Statistical Learning Projekt Arbeit

World Cup Ergebnisse

Younes Bouras

Januar 2023

Dozentin: Prof. Dr. Christina Erlwein-Sayer

Contents

1 Introduction

Our lives today have become easier in many ways, a big reason for that is the incorporation of different analytical programming languages in our day to day lives, these languages help us by providing tools and methods to analyze complex and huge sums of data. Through these analyses we can prepare a population of the data (which is called the “training data”) with which we can build predictive Models that can forecast future events and outcomes. The reason for that is to help organizations plan and prepare for future events.

There are many other ways that using analytical programming Languages have helped us in our lives, one example of which is automating processes in order to avoid everyday repetitive tasks that were previously done manually.

R, the language that we used in our study is the one of the most used and powerful tools today for data analysis, visualization and modeling, it has become a popular choice to go to for data scientists. One of the main advantages of it is its many built-in and external packages and libraries, than can be easily installed and used, an example of these libraries is “dplyr” which provides a set of tools to manipulate and edit the data into our likings.

In our study, we focused on analyzing the data from our database (History of the World Cups that happened from 1930 until 2018) which was consisted of 27 data frames based on different subjects, such as all of the goals scored in the world cup dating all the way back to the first goal in 1930 , which was scored by “Laurent Lucien” for his national French team in a game that ended 4-1 for the blues against Mexico.

After we were done with the analytics part of our study we focused on introducing two methods of machine learning in order to build models, the so-called Neural Network and Decision Trees that can predict outcomes. We then compared both of them at the end to each other to see which method was the best one in our case. But before we talk more about the technical part of the study, let us take an overall look on the world cup.

2 The World Cup

The FIFA World Cup is the most watched and followed sporting event in the world. It is an international football competition in which the national teams of FIFA (Fédération Internationale de Football Association) participate against each other. There are 5 stages in the World Cup: group stage, round of 16, quarter-final, semi-final and final.

The first World Cup was held in 1930 in Uruguay. And since then the tournament has been held every four years, with the exception of 1942 and 1946, when it was not held due to World War II. The World Cup has grown in popularity over the years, and today it is estimated that over one billion people watched the final match of the tournament.

The 32 teams that participate are determined by a qualification phase that belongs to each continent. The tournament lasts for around a month where all the teams compete for the title in the home of the host country. So far, there has been a total of 22 tournaments in which 80 different national teams competed at.

One of the most exciting aspects of the World Cup is the diversity of the participating teams. The tournament features teams from all over the world, each with their own unique style of play and cultural background. This diversity not only makes for exciting matches but also helps to promote unity and understanding among different nations.

On a global scale, the World Cup can serve as a powerful tool for promoting unity and understanding among different nations. The tournament brings together teams from all over the world, and the diversity of the participating teams can help to break down cultural barriers and promote mutual respect and understanding. However, the World Cup can also have negative impacts. The cost of hosting the tournament can be substantial and can lead to financial strains on the host country. Additionally, the tournament can also lead to the displacement of local residents and the destruction of natural habitats.

3 Data Analysis

Before starting with our study, we need to understand what a database is. It is collection of organized data that is stored in a structured format. One of the primary uses of a database is to store, provide and manage informations and statistics based on large amounts of data. The process of transforming huge data sets into meaningful insights is called “Data Analysis”, this allows businesses and individuals to gain insights, make data driven decisions and improve operations.

In our Database, we had 27 different data frames, some of which that we will be using in our data analysis, visualization part are :

- All goals that were scored in the history of the World Cup
- All of the Players that played at least a game of the World Cup
- All of the Bookings (yellow cards and red cards)
- The tournament standings (top 4 Teams) of all World Cups dating from 1930 to 2018
- All of the Stadiums that were a part of the World Cup
- All of the World Cup champions
- All of the teams that participated at least once in the World Cup.

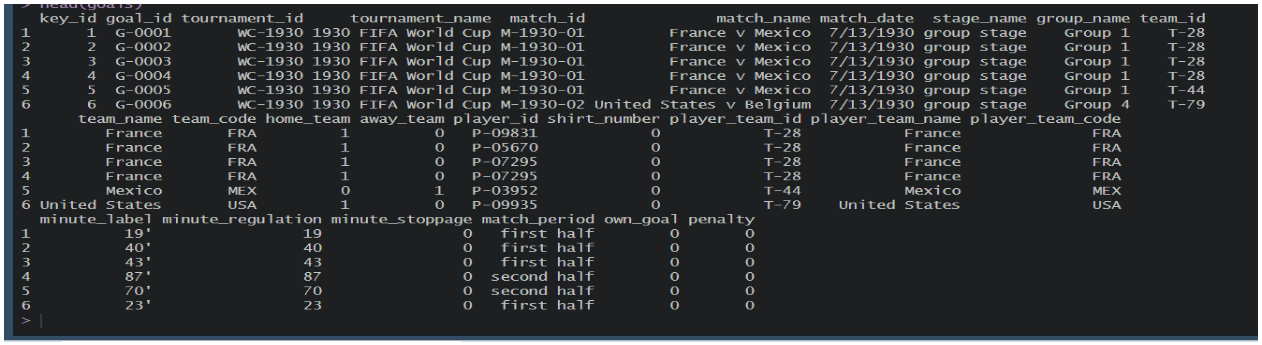
As it was stated in the introduction, we have used the programming language R for our data analysis. So in order to start we needed to import the following packages and libraries into R :

- readr : in order to quickly import big sums of data into R
- lifecycle : in order to successfully install the package dplyr
- dplyr : in order to easily analyse the data using different functions, some of wich are the group_by, select, arrange etc.

I/ Top Goal Scorers:

Thierry Henry, one of greatest football players once said “It's not about me, it's about how I can help my team to achieve more. And I do that through scoring goals”. When we think of football, the first thing that comes to mind is trying to figure out the players that scored the most amount of goals, that is why in order to figure it out we needed to access 2 of our data frames. “Goals” and “Players”, so let us check out what both of them had in store for us.

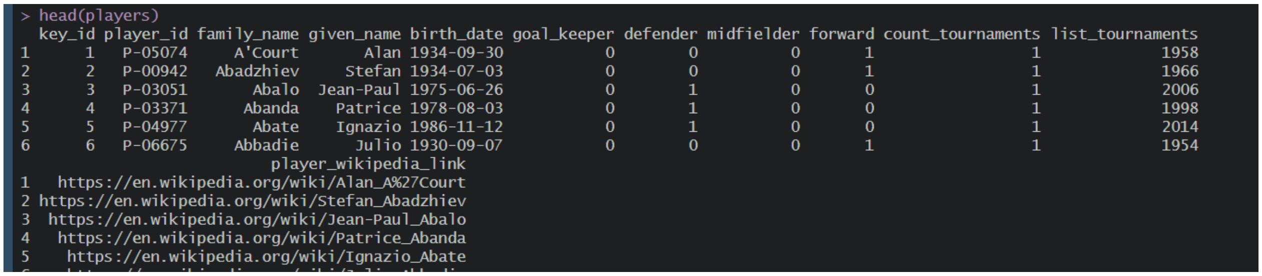
1/Goals:



```
> head(goals)
  key_id goal_id tournament_id     tournament_name  match_id              match_name match_date  stage_name group_name team_id
1      1  G-0001       WC-1930 1930 FIFA World Cup M-1930-01         France v Mexico  7/13/1930 group stage    Group 1    T-28
2      2  G-0002       WC-1930 1930 FIFA World Cup M-1930-01         France v Mexico  7/13/1930 group stage    Group 1    T-28
3      3  G-0003       WC-1930 1930 FIFA World Cup M-1930-01         France v Mexico  7/13/1930 group stage    Group 1    T-28
4      4  G-0004       WC-1930 1930 FIFA World Cup M-1930-01         France v Mexico  7/13/1930 group stage    Group 1    T-28
5      5  G-0005       WC-1930 1930 FIFA World Cup M-1930-01         France v Mexico  7/13/1930 group stage    Group 1    T-44
6      6  G-0006       WC-1930 1930 FIFA World Cup M-1930-02 United States v Belgium  7/13/1930 group stage    Group 4    T-79
      team_name team_code home_team away_team player_id shirt_number player_team_id player_team_name player_team_code
1        France       FRA         1         0   P-09831            0           T-28           France              FRA
2        France       FRA         1         0   P-05670            0           T-28           France              FRA
3        France       FRA         1         0   P-07295            0           T-28           France              FRA
4        France       FRA         1         0   P-07295            0           T-28           France              FRA
5        Mexico       MEX         0         1   P-03952            0           T-44           Mexico              MEX
6 United States       USA         1         0   P-09935            0           T-79    United States              USA
  minute_label minute_regulation minute_stoppage match_period own_goal penalty
1          19'                19               0   first half        0       0
2          40'                40               0   first half        0       0
3          43'                43               0   first half        0       0
4          87'                87               0  second half        0       0
5          70'                70               0  second half        0       0
6          23'                23               0   first half        0       0
> 
```

This data frame showed us every goal that was scored in the history of the World Cup, the important columns for us were player_id.

2/Players:

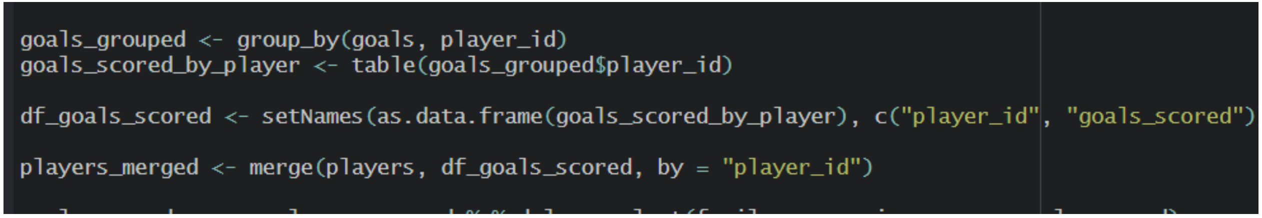


```
> head(players)
  key_id player_id family_name given_name birth_date goal_keeper defender midfielder forward count_tournaments list_tournaments
1      1   P-05074     A'Court       Alan 1934-09-30           0        0          0       1                 1             1958
2      2   P-00942   Abadzhiev     Stefan 1934-07-03           0        0          0       1                 1             1966
3      3   P-03051       Abalo  Jean-Paul 1975-06-26           0        1          0       0                 1             2006
4      4   P-03371      Abanda    Patrice 1978-08-03           0        1          0       0                 1             1998
5      5   P-04977       Abate    Ignazio 1986-11-12           0        1          0       0                 1             2014
6      6   P-06675     Abbadie     Julio 1930-09-07            0        0          0       1                 1             1954
                      player_wikipedia_link
1   https://en.wikipedia.org/wiki/Alan_A%27Court
2 https://en.wikipedia.org/wiki/Stefan_Abadzhiev
3  https://en.wikipedia.org/wiki/Jean-Paul_Abalo
4   https://en.wikipedia.org/wiki/Patrice_Abanda
5    https://en.wikipedia.org/wiki/Ignazio_Abate
```

This data frame showed us every player that played at least a game in the history of the World Cup, the important columns for us were the name columns and the player_id.

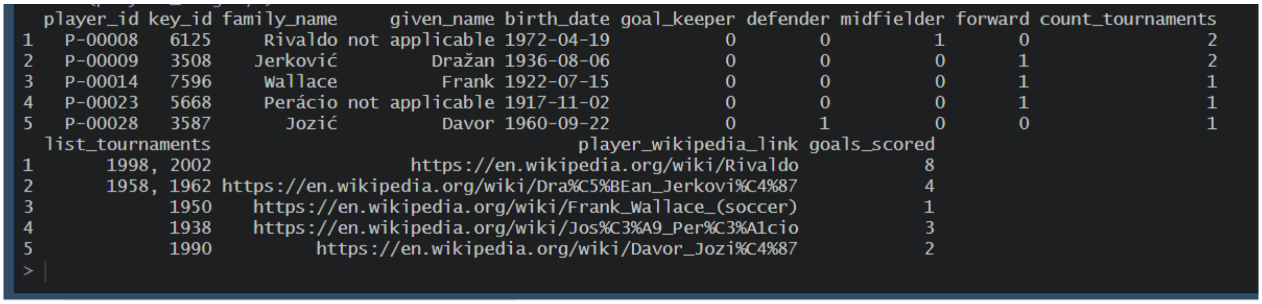
The idea to see how many goals each players scored, is to group the data frame “goals” by the column “player_id” which will give us a table “goals_scored_by_player” showing us how many times a unique player_id value occurred in the data frame, which is perfect for us since we are looking for the exact same thing.

After that, we are going to have to merge our result with the data frame “players” in order to add a new column to it, which we will call “goals_scored” which contains the amount of times as previously mentioned a single unique “player_id” value occurred in “goals”. But to be able to do that we have to transform our table into a data frame using the as.data.frame() function.



```
goals_grouped <- group_by(goals, player_id)
goals_scored_by_player <- table(goals_grouped$player_id)

df_goals_scored <- setNames(as.data.frame(goals_scored_by_player), c("player_id", "goals_scored")

players_merged <- merge(players, df_goals_scored, by = "player_id")
```

This is what the new data frame “players_merged” looked like (it consisted of “players” + the new column “goals_scored”):

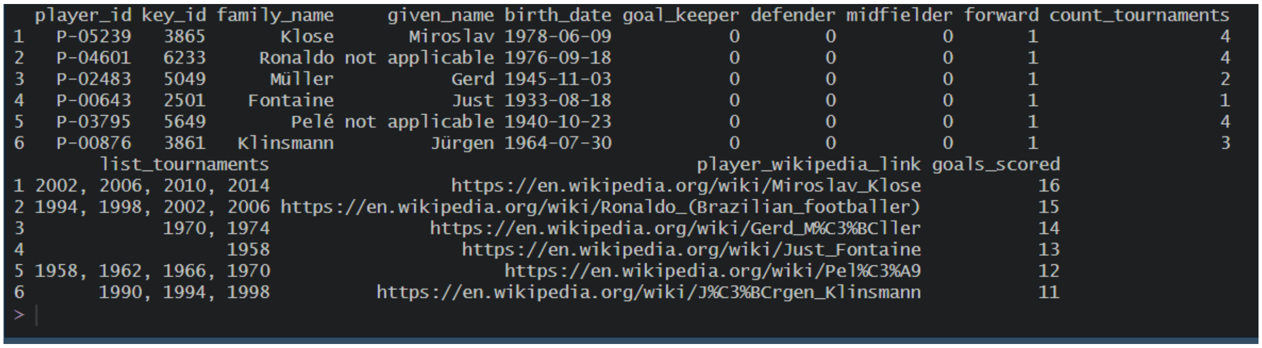


```
  player_id key_id family_name     given_name birth_date goal_keeper defender midfielder forward count_tournaments
1   P-00008   6125     Rivaldo not applicable 1972-04-19           0        0          1       0                 2
2   P-00009   3508    Jerković         Dražan 1936-08-06           0        0          0       1                 2
3   P-00014   7596     Wallace          Frank 1922-07-15           0        0          0       1                 1
4   P-00023   5668     Perácio not applicable 1917-11-02           0        0          0       1                 1
5   P-00028   3587       Jozić          Davor 1960-09-22           0        1          0       0                 1
  list_tournaments                          player_wikipedia_link goals_scored
1       1998, 2002              https://en.wikipedia.org/wiki/Rivaldo            8
2       1958, 1962 https://en.wikipedia.org/wiki/Dra%C5%BEan_Jerkovi%C4%87            4
3             1950    https://en.wikipedia.org/wiki/Frank_Wallace_(soccer)            1
4             1938    https://en.wikipedia.org/wiki/Jos%C3%A9_Per%C3%A1cio            3
5             1990          https://en.wikipedia.org/wiki/Davor_Jozi%C4%87            2
> 
```

To sum up what we did so far, we have extracted how many times every unique “player_id” value (from “goals”) has occurred and we have added the column “goals_scored” in the data frame which contained how many goals each player scored in the world cup.

Since we were looking for the players that scored the most amount of goals, we had to sort our new data frame in a descending order by using the arrange(desc()) function. Meaning that the line of top scorer of the world cup history will be the new first line in our data frame, and the second top scorer's line will be the new second line and so on.

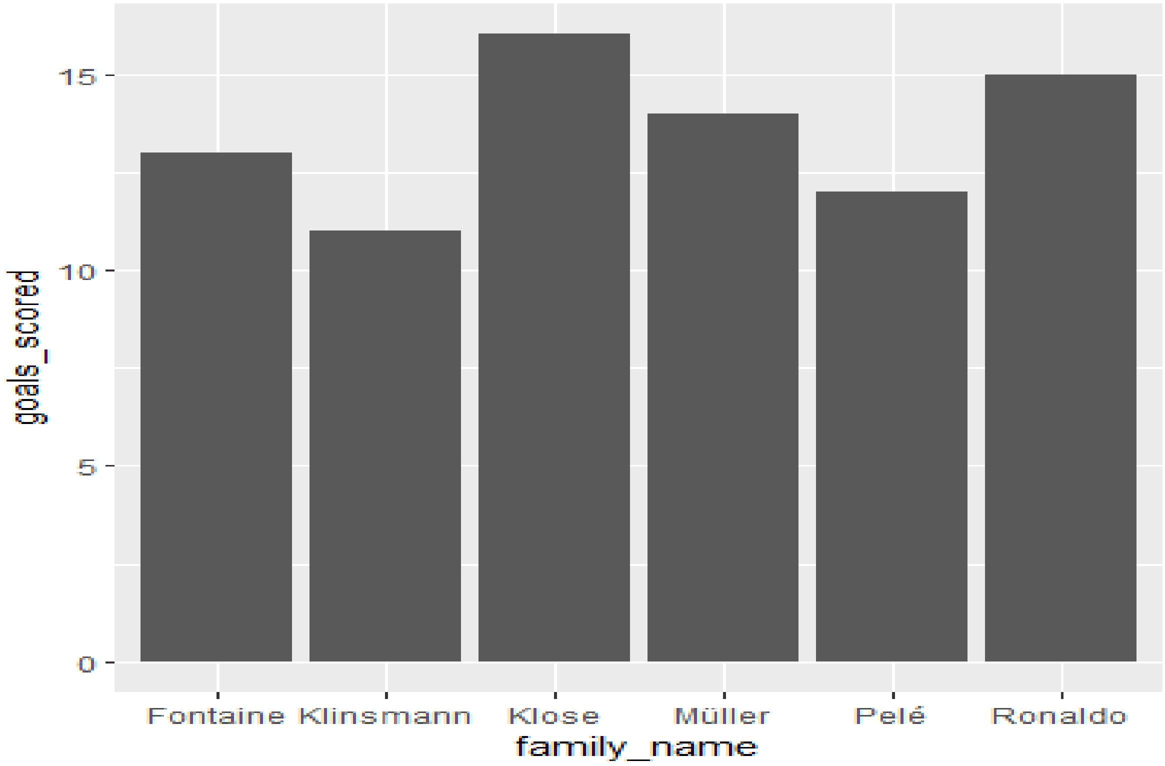
Finally, we had a data frame that looked like this :



```
  player_id key_id family_name     given_name birth_date goal_keeper defender midfielder forward count_tournaments
1   P-05239   3865       Klose       Miroslav 1978-06-09           0        0          0       1                 4
2   P-04601   6233     Ronaldo not applicable 1976-09-18           0        0          0       1                 4
3   P-02483   5049      Müller           Gerd 1945-11-03           0        0          0       1                 2
4   P-00643   2501    Fontaine           Just 1933-08-18           0        0          0       1                 1
5   P-03795   5649        Pelé not applicable 1940-10-23           0        0          0       1                 4
6   P-00876   3861   Klinsmann         Jürgen 1964-07-30           0        0          0       1                 3
        list_tournaments                                        player_wikipedia_link goals_scored
1 2002, 2006, 2010, 2014                 https://en.wikipedia.org/wiki/Miroslav_Klose           16
2 1994, 1998, 2002, 2006 https://en.wikipedia.org/wiki/Ronaldo_(Brazilian_footballer)           15
3             1970, 1974               https://en.wikipedia.org/wiki/Gerd_M%C3%BCller           14
4                   1958                  https://en.wikipedia.org/wiki/Just_Fontaine           13
5 1958, 1962, 1966, 1970                      https://en.wikipedia.org/wiki/Pel%C3%A9           12
6       1990, 1994, 1998        https://en.wikipedia.org/wiki/J%C3%BCrgen_Klinsmann           11
> 
```

From which we could see, that the player with the most amount of goals scored was "Miroslav Klose". The German player scored a total amount of 16 goals in the history of the world cup and after him came the Brazilian legend Ronaldo with 15 goals.

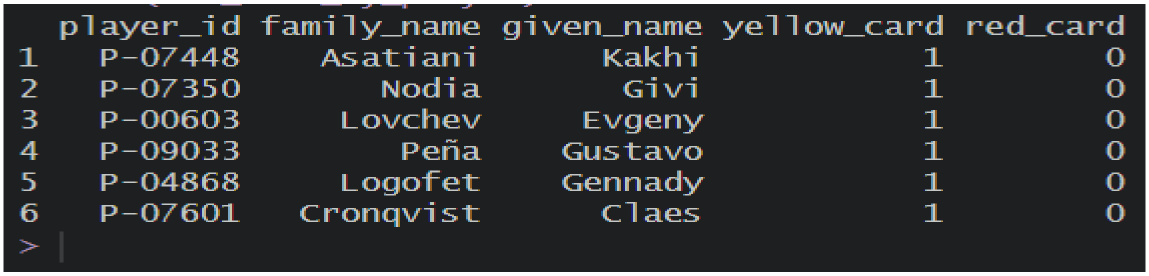
For the sake of making the most out of our analysis, we had to visualize our results using the ggplot() function with "family_name" as our X axis and "goals_scored" as our Y axis.



II/ Top Bookings:

Before starting our second analysis part, it is important to understand what the term “booking” means in football. A booking refers to when a player is shown a yellow or red card by the referee for committing a foul or unsportsmanlike behavior. A yellow card is a warning and a red card results in the player being sent off and unable to play for the remainder of the game. Two yellow cards in a match result in a red card.

With that being said, let us check out what the important columns of the data frame “bookings” look like :

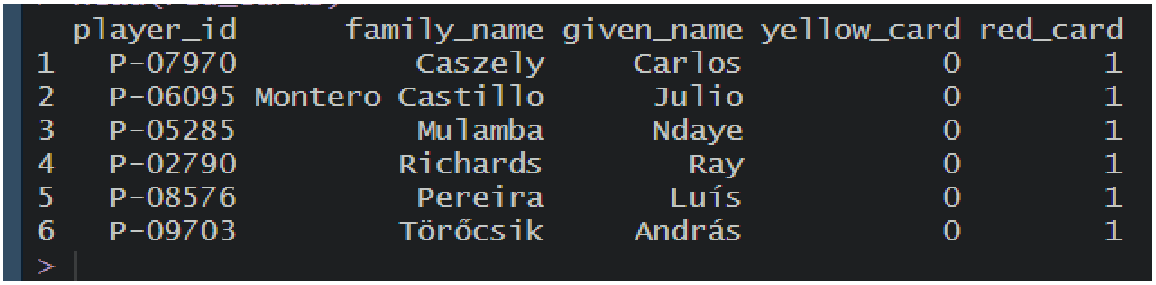


```
  player_id family_name given_name yellow_card red_card
1   P-07448    Asatiani      Kakhi           1        0
2   P-07350       Nodia       Givi           1        0
3   P-00603     Lovchev     Evgeny           1        0
4   P-09033        Peña    Gustavo           1        0
5   P-04868     Logofet    Gennady           1        0
6   P-07601   Cronqvist      Claes           1        0
> 
```

If we take a look at the two last two columns, we can see that the booking was either a yellow card or a red card.

Therefore, after understanding everything, we have decided to focus on the players that were given the most amount of red cards. In order to do that, we needed to extract the lines of the data frame which state that the booking was a red card and not a yellow car and add them into a new data frame called “red_cards”. For this, we needed to use the “%>% filter(red_card > 0)” function (that belongs to the dplyr package we talked about earlier).

This is what “red_cards” looks like :

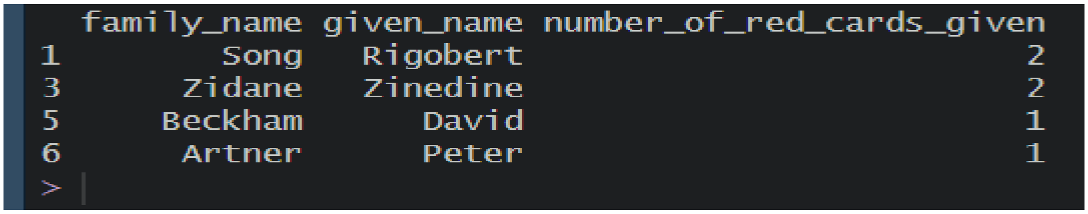


```
  player_id      family_name given_name yellow_card red_card
1   P-07970          Caszely     Carlos           0        1
2   P-06095 Montero Castillo      Julio           0        1
3   P-05285          Mulamba      Ndaye           0        1
4   P-02790         Richards        Ray           0        1
5   P-08576          Pereira       Luís           0        1
6   P-09703         Törőcsik     András           0        1
> 
```

We can actually use the same method we used for the first part of our data analysis, meaning we can group this data frame by “player_id” and merge it to “players” with adding a new column stating how many times the player was given a red cards. But since we had a wide range of functions from the installed packages we decided to use the count() function instead.

Even though the function is a different one but the direction we are going is actually the same, we counted how many times a unique “player_id” occurred in “red_cards” then we merged our result with the “bookings” data frame by adding a new column "number_of_red_cards_given", then selecting the important columns which are : family_name, given_name, number_of_red_cards_given and finally printing the results in a descending way to show us the players with the most amount of red cards.

So here were the results :



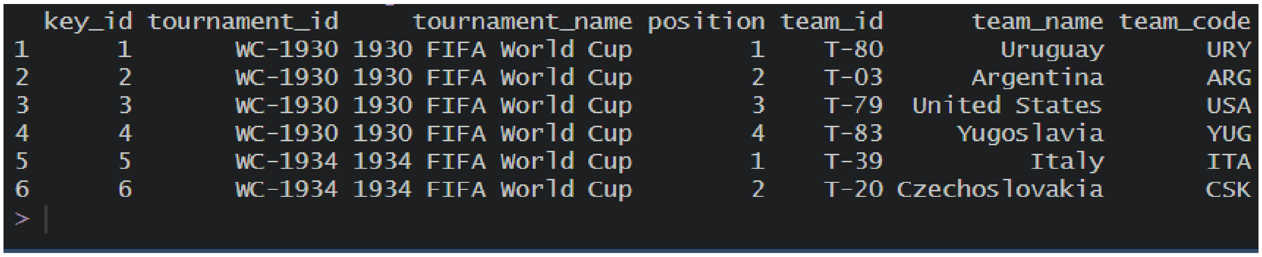
```
  family_name given_name number_of_red_cards_given
1        Song   Rigobert                         2
3      Zidane   Zinedine                         2
5     Beckham      David                         1
6      Artner      Peter                         1
> 
```

“Song Rigobert” and “Zinedine Zidane” were the two players with the most amount of red cards. Both players received 2 red cards in the history of the world cup.

III/ World Cup Champions:

Participating in the World Cup is an honor for national teams, as it is an opportunity to compete against the best teams in the world and showcase their talent on the global stage. Winning the World Cup is considered the ultimate achievement in football, and it can bring immense pride and joy to a country and its people.

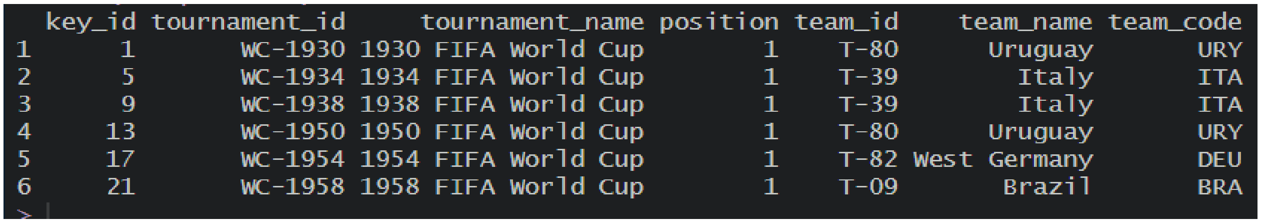
Therefore, we couldn’t make a whole study about the World Cup without covering which teams had the privilege of calling themselves forever as world champions. So let us see what our data frame “tournament_standings” had in store for us through the head() function:



```
  key_id tournament_id     tournament_name position team_id      team_name team_code
1      1       WC-1930 1930 FIFA World Cup        1    T-80        Uruguay       URY
2      2       WC-1930 1930 FIFA World Cup        2    T-03      Argentina       ARG
3      3       WC-1930 1930 FIFA World Cup        3    T-79  United States       USA
4      4       WC-1930 1930 FIFA World Cup        4    T-83     Yugoslavia       YUG
5      5       WC-1934 1934 FIFA World Cup        1    T-39          Italy       ITA
6      6       WC-1934 1934 FIFA World Cup        2    T-20 Czechoslovakia       CSK
> 
```

From what we can see, each line covered a team that made it to the top 4 of all tournaments that happened. If we take a look at the “position” column, we can figure out that the team with the position 1 won the World Cup of the year which was mentioned in the “tournament_name” column. For example (line 5) : “Italy” won the 1934 World Cup.

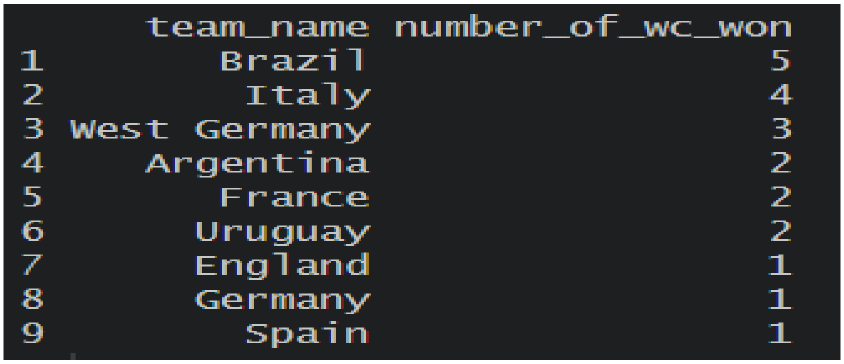
So to start with the process of extracting the teams that became champions, we must use the filter function in the following way “%>% filter(position == 1)” and create a new data frame “df_champions” with it.



```
  key_id tournament_id     tournament_name position team_id    team_name team_code
1      1       WC-1930 1930 FIFA World Cup        1    T-80      Uruguay       URY
2      5       WC-1934 1934 FIFA World Cup        1    T-39        Italy       ITA
3      9       WC-1938 1938 FIFA World Cup        1    T-39        Italy       ITA
4     13       WC-1950 1950 FIFA World Cup        1    T-80      Uruguay       URY
5     17       WC-1954 1954 FIFA World Cup        1    T-82 West Germany       DEU
6     21       WC-1958 1958 FIFA World Cup        1    T-09       Brazil       BRA
> 
```

And now the process of figuring out how many times each team won the tournament is the exact same like the part with that of counting how many times players received red cards. It is through the count()function (with the arrange function).

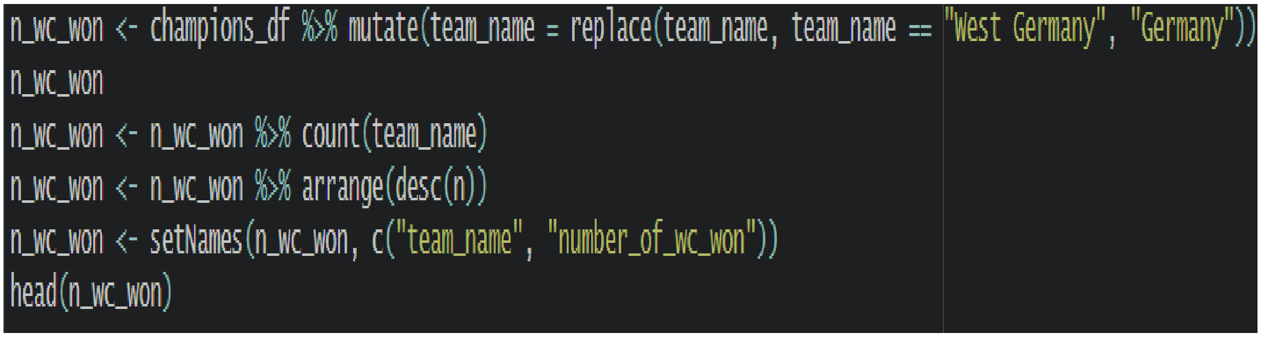
So let us take a look at the code and the result.



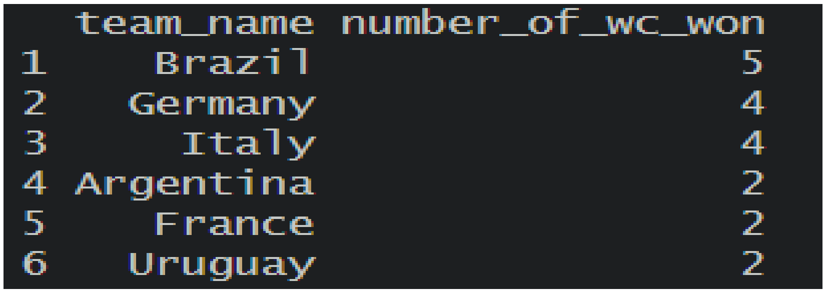
```
     team_name number_of_wc_won
1       Brazil                5
2        Italy                4
3 West Germany                3
4    Argentina                2
5       France                2
6      Uruguay                2
7      England                1
8      Germany                1
9        Spain                1
```

We can see that Brazil was the team that won the most amount of World Cup trophies, but there was something tricky about the result. The “West Germany” and “Germany” values had no correlation between each other even though it is technically the same country.

That is why we needed to use the mutate() function in order to replace “West Germany” to “Germany” so that both values could be counted as one.



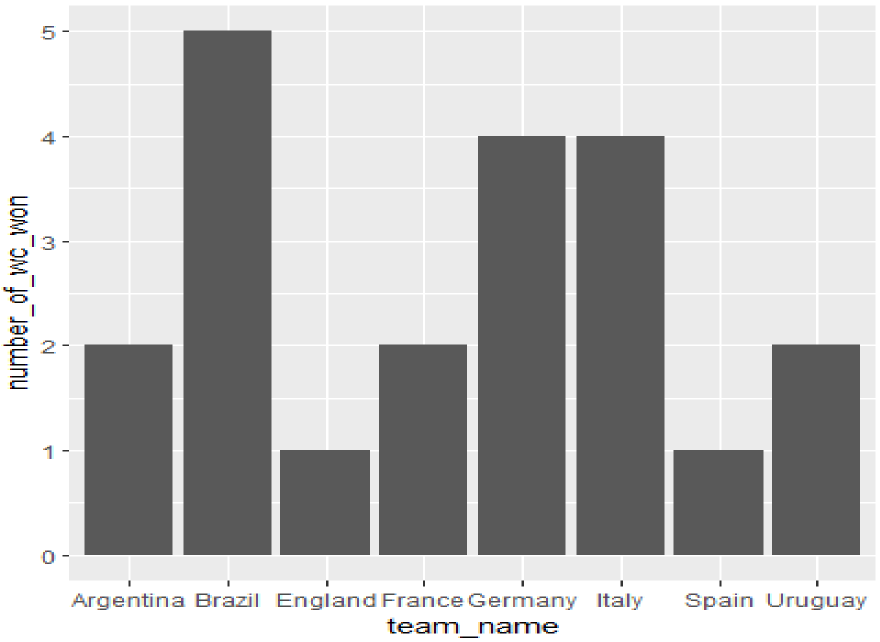
```
n_wc_won <- champions_df %>% mutate(team_name = replace(team_name, team_name == "West Germany", "Germany"))
n_wc_won
n_wc_won <- n_wc_won %>% count(team_name)
n_wc_won <- n_wc_won %>% arrange(desc(n))
n_wc_won <- setNames(n_wc_won, c("team_name", "number_of_wc_won"))
head(n_wc_won)
```



```
  team_name number_of_wc_won
1    Brazil                5
2   Germany                4
3     Italy                4
4 Argentina                2
5    France                2
6   Uruguay                2
```

Finally, we were able to see that “West Germany” and “Germany” were considered as one value, and the “number of world cup won” for it was : 3+1 = 4.

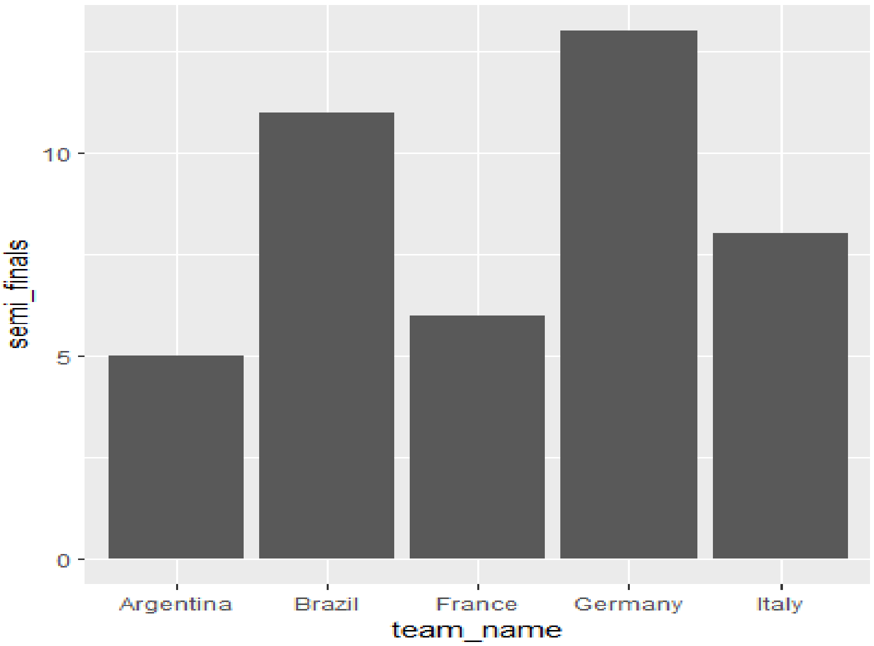
So now we can visualize our results in the following diagram:



IV/The 4 Teams that made it the most to the Semi Final:

The data frame that we used for this process is actually the same like the one we had used for finding out which teams won the competition (“tournament_standings”).

Figuring out the 4 teams that made it the most to the Semi Final is done through replacing “West Germany” to “Germany”, counting the data frame by the team_name value and then sorting the result in a descending form while extracting the first 4 lines at the end through the head(,4) function.

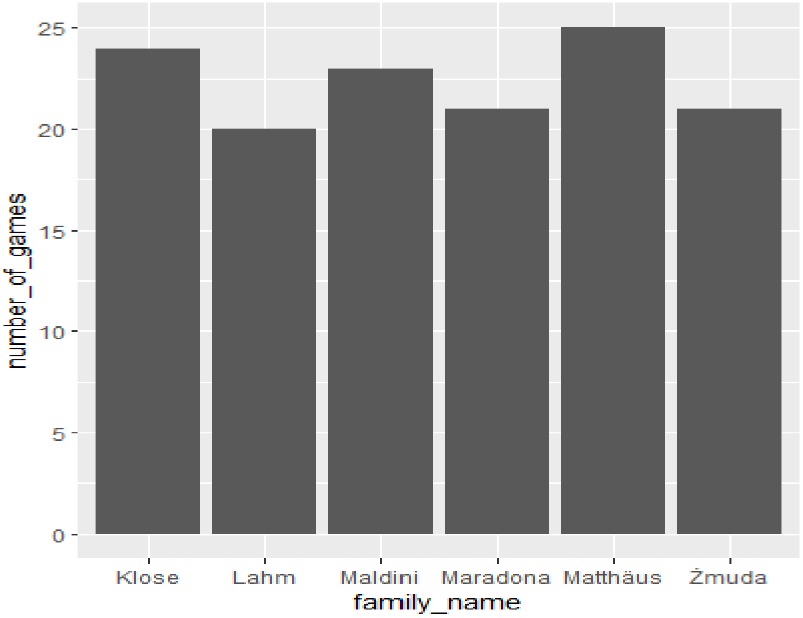


V/ Players with the most amount of games played:

The data frame “player_appearances” was the one that helped us figure out the desired results.

The process was done through extracting the important columns, counting the data frame by the “player_id”, then merging the result with the original data frame. However the output of our data frame seemed to have a lot of duplicates, this is why we had to remove them through the: “players_games_merged <- players_games_merged[!duplicated(players_games_merged),]” command.

In order to check the top results, we then sorted our final result.



4 Machine Learning Methods

Machine learning is a subset of artificial intelligence (AI). It is focused on teaching computers to learn from data and to improve with experience instead of being precisely programmed to do so. In machine learning, algorithms are trained to find patterns and correlations in large data sets and to make the best decisions and predictions based on that analysis.

Machine learning is consisted of 3 different types of algorithms : reinforced, unsupervised and **supervised learning**. The one that we decided to cover was the supervised learning.

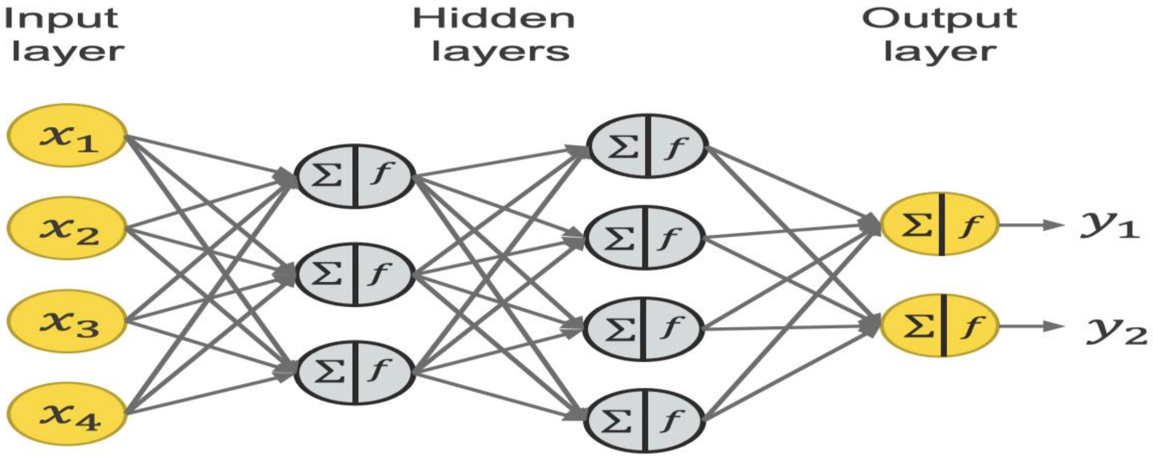
Supervised learning is one of the most common types of machine learning, it involves using data to train a model to make predictions or decisions. In supervised learning, the data is labeled, meaning it has both input and output variables. The model is trained using this labeled data and the goal is to make prediction.

Supervised learning algorithms are used in many applications such as image classification, speech recognition, natural language processing, and recommendation systems. Some popular supervised learning algorithms include Decision Trees and Neural Network, which we built in our study to predict outcomes. We will explain how both methods work, how we implemented them in R, what exactly we wanted to predict and finally we will compare the results of the prediction that came from both methods.

I / Neural Network:

Neural Network is the first method of the two that we chose which belongs to supervised learning, the model's name and structure was inspired by the human brain, mimicking the way that biological neurons signal to one another.

It consists of 3 different layers of interconnected neurons that process and transmit information : input layer, hidden layer and the output layer.



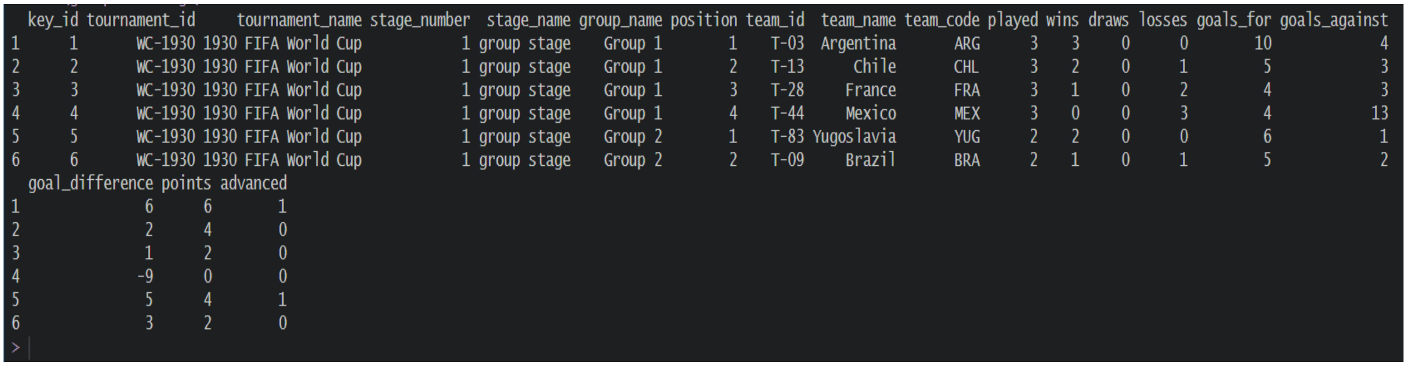
Each neuron that belong to the input layers receives information from the data (the input variables) and passes it to the next "hidden layers" where mathematical operations will be performed, to be precise weights of each connection will be counted. The closer a weight is to 0, the less role it plays in the final prediction. This process is repeated across the next layers, allowing the network to learn and represent complex relationships in the data.

Neural networks can be used for a wide range of tasks, including image recognition, natural language processing, and prediction. They are particularly well suited for tasks that involve large and complex data sets.

So now after we understood the concept of Neural Network and how the method works, we can get to the fun part of implementing it into R. We started through installing the packages and accessing the following libraries:

- NeuralNetTools: is a package which provides some tools for visualizing and interpreting neural networks. This package is designed to be used in conjunction with the next two libraries
- Neuralnet: is a package that provides functions for training a neural network on a given dataset.
- ggplot2: is a package for creating data visualizations.

The data frame we chose for the prediction was “group_standings”, so we are going to take a look at it.

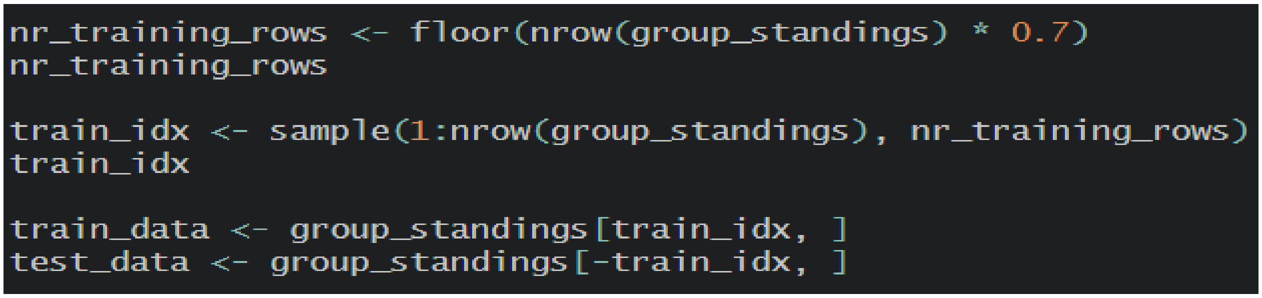


```
  key_id tournament_id     tournament_name stage_number  stage_name group_name position team_id  team_name team_code played wins draws losses goals_for goals_against
1      1       WC-1930 1930 FIFA World Cup            1 group stage    Group 1        1    T-03  Argentina       ARG      3    3     0      0        10             4
2      2       WC-1930 1930 FIFA World Cup            1 group stage    Group 1        2    T-13      Chile       CHL      3    2     0      1         5             3
3      3       WC-1930 1930 FIFA World Cup            1 group stage    Group 1        3    T-28     France       FRA      3    1     0      2         4             3
4      4       WC-1930 1930 FIFA World Cup            1 group stage    Group 1        4    T-44     Mexico       MEX      3    0     0      3         4            13
5      5       WC-1930 1930 FIFA World Cup            1 group stage    Group 2        1    T-83 Yugoslavia       YUG      2    2     0      0         6             1
6      6       WC-1930 1930 FIFA World Cup            1 group stage    Group 2        2    T-09     Brazil       BRA      2    1     0      1         5             2
  goal_difference points advanced
1               6      6        1
2               2      4        0
3               1      2        0
4              -9      0        0
5               5      4        1
6               3      2        0
>
```

The last column stated whether the team advanced from the group stage or not, using the dummy variable : 1 for true and 0 for false. And the other column from “played” to “points” were the different statistics of the team in the group stage.

So if we use our brain for a little bit, we can figure out what the output layer and the input layer(s) in this case. Through the input of the teams’ statistics (how many goals they scored, how many wins they had etc.) we can have a prediction if the team advanced to the next round or not.

Now that we have our input and output layers figured out, we needed to train our model and build it. The first step would be to prepare a Training and Testing data frame from “group_standings”, we decided to take 70 and 30 percent of the total amount of lines respectively for both data frames.

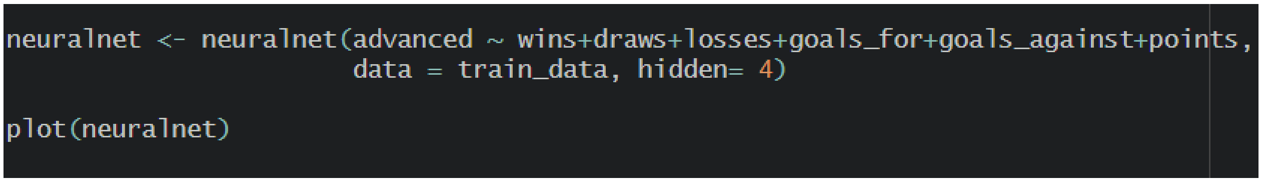


```
nr_training_rows <- floor(nrow(group_standings) * 0.7)
nr_training_rows

train_idx <- sample(1:nrow(group_standings), nr_training_rows)
train_idx

train_data <- group_standings[train_idx, ]
test_data <- group_standings[-train_idx, ]
```

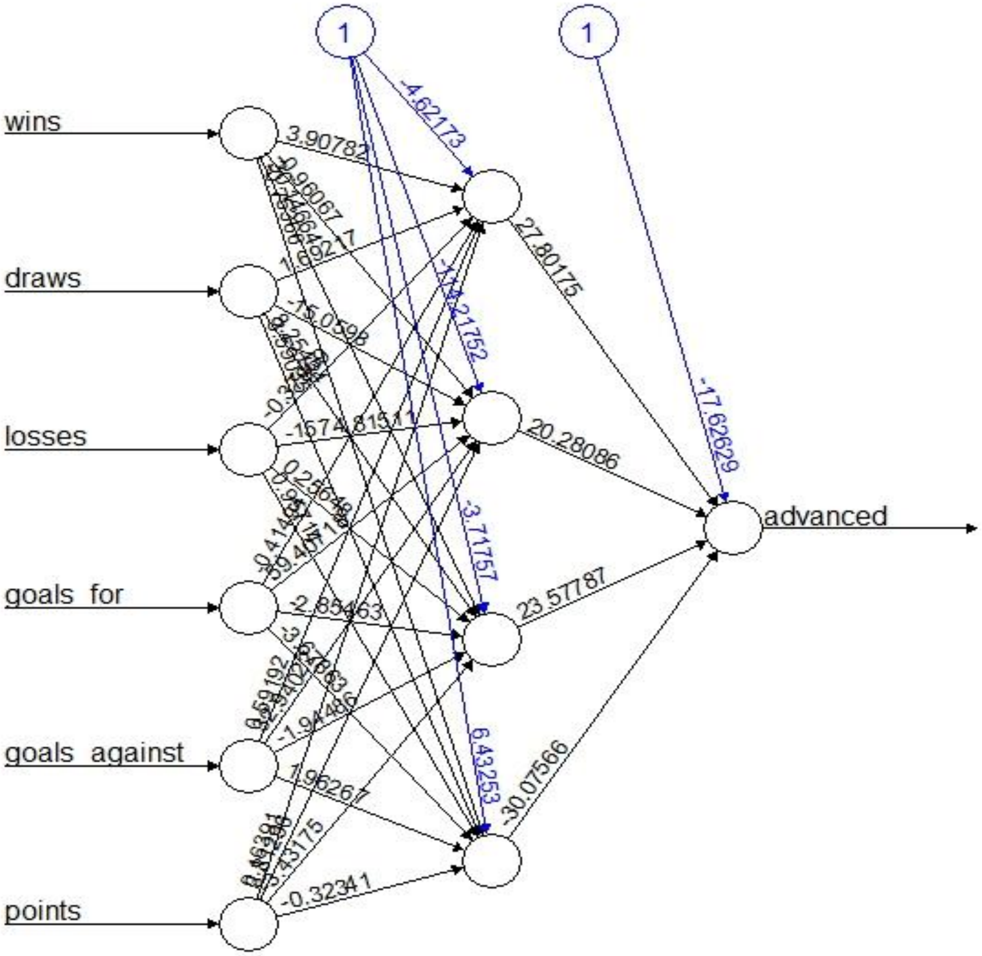
Through the “nrow(group_standings)”, R printed us how many lines the original data frame had (458). We used the floor number to round the number of 458*0.7.

The sample function helped us figure out the indexes for the training data and the testing data, in this example: “sample(1:458, 320)” gave us 320 random numbers that belonged to (1:458), which will be the indexes of the training data frame and the rest of the lines will be added to the testing data frame.

Now we can build and visualize our model through the neuralnet() function, we decided to choose 4 hidden layers.



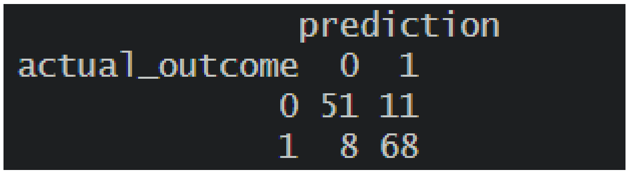
```
neuralnet <- neuralnet(advanced ~ wins+draws+losses+goals_for+goals_against+points,
                       data = train_data, hidden= 4)

plot(neuralnet)
```



Finally our model was ready to make predictions with through the predict() function. We predicted the outcome of the test data frame, which had counted the probability that a team will advance.

After that we compared our predictions with the actual outcome, but in order to do that we needed to modify the values we had from the prediction to match the values of the “advanced” column. We will change our results that were ≥ 0.5 to 1 and the rest to 0.

So here we had our comparison table :



```
              prediction
actual_outcome  0  1
             0 51 11
             1  8 68
```

Out of the 62 “0” values in the test data frame, we predicted 51 correctly. And out of the 76 “1” values we predicted 68 correctly.

So to be precise, our model was : [(51+68)/(51+11+8+68)]*100 = 86,23% accurate.

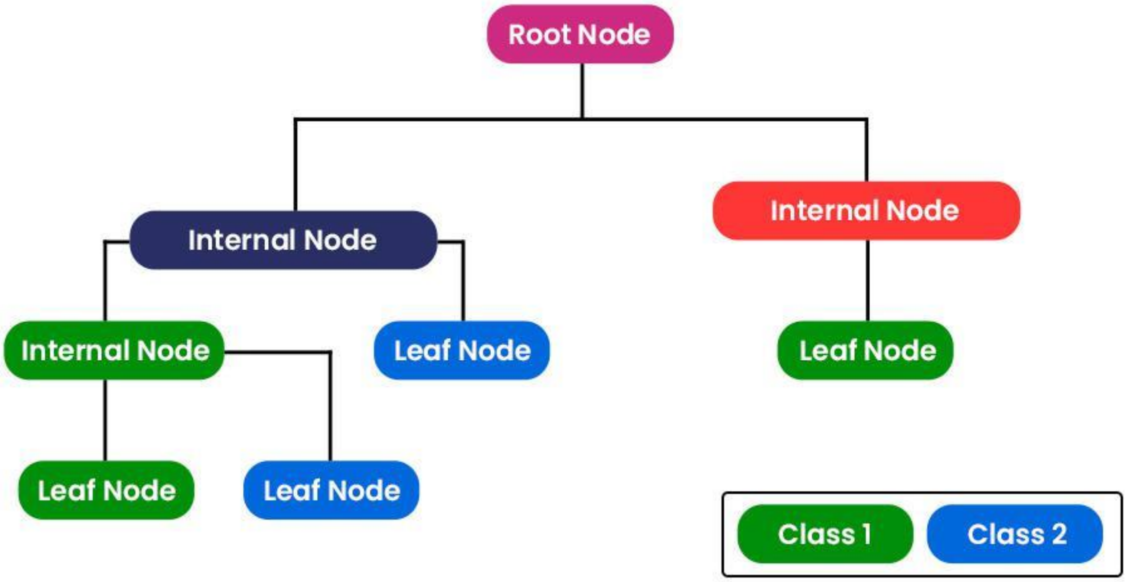
As a bonus for our study, I wanted to predict through our model whether a team with the following statistics will advance to the next stage:

Win = 2; Draws = 0; Losses = 1; Goals Scored =9; Goals Conceded =5; Points =6

Our model gave us a probability of 97,42% that the team will advance to the next round.

II/ Decision Tree:

The decision tree is the second method that we used for our study. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the input data features. The decision tree is a popular choice for many machine learning applications due to their simplicity and interpretability.



The algorithm starts at the root node and compares the value of the input feature to the value stored in the node. If the input feature is smaller or larger than the value stored in the node, it will follow the left or right branch respectively. The process repeats until it reaches a leaf node, which represents the output.

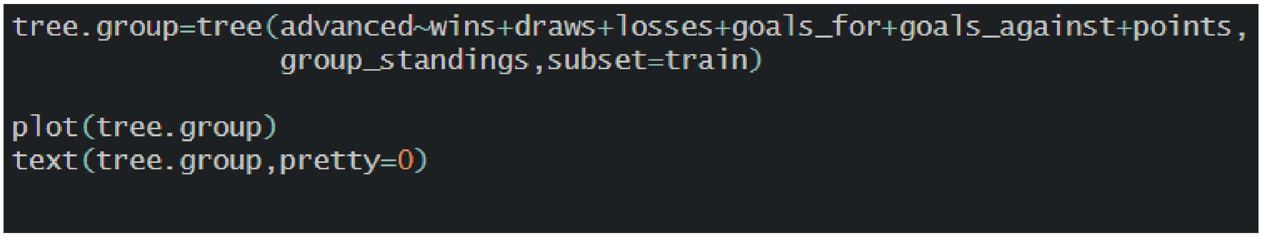
The Packages/Libraries we installed in R are:

- tree: is a package that provides functions for creating and analyzing decision trees.
- ggplot2: is a package for creating data visualizations.

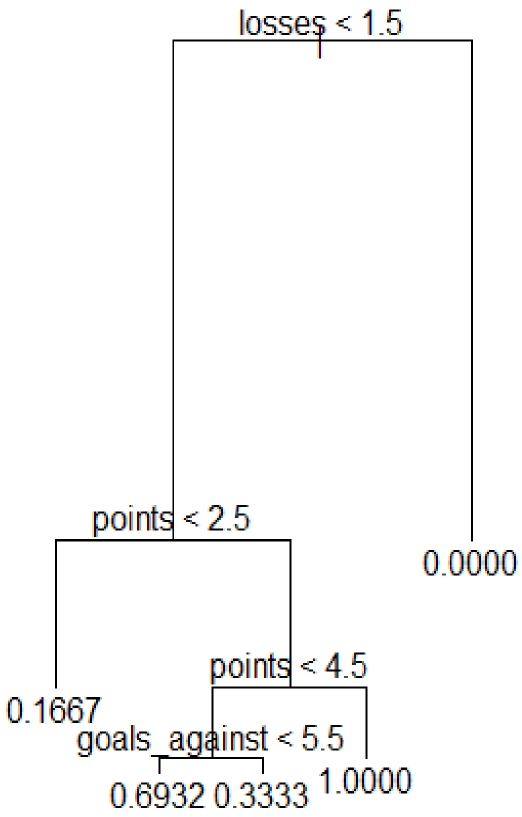
The concept and the goal of building this model, is to figure out the same outcome as the first model we built through “Neural Network” so that we can fairly compare how accurate the two different methods were and which one was the better one for our database.

To put it in words again : we want to predict whether a team will advance into the next round based on its different statistics (wins, draws, losses, goals scored, goals conceded and points collected).

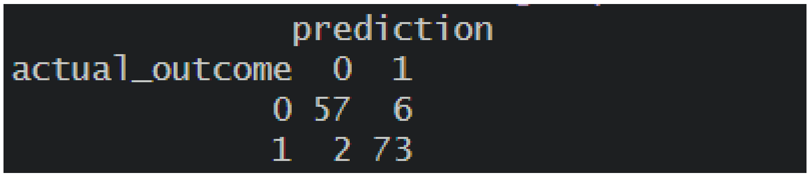
We started that through extracting out training data (70% of the lines of the original “group_standings” data frame) , then building through it our model and finally visualizing it :



```
tree.group=tree(advanced~wins+draws+losses+goals_for+goals_against+points,
                group_standings,subset=train)

plot(tree.group)
text(tree.group,pretty=0)
```



So now that our model was built, we could use it to predict our test data which represented the rest of the original data frame. Then compared it with the actual outcome just like we did in the first method:



```
              prediction
actual_outcome  0  1
             0 57  6
             1  2 73
```

Out of the 63 “0” values in the test data frame, we predicted 57 correctly. And out of the 75 “1” values we predicted 73 correctly.

So to be precise, our model was : [(57+73)/(57+73+2+6)]*100 = 94,20% accurate.

III/ Comparing both Methods:

An accuracy of 94,20% for a decision tree and 86,23% for a neural network indicates that the decision tree performed better than the neural network on the given data set. Accuracy is a commonly used metric to evaluate the performance of a classifier, it is the proportion of correct predictions made by the model. A higher accuracy means that the model is making more correct predictions.

5 Conclusion

In conclusion, the use of the different R libraries that are available (e.g.: dplyr) has shown to be a very effective method for analyzing and visualizing the different outcomes that had happened in the history of the World Cup.

Additionally, the use of historical data and statistical analysis allows for the identification of patterns and trends that can inform future predictions using different machine learning methods, the two ones that we have covered were the Neural Network and the Decision Trees. These machine learning techniques have been able to accurately forecast results such as whether the Team was going to advance from the group stage into the next one. They are powerful machine learning techniques that can be used to model complex patterns and relationships in data.

However, it is important to note that these methods have some limitations such as the need for large amounts of data, the risk of overfitting, and the computational cost. Also, they are not correct 100 percent of the time because there may be other factors that can impact the desired outcomes. That is why further research and analysis is necessary to continue to improve the accuracy of the predictions.

6 References

https://en.wikipedia.org/wiki/FIFA_World_Cup

https://www.sas.com/en_us/insights/analytics/machine-learning.html

https://www.alteryx.com/glossary/supervised-vs-unsupervised-learning#:~:text=Supervised%20and%20unsupervised%20learning%20have,tagged%20with%20the%20right%20answer.&text=A%20classification%20problem%20uses%20algorithms%20to%20classify%20data%20into%20particular%20segments.

https://www.ibm.com/topics/neural-networks#:~:text=Neural%20networks%2C%20also%20known%20as,neurons%20signal%20to%20one%20another.

https://scikit-learn.org/stable/modules/tree.html