FINAL PROJECT

FUNDAMENTAL OF DATA SCIENCE

9 de junio de 2022

domingo menéndez rúa

# INDEX

1. PROBLEM
2. ANALYSIS OF THE DATASET
3. SELECTING THE DATA AND ANALYZING THE DATA
4. MODEL APROACH
5. ACCURACY AND CONFUSSION MATRIX
6. CONCLUSIONS
7. LINK

## 1.UNDERSTANDING OF THE PROJECT

The project consists in predict the results of the match (Win, Draw, Lost home), having as the input data, all the matches which happened before a date, this data was given by an enormous dataset called **European Soccer Database**.

The goal of this project is to obtain using the input data mentioned previously an accuracy of more than a 46%, this means that we need to predict with more than a 46% the result of a match.

## 2.ANALISIS OF THE DATASET

The European Soccer Database contains 7 tables, this tables are: Country, League, Match, Player, Player\_attributes, Team, Team\_attributes.

The Country, League, Player and Team tables contains extra information about their topics, like the long names of the leagues, or the names of the players.

But the interesting tables are the Match table, that contains most of the important information related with the model that we are going to develop, the player\_attributes table that contains FIFA information about the different players of the teams and the Team\_attributes that also contains interesting FIFA information.

* MATCH TABLE

As I said this table contains most of the information related with the matches, it is composed by 115 columns, and they can classify in match attributes like: id, country, league, home\_team\_id, home\_team\_goals… , it also contains predictions of betting houses in terms of Win, Draw or Lose .

* PLAYER\_ATTRIBUTES

This table has 42 columns with information recollected by the FIFA games about each player, this information will be so useful in our model for analyze how good is a team.

## 3.SELECTIN AND ANALIZYNG THE DATA

After look the tables and know which tables are going to be the more useful ones, I started with the selection of the data, for this I use sqlite3 for python.

Texto

Descripción generada automáticamente

For make it quickly and easily I reduce the match\_table to 100 data, as we can see we also introduce a date for obtain only the interesting data.

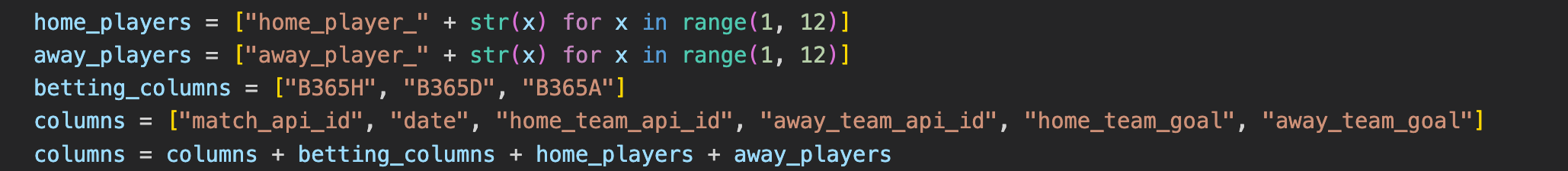
Afer this I calculate some statistics of the match table, using the describe method, this method gives us some useful information like the mean goals of the home team or the mean goals of the away team, that clearly indicates that playing at home gives them advantage on their contrary, because their mean of goals are bigger than the away team

Texto

Descripción generada automáticamente con confianza media

## 

After analyzed all the data, I decided to obtain the most useful columns and eliminate the others, I obtain the home and away player columns that I am going to use it later with the player stats, the bet365 Win Home, Draw and Win Away fields and the most characteristic columns of the table match.



Now I am prepared for starting the model approach.

## 4.MODEL APROACH

After selecting all the interesting columns, I started to analyze what I can do with them so I started to create the features, for this I do two things first obtain the last 10 matches of home\_team and away\_team, and then the home\_team and away\_team against each other 5 last matches, with this I obtain the data that I needed for calculate the features that I want, that are the number of goals scored by the home\_team in that 10 matches , the number of goals scored by the away\_team in that last 10 matches and then the goals scored against each other.

With all this data collected I created the features that are:

* home\_team\_goals\_difference
* away\_team\_goals\_difference
* game\_against\_won
* game\_against\_lost
* home\_team\_games\_won
* away\_team\_games\_won

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

After obtain this data, I prove my model with the random forest classificator, and I obtain a 0.375, so this was insufficient for the requirements, after this I was analyzing the rest of the tables searching, data useful for improving my model, and I saw the player\_attributes table, after see this table and understand it, I decided to use that info and calculate some overall rating for each player of the team generating with this useful information for my model, creating with this a new table called player stats that contains all the ratings for each player in a match.

Una captura de pantalla de una computadora

Descripción generada automáticamente con confianza media

After this I merge both tables, the match features table and the player stats table and I use random forest once again, observing that I obtain a 0.55 that is a good result.

Texto

Descripción generada automáticamente

## 5.ACCURACY AND CONFUSION MATRIX

After analyzing all the data and obtain the features, I prove some different classifiers like (Gaussian NB , KNN Neighbors …. ) and I obtain a 0.6 in terms of accuracy using Random Forest with 100 estimators, after this I prepare the confusion matrix function and display it.

Texto

Descripción generada automáticamente

# CONFUSION MATRIX

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Observing the confusion matrix, we can observe that we have 3 different True Labels that are D , L and W, in the D row we can see that the model has predict that 80% of the Matches that ends in Draw, are a wining for the home\_team and the other 20% loses for the home\_team, this indicates that our algorithm is not the best for predicting Draws, perhaps if we look the L row we can see that our model predict 62.5% of the loses as lost matches, so we can say that is good predicting when the match is Lost by the home team, and for conclude we can see that the algorithm has a 100% of accuracy in terms of wining, so we can say that the got all right.

## 6.CONCLUSIONS

With the input data obtained by the selection and analyzing part I have obtained a 0.6 of accuracy of prediction for the next matches. It means that we can predict with a 60% of sure the result of the next matches. These results are because I have done a good work with the data obtaining good features that really helps the classifier in their job.

During this project I have to deal with a lot of problems, the first problem was the columns selection, because in the match table there were more than 100 columns, so I had to studied all the columns and depreciate all the columns that don´t give me relevant information for the task, after this I fight against what I think it was the most difficult problem of the project, obtain the features needed for the model, after achieve this goal, I tried different classificator and I finally choose Random Forest as the best for this model.

Finally, I have to say that I am happy with the results of the project, I think that reflects a good solution for the problem described in the first part of this document and reflect how the machine learning can help us knowing things that previously we don´t know.

## 7.LINKS

<https://github.com/domens028/FDS_Project.git>