Predicting Premier League teams performance on next season.

José Eduardo Orenday de la Rosa

July 27, 2020

Introduction.

Background.

Premier League is one of the best football leagues in the world, with millions of fans from all around the world. Year by year, thousands of fans are looking for a chance to go to the best possible games of the season, and it implies lots of expenses, so it can turn disappointing to go and see a bad, boring game.

Problem.

Data that might contribute to get a better idea of when your favorite team could have a better performance, based on statistics from their own games on last season.

Interest.

Travel agencies and sports broadcasters would have an interest on this tool, because they could predict what games will have a better quality for their audience's preferences. And they would have the opportunity of offering an extra in their recommendations about when to travel.

Data Acquisition and cleaning.

Data sources.

Dataset including statistics from all matches from season 2018-2019 from the Premier League.

Data displays several attributes (full documentation) :

- Div = League Division
- Date = Match Date (dd/mm/yy)
- HomeTeam = Home Team
- AwayTeam = Away Team
- FTHG = Full Time Home Team Goals
- FTAG = Full Time Away Team Goals
- FTR and Res = Full Time Result (H=Home Win, D=Draw, A=Away Win)

- HTHG = Half Time Home Team Goals
- HTAG = Half Time Away Team Goals
- HTR = Half Time Result (H=Home Win, D=Draw, A=Away Win)

Match Statistics

- HS = Home Team Shots
- AS = Away Team Shots
- HST = Home Team Shots on Target
- AST = Away Team Shots on Target
- HC = Home Team Corners
- AC = Away Team Corners
- HF = Home Team Fouls Committed
- AF = Away Team Fouls Committed
- HY = Home Team Yellow Cards
- AY = Away Team Yellow Cards
- HR = Home Team Red Cards
- AR = Away Team Red Cards

Algo, we used a dataset including the coordinates of all the football stadiums in England, and added queries from Foursquare regarding venues close to the stadiums.

Data Cleaning.

The only two operations performed for data cleaning were encoding of catergorical values (“FTR” and “HTR”).

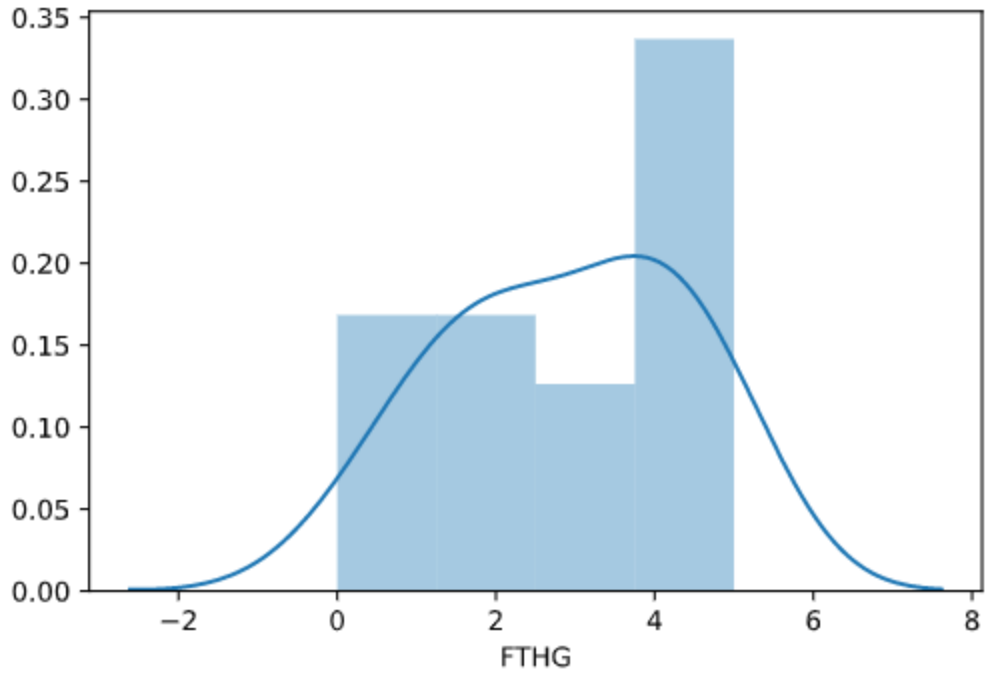
Feature selection.

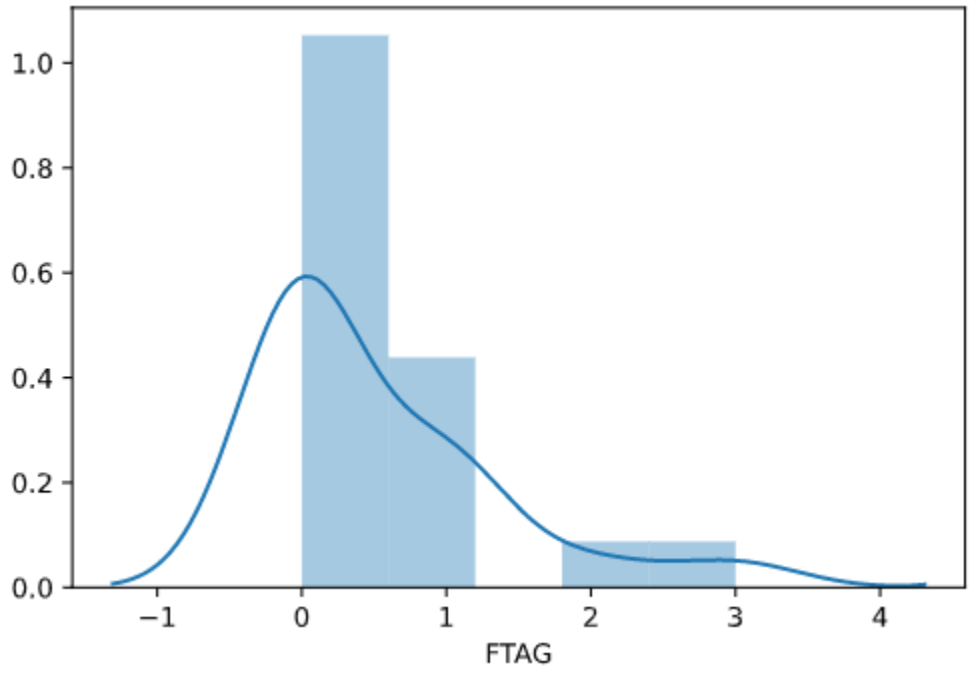
The target value was the Final time result. Fans want to know when their favorite team has better chances of winning. Since we didn´t notice a clear correlation with some specific features, we decided to use all of them, so that we could enrich the source of information for prediction.

Exploratory analysis.

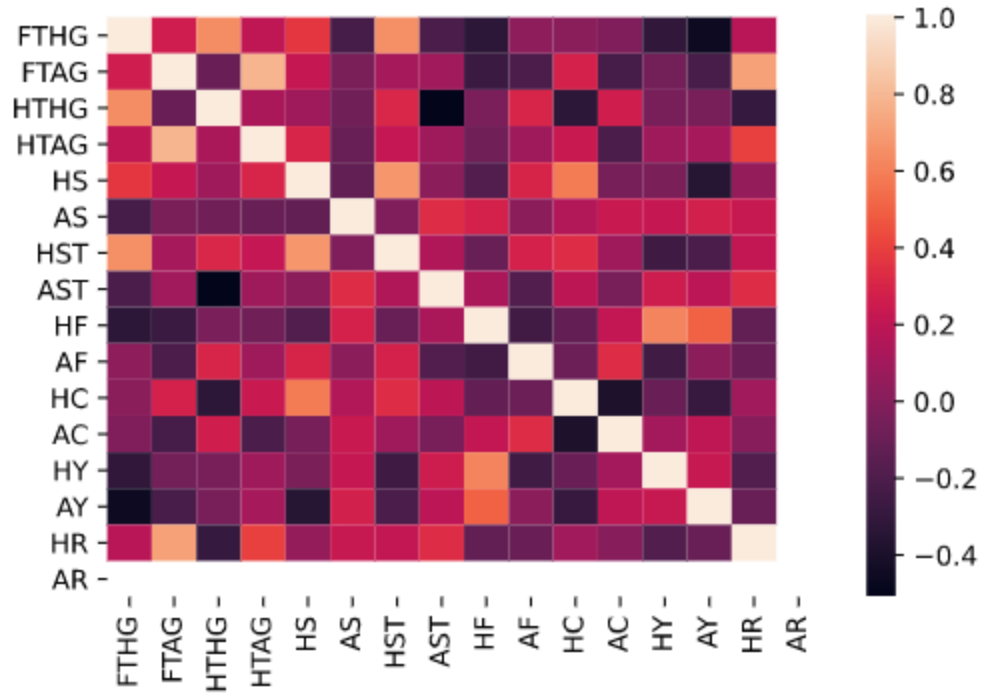
We created a script to count the amount of victories, draws and looses of the selected team. Also, we divided the data in home and visit games.

We performed three visualizations in order to know the goals distribution for each team across all the season.





We also created a heatmap to get a better idea of the correlation of all the features with the target value.



Modeling.

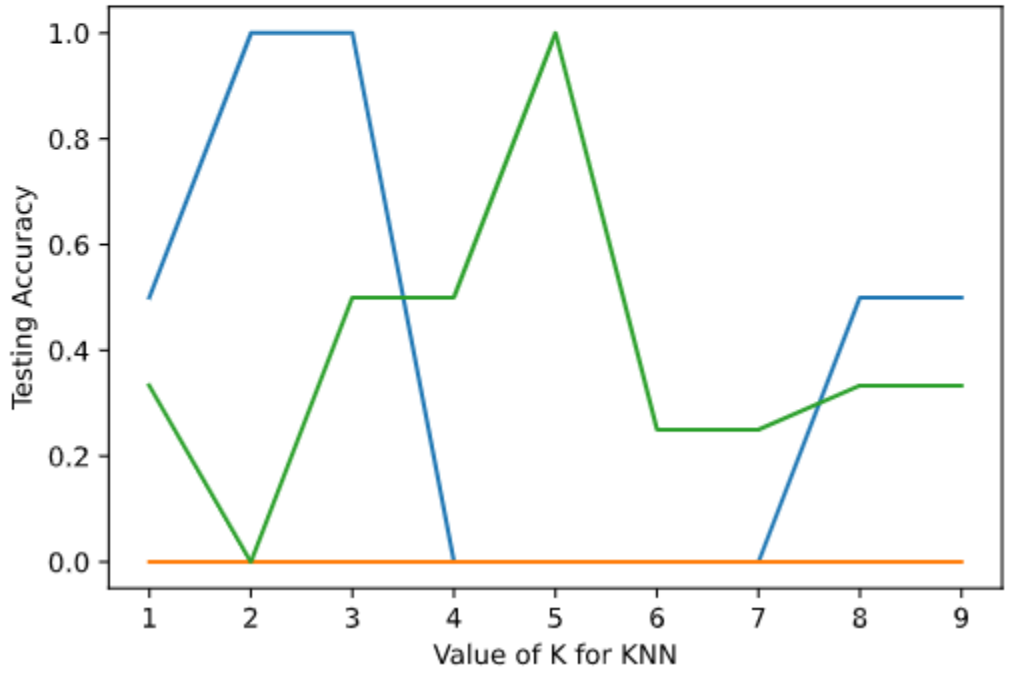
We used a four classifiers (Decision tree, SVM, Logistic Regression and KNN) in order to get the option with the best accuracy. Each one with proper testing and evaluation methods. Here are the results we've got for the Liverpool F.C. example:

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.56	0.64	NA
Decision Tree	0.56	0.64	NA
SVM	0.56	0.64	NA
Logistic Regression	0.56	0.64	0.73

Test results for each classifier:

KNN

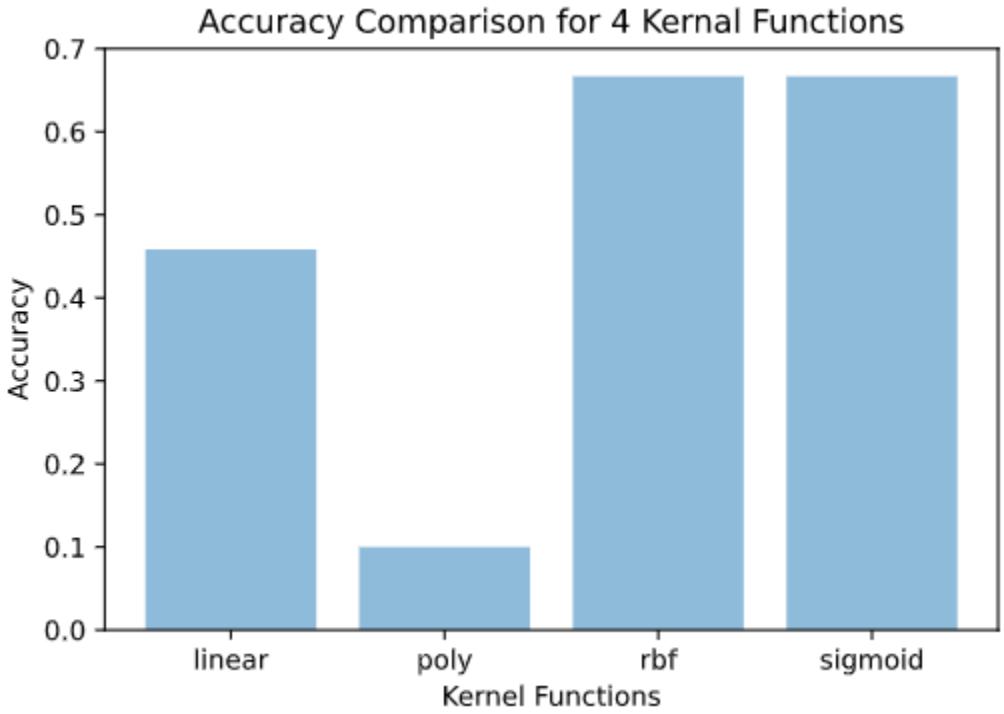
```
Test set Accuracy at k= 1 :  [0.5        0.         0.33333333]
Test set Accuracy at k= 2 :  [1. 0. 0.]
Test set Accuracy at k= 3 :  [1.  0.  0.5]
Test set Accuracy at k= 4 :  [0.  0.  0.5]
Test set Accuracy at k= 5 :  [0. 0. 1.]
Test set Accuracy at k= 6 :  [0.   0.   0.25]
Test set Accuracy at k= 7 :  [0.   0.   0.25]
Test set Accuracy at k= 8 :  [0.5        0.         0.33333333]
Test set Accuracy at k= 9 :  [0.5        0.         0.33333333]
Text(0, 0.5, 'Testing Accuracy')
```



Decision Tree

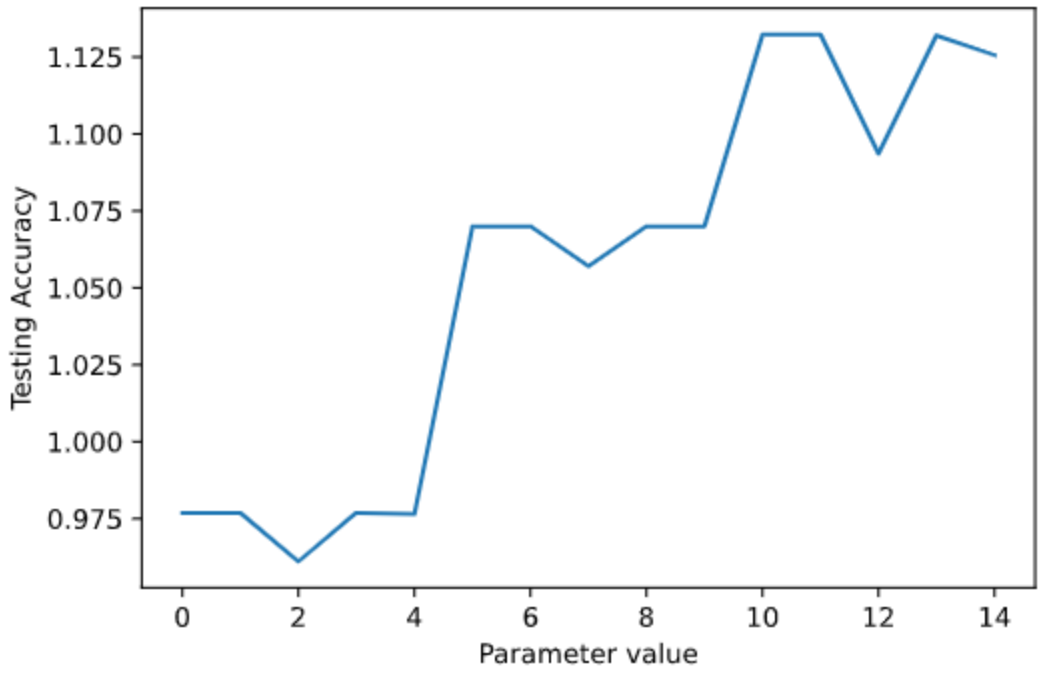
Depth	F1-score	Jacard
d=3	0.100	0.062500
d=4	0.375	0.333333
d=5	0.000	0.000000

SVM



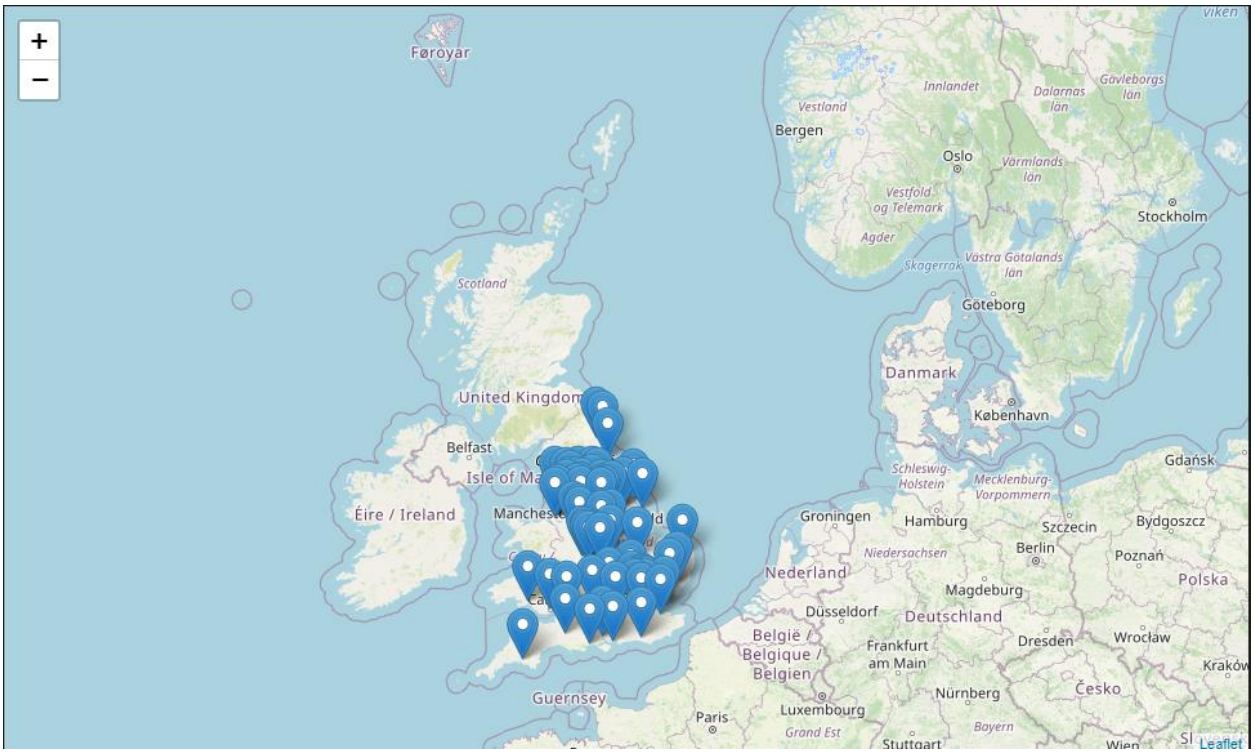
Logistic Regression

```
Test  1 : Accuracy at c = 0.1 solver= lbfgs is :  0.9769077408599538
Test  2 : Accuracy at c = 0.1 solver= liblinear is :  0.9611040296223794
Test  3 : Accuracy at c = 0.1 solver= sag is :  0.9769116279156997
Test  4 : Accuracy at c = 0.1 solver= saga is :  0.9764579108439202
Test  5 : Accuracy at c = 0.01 solver= newton-cg is :  1.0699895923379548
Test  6 : Accuracy at c = 0.01 solver= lbfgs is :  1.0699906453604398
Test  7 : Accuracy at c = 0.01 solver= liblinear is :  1.0570132473150786
Test  8 : Accuracy at c = 0.01 solver= sag is :  1.0699678985880043
Test  9 : Accuracy at c = 0.01 solver= saga is :  1.069944929943342
Test  10 : Accuracy at c = 0.001 solver= newton-cg is :  1.1323091250114061
Test  11 : Accuracy at c = 0.001 solver= lbfgs is :  1.1323082053033673
Test  12 : Accuracy at c = 0.001 solver= liblinear is :  1.0935822798162471
Test  13 : Accuracy at c = 0.001 solver= sag is :  1.1319695856559622
Test  14 : Accuracy at c = 0.001 solver= saga is :  1.125641937460775
Text(0, 0.5, 'Testing Accuracy')
```



Additional data:

We created a map locating all the stadiums in England, adding queries for restaurants, bus stations, metro stations and different amenities close to the stadiums, in order to give the user better tools for a better experience if they decide to travel.



	name	categories	address	lat	lng	labeledLatLngs	distance	postalCode	cc	city	state	co
0	Park Plaza	Hotel	41 Maid Marian Way, Nottingham	52.952956	-1.153436	[{'label': 'display', 'lat': 52.952956, 'lng':...	44773	NG1 6GD	GB	Nottingham	Nottinghamshire	U Ki
1	Holiday Inn Express	Hotel	Brayford Enterprise Park, Ruston Way	53.225963	-0.551053	[{'label': 'display', 'lat': 53.22596289759921...	47926	LN6 7DB	GB	Lincoln	Lincolnshire	U Ki
2	Best Western Orton Hall Hotel	Hotel	The Village,Orton Longueville	52.554027	-0.278955	[{'label': 'display', 'lat': 52.55402727610075...	32147	PE2 7DN	GB	Cambridgeshire	Cambridgeshire	U Ki
3	Radisson Blu Hotel, East Midlands Airport	Hotel	Herald Way, Pegasus Business Park	52.825237	-1.308500	[{'label': 'display', 'lat': 52.825237, 'lng':...	51799	DE74 2TZ	GB	East Midlands	NaN	U Ki
4	Crowne Plaza Nottingham	Hotel	Wollaton St	52.955481	-1.153474	[{'label': 'display', 'lat': 52.9554811, 'lng'...	44885	NG1 5RH	GB	Nottingham	Nottinghamshire	U Ki

	name	categories	address	lat	lng	labeledLatLngs	distance	cc	city	state	country	formattedA
0	Mansfield Bus Station	Bus Station	Quaker Way	53.143125	-1.197828	[{'label': 'display', 'lat': 53.14312477867679...	58661	GB	Mansfield	Nottinghamshire	United Kingdom	[Quake Mans Nottingham Un
1	Victoria Bus Station	Bus Station	Mansfield Rd.	52.958911	-1.148613	[{'label': 'display', 'lat': 52.95891120093947...	44738	GB	Nottingham	Nottinghamshire	United Kingdom	[Mansfiel Notti Nottingham
2	Lincoln Central Bus Station	Bus Station	Melville St.	53.226493	-0.538597	[{'label': 'display', 'lat': 53.226493, 'lng':...	47980	GB	Lincoln	Lincolnshire	United Kingdom	[Melvill Li Lincoln LN5 7
3	Queensgate Bus Station	Bus Station	Queensgate Shopping Centre	52.574695	-0.246425	[{'label': 'display', 'lat': 52.57469459738859...	31577	GB	Peterborough	Peterborough	United Kingdom	[Quee Shopping C Peterbo
4	Milton Keynes Coachway	Bus Station	NaN	52.455306	-1.136969	[{'label': 'display', 'lat': 52.45530622092375...	55315	GB	NaN	NaN	United Kingdom	[United Ki

	name	categories	address	lat	lng	labeledLatLngs	distance	postalCode	cc	city	state	country	form
0	The Park Tunnel (Derby Road Entrance)	Metro Station	Derby Road	52.955212	-1.158468	[{'label': 'display', 'lat': 52.95521173418335...	45182	NG1	GB	Nottingham	Nottinghamshire	United Kingdom	Nott
1	A1(M) /A14 Junc14 S/B	Metro Station	NaN	52.380168	-0.255667	[{'label': 'display', 'lat': 52.38016792233747...	50077	NaN	GB	NaN	NaN	United Kingdom	[Uni

Conclusions:

It is very hard to predict the result of a match based only on hard match data. Prediction would increase if we added information regarding ball possession, team position and tactic styles. Unfortunately, we could not find information regarding those features. Anyway, we reached a maximum accuracy of .65, which gives us very good chances to have predictions with good results.