ISyE 6740 – Spring 2021
Project Proposal (or Final Report)

Team Member Names: Cristina Lisset Zenteno Garcia

Project Title: Possibility of soccer teams of classifying to the World Cup

1. Problem Statement:

Classifying to the World Cup is one of the major events for soccer fans around the world. It consists of several rounds of matches 1 vs 1 where the teams could either lose, tie or win, and the team will receive a score and at the end the first five with the highest score will qualify to be part of the World Cup.

So, it would be interesting to predict the probability of a team to classify to the World Cup. There are many attributes that could influence on this, like for example the performance of the team on each season and the performance of the players, so I'll try to find some characteristics of performance from squad or players that have a high impact or can ensure a team will classify to the World Cup

The scope for this project will be limited to the World Cup Qualifier matches in South America for the last three qualifiers including the current one which is still being played.

2. Data Source:

For the project report I listed some sources to use, but the links for the detail result of the matches from previous qualifier competitions are no longer available, so it was necessary to look for a different source: https://fbref.com/ , however there was not similar information for the result of matches from the previous qualifiers, most of the detailed information is available for matches after 2019, so only available for the current qualifier.

So, the dataset and the script used to pull the data from the website is described in the appendix section as tables 1 and 2. The main dataset will be table 1 and I'll use table 2 to get the list of squads that classified for the WC.

3. Methodology

There will be at least three steps:

3.1 Processing:

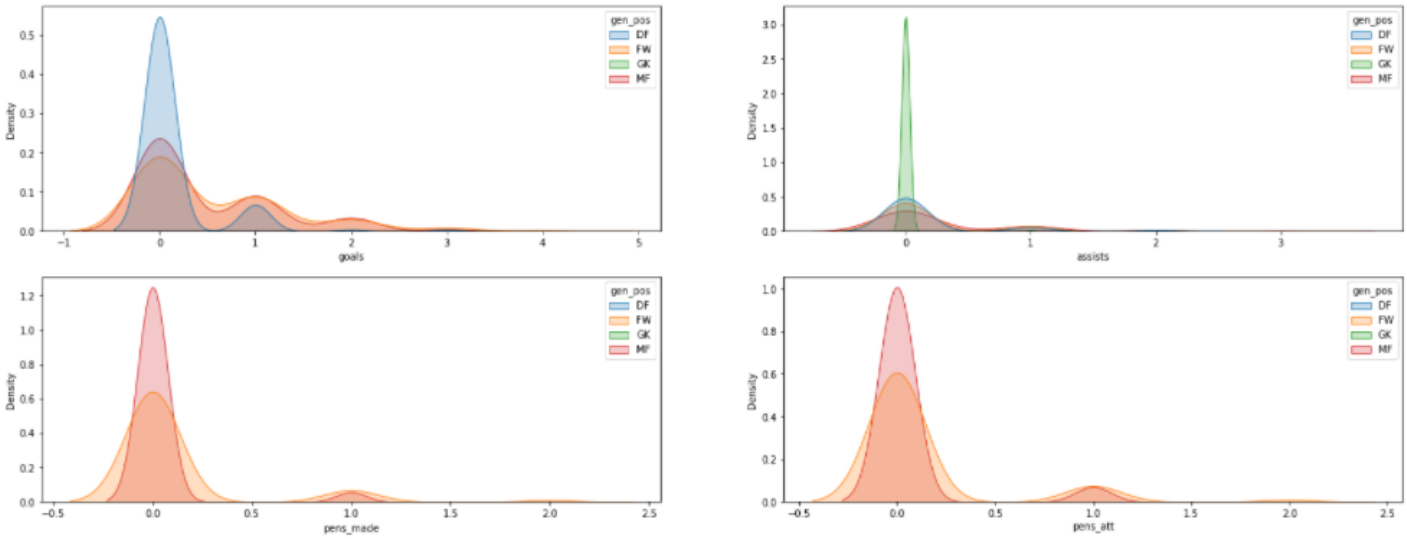
As mentioned before, this step required to develop a scrapping script to collect the data.
The cleaning process included:

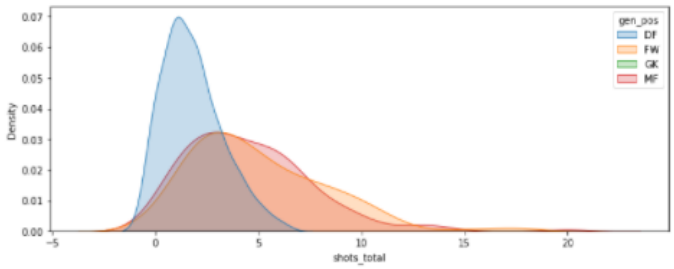
- The removal of players who were listed or called as part of the squad but during matches didn't get to play
- Removed columns that didn't have data
- Append and created columns to dataset:
  - Added the round in the qualifier. The first round from current qualifier has already ended so those matches could be used as training dataset to predict the variables needed to presume results for the second round.
  - Added a column with the points won for each match
  - Added a column with a flag equal to 1 if currently the squad is between the top five, so as for now, it would be classifying for the World cup
  - Grouped different player positions in a more general group:
    "FW","FW,MF","LW","RW" as "FW" or Forward
    "CM","MF","AM","DM","LM","RM" as "MD" or Midfielder
    "DF,MF","DF,FW","LB","CB","DF","RB" as "DF" or Defense
    "GK" stays as Goalkeeper

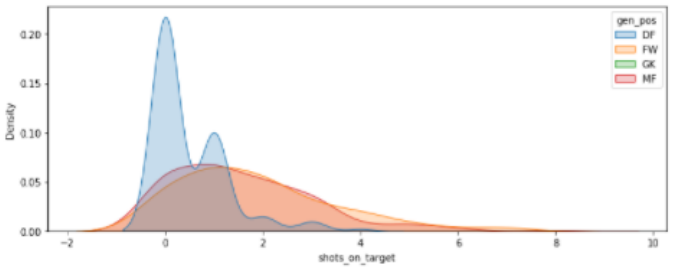
3.2 Exploration:

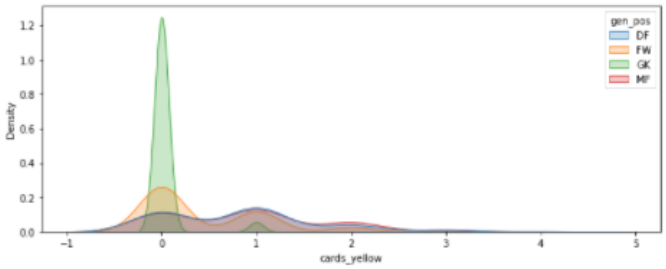
Plotted the distribution for each column to evaluate if there was a significant difference in frequency between the different group positions.

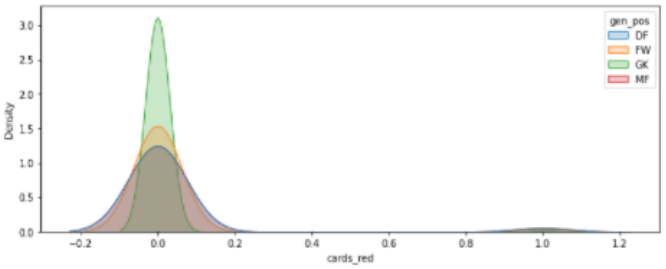
It was notorious that some variables were only related to the goalkeeper as 'own_goals', 'shots_on_target_against', 'goals_against_gk', 'saves'. The others are related to the other positions without significant difference between them, so, for the rest of the analysis I considered convenient to focus on the Forward, Midfielders and Defense positions as they have a more active performance and can have a higher influence for the result of the match. This doesn't mean the goalkeeper doesn't play an important role, but this would require a different analysis

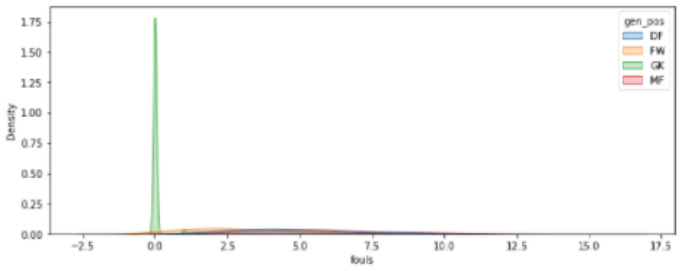


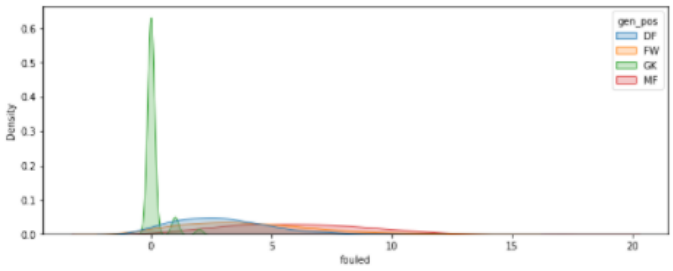

gen_pos
DF
FW
GK
MF
Density
shots_total

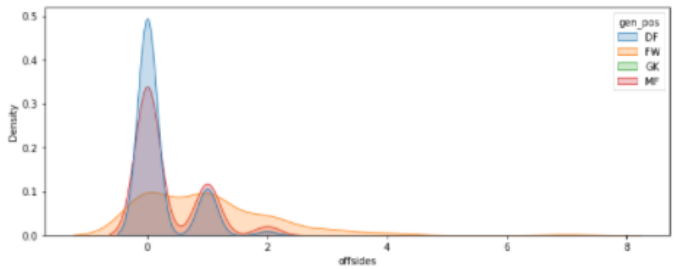

gen_pos
DF
FW
GK
MF
Density
shots_on_target

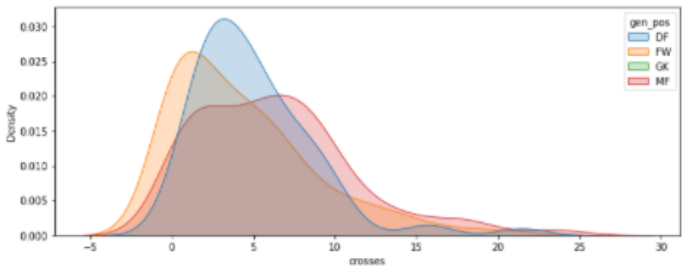

gen_pos
DF
FW
GK
MF
Density
cards_yellow

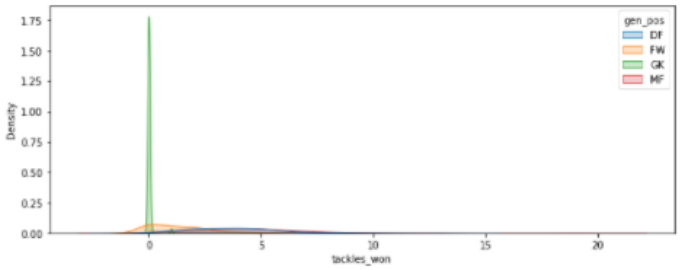

gen_pos
DF
FW
GK
MF
Density
cards_red

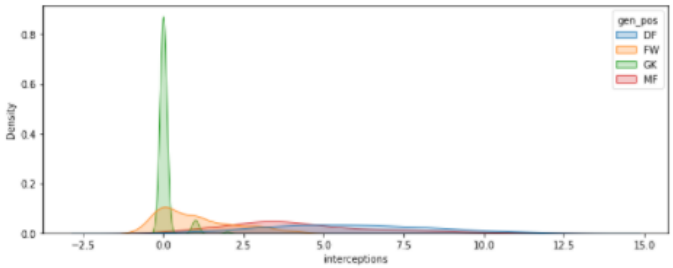

gen_pos
DF
FW
GK
MF
Density
fouls

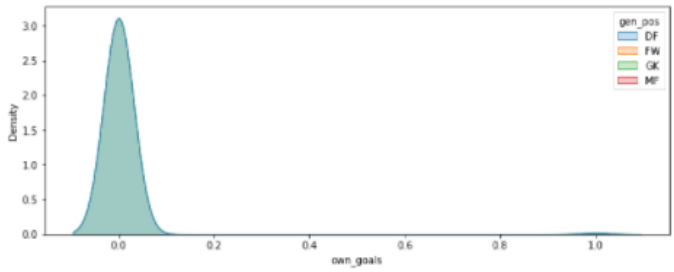

gen_pos
DF
FW
GK
MF
Density
fouled

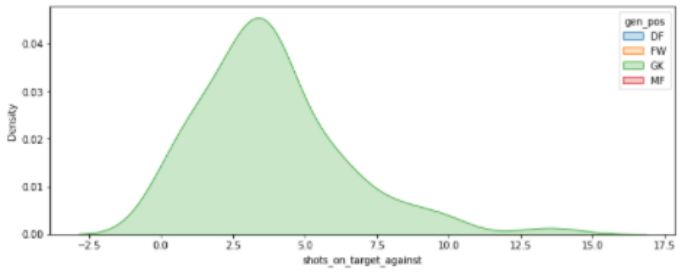

gen_pos
DF
FW
GK
MF
Density
offsides

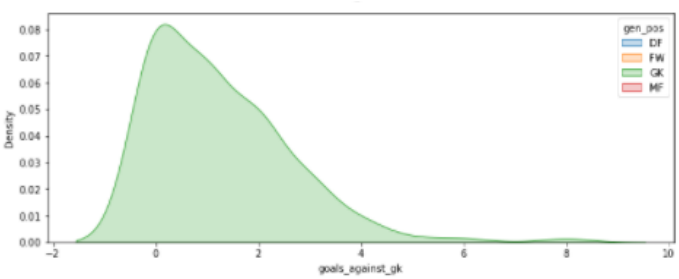

gen_pos
DF
FW
GK
MF
Density
crosses


gen_pos
DF
FW
GK
MF
Density
tackles_won


gen_pos
DF
FW
GK
MF
Density
interceptions


gen_pos
DF
FW
GK
MF
Density
own_goals


gen_pos
DF
FW
GK
MF
Density
shots_on_target_against

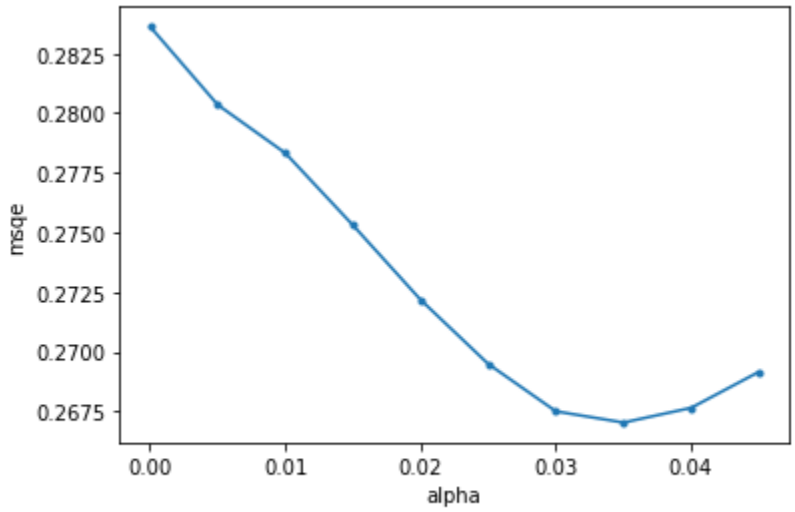


Additional variables were created as it was necessary to convert them from string to integers to use them in the models: 'venue','formation','gen_pos' using the get_dummy() function. The final variables are listed in the appendix as table 3. Also, the data was standardized

Linear Regression:

First, I used a lasso regression to select the most relevant features. I choose lasso because, as was seen in the lectures and homework, the dataset has sparse data, the values are relative low and in most cases equal to 0.

Various values of alpha were tested and the one which gave the lowest MSE was alpha = 0.035




```
Best alpha: 0.035010000000000006 , MSE: 0.26702569976739515
```

According to these, the selected features with a coefficient different to zero are the following:

```
Lasso model:
'goals: 0.034',
'assists: 0.017',
'pens_att: -0.001',
```

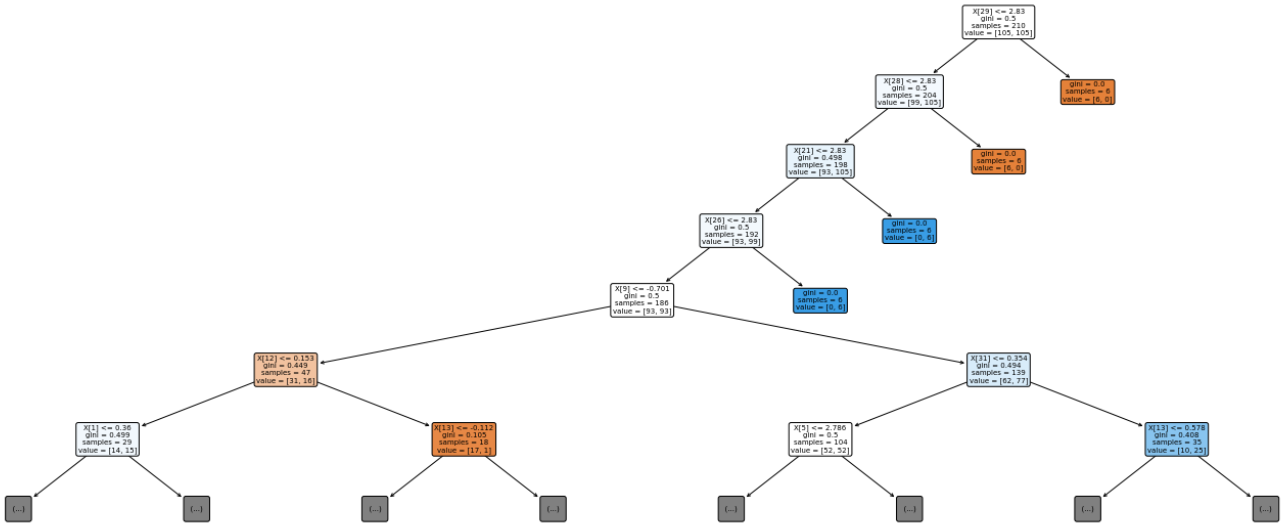
```
'fouls: 0.015',
'fouled: 0.005',
'formation_3-5-2: -0.022',
'formation_4-2-2-2: 0.047',
'formation_4-3-2-1: 0.025',
'formation_4-4-1-1: 0.046',
'formation_4-4-2: -0.024',
'formation_4-5-1: -0.047',
'formation_5-3-2: -0.048',
'formation_5-4-1: -0.022'
```

Some variables are pretty obvious for the result of the match, goals has one of the highest coefficients, but the variables related to the formations are interesting, as this could mean that squads that follow these formation could have a higher probability of qualifying for the WC.

CART:

When using CART the top levels or the tree are the variables related to the squad formations: 'formation_4-2-2-2','formation_4-5-1', 'formation_5-3-2'

So again, this could mean that the formation have a high influence over the match result and at the end over the probability of playing at the WC



4. Evaluation and Final Results:

Finally, if we see the results of the different metric for both models, we can see that CART performed slightly better than regression 58% vs 53%, which in general isn't that high.

	Accuracy	Precision	Recall	F1 score
Regression	52.94%	52.31%	34.34%	41.46%

CART	58.33%	57.29%	55.56%	56.41%

Accuracy: It's one of the main metrics that will let us see the performance of the model
Precision: It's the percentage of results that are relevant
Recall: It's the percentage of the total results correctly classified
F1 score: Also important as it's more useful than accuracy, it considers false positives and false negatives

This could have different reasons:

- A regression model is not ideal when the result has a binary value: qualified or not qualified. In this case a classification model would have been more suitable.
- CART and Lasso gave similar results in the sense of which were the most relevant variables
- The need to find more data that can explain better the results to classify for the WC. Just working with data from the current qualifier was a great restriction.

5. Appendix

5.1 Table 1

Script: https://github.com/criztinazg/wcq/blob/main/Scrapping.ipynb
Generated file: wcq_matches_logs.xlsx

COLUMN	DESCRIPTION
WCQ	world cup qualifier
DATE	date of the match
RESULT	
SQUAD_HOME_AWAY	
VENUE	
SQUAD	
OPPONENT	
FORMATION	
POSSESSION	
PLAYER	
SHIRTNUMBER	
POSITION	
AGE	
MINUTES	
GOALS	# goals scored in match
ASSISTS	# assists in match
PENS_MADE	# penalties made
PENS_ATT	# penalties attempted
SHOTS_TOTAL	# total shots In match
SHOTS_ON_TARGET	# total shots to the goal post
CARDS_YELLOW	# yellow cards
CARDS_RED	# red cards

FOULS	# fouls
FOULED	# times the player was fouled
OFFSIDES	# offsides
CROSSES	# crosses
TACKLES_WON	# tackles_won
INTERCEPTIONS	# interceptions
OWN_GOALS	# own goals
PENS_WON	# penalties the goalkeeper stopped
PENS_CONCEDED	# penalties the goalkeeper couldn't stop
SHOTS_ON_TARGET_AGAINST	# shots to the goal post made by opposite team
GOALS_AGAINST_GK	# goals scored in match by opposite squad
SAVES	# shots to the goal post made by opposite team saved by goalkeeper
SAVE_PCT	% shots to the goal post saved by goalkeeper

5.2 Table 2

Script: https://github.com/criztinazg/wcq/blob/main/Scrapping.ipynb
Generated file: wcq_final.xlsx

COLUMNS	DESCRIPTION
WCQ	**world cup qualifier**
RANK	
SQUAD	
SQUAD_DETAIL_LINK	
GAMES	**#matches played**
WINS	#matches won
DRAWS	#matches tied
LOSSES	#matches lost
GOALS_FOR	#goals in favor
GOALS_AGAINST	#goals against
GOAL_DIFF	#difference in goals, in favor minus against
POINTS	#total points

5.3 Table 3

Dataset created for the models

COLUMNS	DESCRIPTION
GOALS	# goals scored in match
ASSISTS	# assists in match
PENS_MADE	# penalties made
PENS_ATT	# penalties attempted
SHOTS_TOTAL	# total shots In match
SHOTS_ON_TARGET	# total shots to the goal post

CARDS_YELLOW	# yellow cards
CARDS_RED	# red cards
FOULS	# fouls
FOULED	# times the player was fouled
OFFSIDES	# offsides
CROSSES	# crosses
TACKLES_WON	# tackles_won
INTERCEPTIONS	# interceptions
VENUE_AWAY	# squad played as visitant
VENUE_HOME	# squad played as host
FORMATION_3-1-4-2	# dummy variables for squad formation
FORMATION_3-4-1-2	
FORMATION_3-4-3	
FORMATION_3-5-2	
FORMATION_4-1-4-1	
FORMATION_4-2-2-2	
FORMATION_4-2-3-1	
FORMATION_4-3-1-2	
FORMATION_4-3-2-1	
FORMATION_4-3-3	
FORMATION_4-4-1-1	
FORMATION_4-4-2	
FORMATION_4-5-1	
FORMATION_5-3-2	
FORMATION_5-4-1	
GEN_POS_DF	# dummy variables for player position
GEN_POS_FW	
GEN_POS_MF	
QUALIFIED	flag squad is between the top 5 and is a candidate to play at the World Cup

5.4 Code

Script: https://github.com/criztinazg/wcq/blob/main/project.ipynb