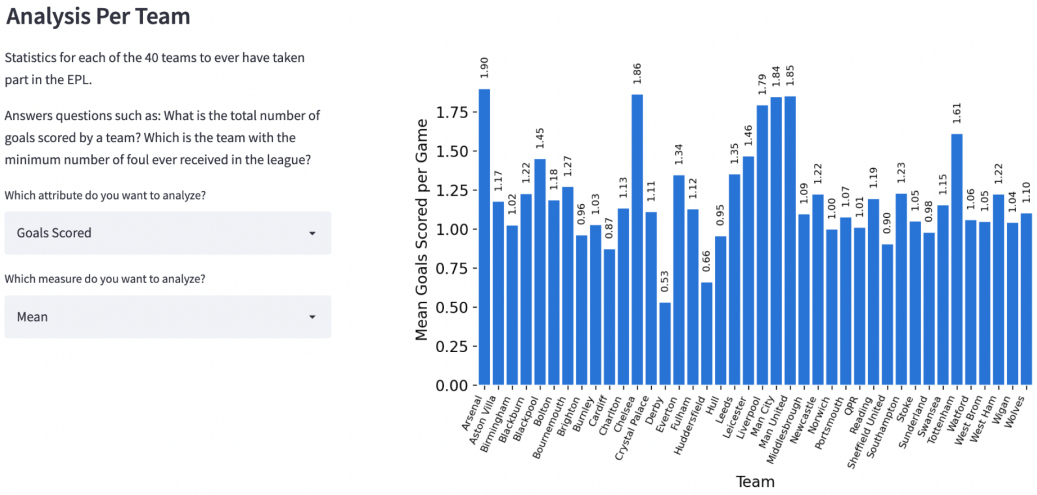


Figure 3: Analysis Per Team

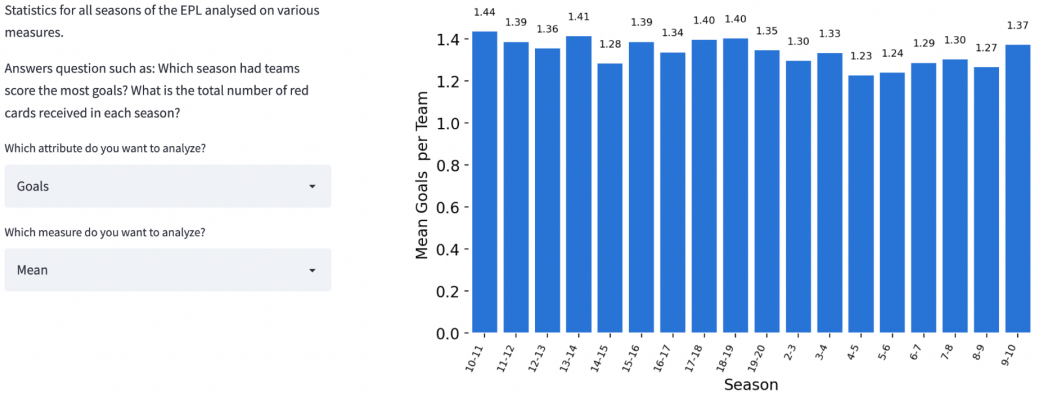


Figure 4: Analysis Per Season

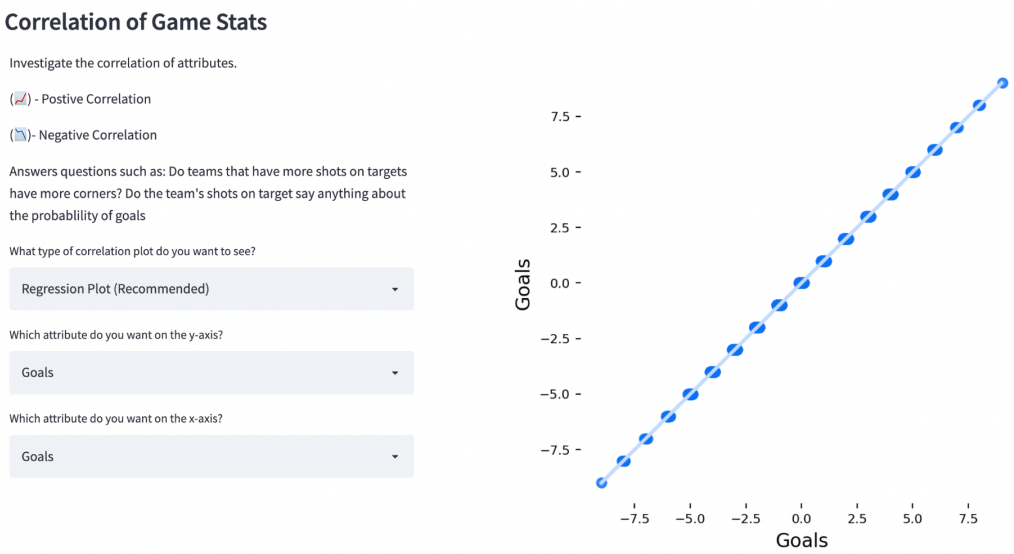


Figure 5: Correlation Of Game Stats

Chapter 5
Conclusions and Learnings

5.1 Conclusions

- Machine learning can be used to predict parameters in different fields, including sports, especially football. An example of English Premier League (EPL) is indicated with this research that the possibility of finding a classifier that can predict the upcoming matches by accuracy of up to 65% can be gained. And also there is lots of work to do and research more on the features in the future.

- The result of this project which has given us the prediction model for the full time result of EPL football matches has the accuracy of 65% for KNN, 56% for Linear SVC and 56% for Logistic Regression.

- Inference from visualizations - countless questions can be answered from the website such as:
    - What is the total number of goals scored by a team?
    - Which is the team with the minimum number of foul ever received in the league?
    - Do teams that have more shots on targets have more corners?
    - What is the total number of red cards received in each season?

5.2 Scope for future work

This project has provided us with the most important prediction feature of any sports website which is based upon whether the Home Team/Away Team will win, lose or will it be draw. Even though this is the feature on every sports website that the user is curious about, the project can be improved by predicting multiple other features, number of expected goals, player statistics and more.

- The implementation of a full fledged website that allows the users to not only gain insightful knowledge about the player statistics and result of every football match but also allow them to contribute to the betting odds to predict who will win.
- The website can show analysis for multiple leagues, the logic of which can be built on top of the EPL league, done up until now.

5.3 Learnings

Sports prediction is a huge industry in itself at the moment. One of the major use case for building this website is to provide a one stop solution for football enthusiasts to view all data till date, see predictions for upcoming matches and show analysis according to the team on the attribute and measure selected by the user. This is why we decided to implement a project that could use machine learning algorithms and create a user-friendly website which could show the predictions of the upcoming week and other statistics. The project gave us an idea on how to develop a machine learning project from scratch. While building the website, we encountered a lot of errors and it was a difficult task to resolve all the errors slowly and gradually.

6 References:

1. Saiedy, Sayed & Qachmas, Muhammad & Amanullah, Faqiri. **Predicting EPL Football Matches Results Using Machine Learning Algorithms**. International Journal of Engineering Applied Sciences and Technology, 2020, 5. 83-91. 10.33564/IJEAST.2020.v05i03.013.

2. Ishan Jawade1, Rushikesh Jadhav2, Mark Joseph Vaz3, Vaishnavi Yamgekar4. **Predicting Football Match Results using Machine Learning,** Volume 08 Issue 07, 2021, ISSN: 2395-0056

3. Rory P. Bunker, Fadi Thabtah, **A machine learning framework for sport result prediction**, Applied Computing and Informatics, Volume 15, Issue 1, 2019, Pages 27-33,ISSN 2210-8327.

4. https://understat.com/league/EPL/2021

5. https://www.football-data.co.uk/

6.https://medium.com/cem-kilicli/estimating-football-game-results-with-statistics-3ea1ee9abd2c

7.https://github.com/llSourcell/Predicting_Winning_Teams/blob/master/Prediction.ipynb

8. https://docs.streamlit.io/library/api-reference

9. https://www.kaggle.com/code/aj7amigo/epl-data-analysis-2009-2019/notebook

10. https://scikit-learn.org/stable/

11. https://www.crummy.com/software/BeautifulSoup/bs4/doc/

12. https://github.com/tdenzl/BuLiAn

7 Acknowledgments:

We would like to express our gratitude towards our peers & faculty of KJSCE for their kind cooperation and encouragement which helped us in the completion of this project. We would like to express our special gratitude and thanks to industry persons for giving us such attention and time.