

Astana IT University

Advanced Programming | Python

Project: Football Match Predictor

Students Name: Zhandarbek Saginov, Orymbayev Ali

Group: IT-2101

Supervisor: Sultanmurat Yeleu

February 2023

Table of Content

[**Introduction** 3](#_Toc127823438)

[**Problem** 3](#_Toc127823439)

[**Literature Review and References** 3](#_Toc127823440)

[**Current Work** 3](#_Toc127823441)

[**Data and Methods** 3](#_Toc127823442)

[**Information about the Data** 3](#_Toc127823443)

[**ML Model Description** 4](#_Toc127823444)

[**Results** 5](#_Toc127823445)

[**Discussion** 5](#_Toc127823446)

[**Critical Review of the Results** 5](#_Toc127823447)

[**Limitations and Future Work** 6](#_Toc127823448)

# **Introduction**

## **Problem**

Undoubtedly football is the most popular sport in the world. The purposes of football match prediction are betting, fan engagement and data analysis. Betting and fan engagement are close things, both serve for people entertainment in the sport, while for data analysis is another field for investigation which helps coaches and players to better estimate their strong sides and prepare for matches. This project is going to predict the outcomes of matches using neural networks depending on the past performance of teams.

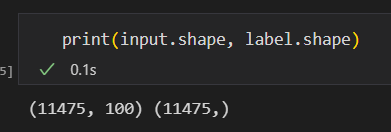
## **Literature Review and References**

All club statistics were scraped using bs4.BeautifulSoup from <https://understat.com/>. The website provides useful statistics of every football game in TOP-5 Leagues (EPL, Serie A, Bundesliga, Ligue 1, La Liga).

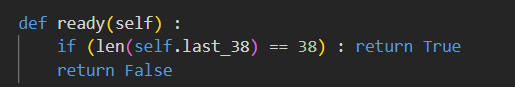
GitHub solution <https://github.com/krishnakartik1/LSTM-footballMatchWinner>.

**Current Work**

The model takes 60% of training data, and the other 40% used for validation. Giving the result of 51~% accuracy on training and 49~% accuracy on validation. Since football games do not only depend on statistics, these factors is enough to predict with 50% accuracy. Also as shown:



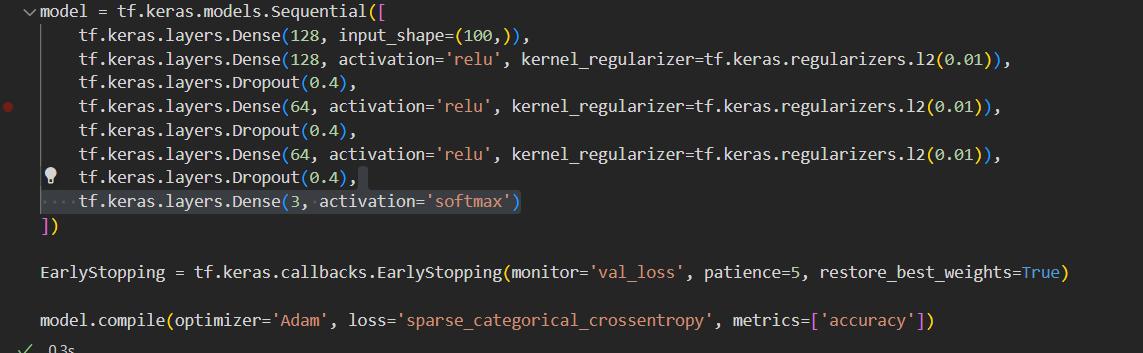
We only used ~11000 data for training while having ~15000 games. Since the single input includes the statistics of past 38 matches (38 – one season), we did not pass the data until we gather 38 matches for 1 team:



inside Class ‘Club’

In general:

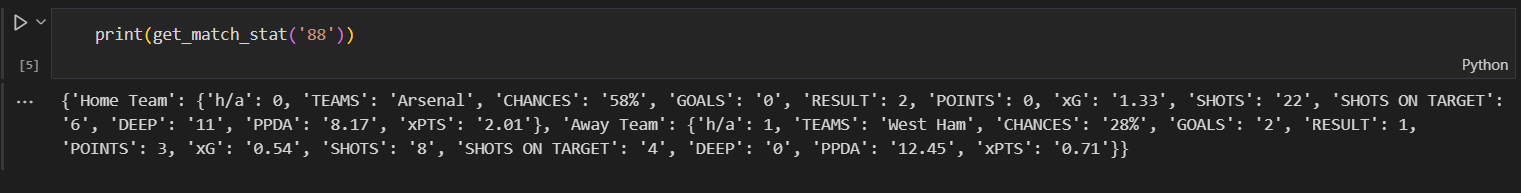
The model:



# **Data and Methods**

## **Information about the Data**

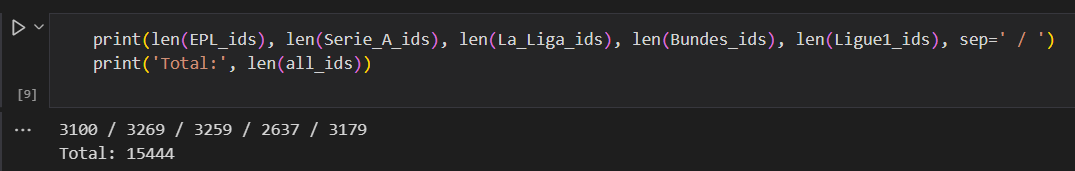
The data includes the average performance of teams’ last 5, 10, 15, 25, 38 games (the most valuable ones are last 5 and 10 games, since it shows current condition of teams). So, every team has an array of size 50 showing their stats. The data contains matches from August 2014 until Today.



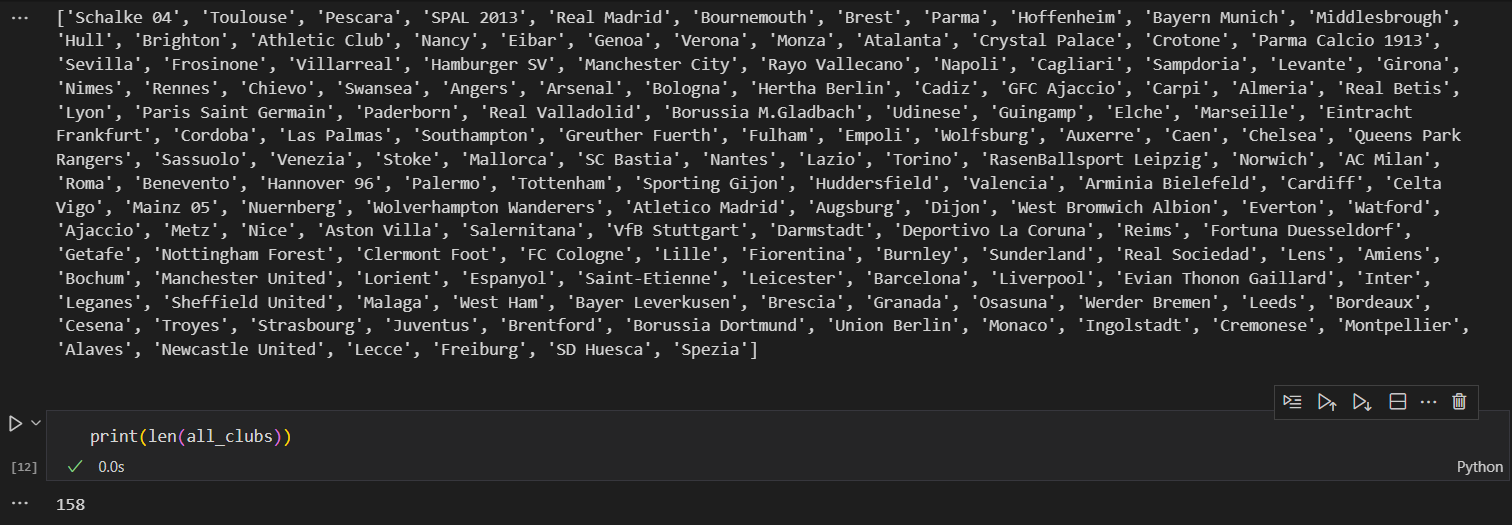
You can see the following information:

* ‘h/a’ – indicates, which club is being the host in their stadium and which team is being guest to first team’s stadium.
* ‘TEAMS’ – shows team name (not included in stats)
* ‘CHANCES’ – shows the percentage of dangerous attacks over all attacks
* ‘GOALS’ – shows the number of goals scored
* ‘RESULT’ – shows the result of the match (0 – draw, 1 – won, 2 – lost)(not included in stats)
* ‘POINTS’ – points earned in the game (0 – lost, 1 – draw, 3 – won)
* ‘xG’ – expected goals. Every shot in the game is estimated by some algorithm which depends on many factors (goal complexity, distance to goal, player skill, goalkeeper skill, shot convenience and so on)
* ‘SHOTS’ – number of shots
* ‘SHOTS ON TARGET’ – number of shots that were striking into gates (not counting crossbars or deflections away from gates)
* ‘DEEP’ – number of passes accomplished in opponents goal area
* ‘PPDA’ – passes allowed per defensive action in the opponent’s half
* ‘xPTS’ – expected points, may rely on expected goals, and other indicators

One input contains [first\_team\_stat, second\_team\_stat], then used in model training



In total we have 15444 matches or data



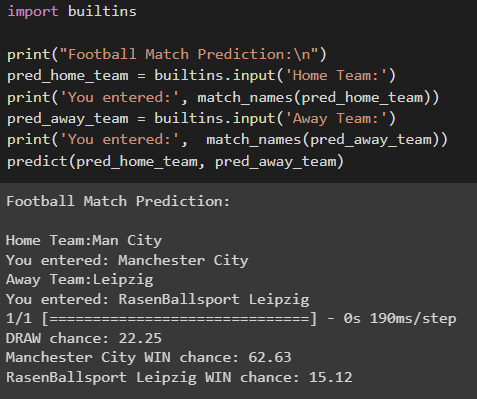
With 158 Clubs

## **ML Model Description**

This model is a neural network designed to predict the outcomes of football matches. The model itself consists of several dense layers, each with its own activation function and dropout regularization to prevent overfitting. The first two layers have 128 neurons and use the ReLU activation function, while the next two have 64 neurons and also use ReLU activation. The final layer has 3 neurons and uses the softmax activation function to output the probabilities of each possible outcome of a football match. The model is compiled with the Adam optimizer and the sparse\_categorical\_crossentropy loss function, which is commonly used for multi-class classification problems. The accuracy metric is used to evaluate the model's performance. Additionally, the EarlyStopping callback is used to stop training if the validation loss does not improve for 5 epochs, preventing overfitting and saving the best weights.

# **Results**

Based on the data input for the football match between Manchester City and RasenBallsport Leipzig, the code predicts of each team winning or the match ending in a draw.

****

The overall results indicate that Manchester City is favored to win the match with a chance of 62.63%, while RasenBallsport Leipzig has a lower chance of winning at 15.12%. The probability of the match ending in a draw is 22.25%.

# **Discussion**

## **Critical Review of the Results**

The results of the match predictions using the provided data show that Manchester City has a higher probability of winning than RasenBallsport Leipzig. The predicted probabilities for each team and the draw seem to be based on a neural network model, which has been trained on historical data. However, It is important to note that such predictions can be influenced by many other factors, such as player injuries, team form, weather conditions, tactical changes, and other unforeseen circumstances. Therefore, while these predictions provide some insight into the likely outcome of the match, they should be viewed as estimates, rather than definitive results.

## **Limitations and Future Work**

1)The project only relies on statistical information of teams, it does not include advanced stats or human factors, mental, social and other states of teams. Also, it does not estimate players skill, or starting XI (sometimes coaches may use alternative squads for invaluable games). However, depending on stats and additional human thought may bring to closest outcome.

2) The model still needs boost in accuracy. This can be accomplished by adding more training data, and using better model architecture.

3) Limited use. This project only predicts the match outcomes of teams that ARE in TOP-5 Leagues, not including the rest of the World. Data of 2014-2022 years were used only for training, only use teams that exist in TOP-5 Leagues right now, since it shows most actual and new statistics.

4) Future expansion of information.