World Cup Tactics Analysis

Senior Project

Name: ***Jaime Gutiérrez de Calderón Martínez***

Major: ***Mathematics and Engineering Science***

Millikin ID: ***905640***

Tutor: ***Dr. Cassidy Krause, Professor of Mathematics and Computer Science***

INDEX

1. INTRODUCTION

The Soccer World Cup is one of the most important sporting events in the world, bringing together the best national soccer teams to compete for the title of world champion. The first edition of the World Cup was held in 1930 in Uruguay, and since then it has become one of the most followed sporting events worldwide.

Over the years, soccer has evolved and undergone significant changes in tactics and the way the game is played. In the early editions of the World Cup, teams tended to play with a more defensive approach, using a formation that included five defenders and a striker. However, over time, teams began to adopt a more attacking approach, using more balanced formations with a greater presence in the center of the field.

In recent decades, soccer tactics have evolved further, with an increasing focus on possession and building play from the back. Teams have begun to use more flexible formations, with a greater emphasis on mobility and the ability to press and win the ball back in midfield.

World Cup soccer has witnessed a steady evolution in tactics and gameplay over the years. From the defensive formations of the 1930s to the more attacking and flexible tactics of today, soccer has proven to be a constantly evolving sport, driven by the creativity and innovation of coaches and players.

Nowadays, data analysis and the use of big data are fundamental tools in professional soccer. Soccer clubs use huge banks of information for decision-making in various areas, such as training planning, nutrition, the transfer market and game analysis, among others. The main objective is to maximize team performance at all levels, minimizing errors and saving time and costs.[1]

Data analytics tools provide valuable information that enables clubs to make decisions based on facts and statistics, rather than relying solely on intuition or subjective judgment. This makes it possible to identify patterns, strengths and weaknesses of both one's own team and opponents, which can be key to devising effective game

[1] Días, Ricky García. "El Análisis De Datos Revoluciona El Fútbol: ¿Qué Es Y Cómo Se Aplica En España?" *Cadena SER*, 3 Feb. 2022, cadenaser.com/ser/2021/11/21/deportes/1637491611_014299.html.

strategies. In addition, data analysis is also used to evaluate the performance of individual players, which can be of great help in making decisions related to the transfer market.

In summary, the importance of big data and data analytics in today's soccer lies in its ability to provide accurate and objective information for decision-making in multiple areas of the sport. This allows clubs to be more efficient, minimize errors and maximize their performance at all levels.

2. PROJECT

Soccer is a complex and dynamic sport that depends on multiple factors to achieve success. In this final project, we aim to analyze data from World Cup soccer matches to identify key patterns and factors that influence team performance and success. This work will analyze the impact of various factors on soccer World Cup matches, such as team formation, average age of players, and ball possession. Using predictive models, we will examine which of these factors have the greatest influence on the outcome of the match.

Furthermore, through the study of the data and analysis of the results, we hope to better understand the dynamics of World Cup matches and provide insights that can be used by coaches and players to improve their performance. Ultimately, this work aims to provide a deeper and more rigorous view of soccer as a sport and to contribute to the development of new strategies and approaches for success in the World Cup.

3. Literature Review

Soccer is one of the most popular sports in the world, with the FIFA World Cup being the biggest and most prestigious tournament in the sport. The tournament is held every four years, and it attracts millions of viewers from around the world. As a result, the tournament has been the subject of numerous research studies in recent years. In this section, we will provide a comprehensive review of the existing literature on soccer World Cup matches and the factors that affect their outcome.

One of the most extensively studied factors in soccer World Cup matches is team formation. Team formation refers to the arrangement of players on the field, including the number of players in each position, the distance between players, and the overall tactical approach of the team. The study Forcher et al. ("The Influence of Tactical Formation on Physical and Technical Match Performance in Male Soccer: A Systematic Review")[2] found that teams that used a 4-4-2 formation had a higher chance of winning than teams that used other formations. Similarly, another study found that teams that used a more defensive formation, such as 5-4-1, had a higher chance of winning against stronger opponents.

Another factor that has been extensively studied in soccer World Cup matches is the average player age. Younger teams are generally thought to have more energy and enthusiasm, while older teams are considered to be more experienced and tactically astute. In this research *(¿Cómo Afecta La Edad a Nuestro Rendimiento Deportivo?* – Altea Sport Clinic) found that younger teams had a higher chance of winning, especially in the earlier stages of the tournament. However, as the tournament progressed, the experience of older teams became more valuable.

Possession of the ball is another factor that has been studied extensively in soccer World Cup matches. Possession of the ball refers to the amount of time that a team has the ball during a game. It is known that teams that had more possession of the ball had a higher chance of winning, especially in the later stages of the tournament[3]. Similarly, a study by Redwood-Brown et al. ("Effects of Playing Position, Pitch Location, Opposition Ability and Team Ability on the Technical Performance of Elite Soccer Players in Different Score Line States")[4] found that teams that had more successful passes had a higher chance of winning.

---

[2] "The Influence of Tactical Formation on Physical and Technical Match Performance in Male Soccer: A Systematic Review." *International Journal of Sports Science & Coaching*, SAGE Publishing, May 2022, p. 174795412211013. https://doi.org/10.1177/17479541221101363.

[3] Glasser, Hana. "The Problem With Possession: The Inside Story of Soccer's Most Controversial Stat." *Slate Magazine*, 27 June 2014, slate.com/culture/2014/06/soccer-possession-the-inside-story-of-the-games-most-controversial-stat.html.

[4] "Effects of Playing Position, Pitch Location, Opposition Ability and Team Ability on the Technical Performance of Elite Soccer Players in Different Score Line States." *PLOS ONE*, vol. 14, no. 2, Public Library of Science, Feb. 2019, p. e0211707. https://doi.org/10.1371/journal.pone.0211707.

While the above factors have been studied extensively, there are still several research gaps that need to be addressed. For example, the impact of player injuries on the outcome of soccer World Cup matches has not been extensively studied. Additionally, the impact of psychological factors, such as team motivation and pressure, on the outcome of the matches needs to be explored further.

After analyzing the importance of training factors, ball possession and the average age of the players, it was decided to compare the playing styles of two of the most successful coaches in the sport, Pep Guardiola and José Mourinho, to find out which of them is more successful in this competition.

Jose Mourinho and Pep Guardiola are two soccer coaches with different tactical approaches. In terms of the formation they use, Mourinho tends to prefer a more defensive and organized approach, with a typical 4-2-3-1, 4-3-1-2 or 5-3-2 formation, while Guardiola tends to use a more attacking formation, such as a 4-3-3, 3-4-1-2 or a 3-4-3, emphasizing ball possession and the passing game.

In terms of the focus on ball possession, Guardiola is known for his focus on "tiki-taka," a style of play that emphasizes ball control and quick ball movement with a lot of short, quick passing, while Mourinho tends to favor a more counter-attacking strategy, focusing on solid defense and quick counter-attacking. Guardiola's teams tend to get more than 57% possession while Mourinho's teams tend to hand possession to the opposition in a way that usually gets less than 50%.

Finally, in terms of player age, Guardiola has often been criticized for relying on young players and using a long-term development philosophy, while Mourinho has been criticized for being more likely to rely on experienced and veteran players, looking for short-term results rather than a long-term project.

After analyzing this data, we proceeded to look for the best predicting model to be able to know, from the variables of a match, if a team won or lost that match. “Predictive modeling uses known results to create, process, and validate a model that can be used to forecast future outcomes”[5].

[5] Frankenfield, Jake. “Predictive Modeling Definition.” *Investopedia*, Feb. 2023, www.investopedia.com/terms/p/predictive-modeling.asp.

In order to predict the outcome of soccer World Cup matches, we used several predictive models, including linear regression, decision trees, Bernoulli Navie Bayes, K-Nearest Neighbors, XGBoost, LightGBM, Gradient Boosting and MLP and neural networks. We used the data collected from the previous World Cup matches to train the models and then used them to predict the outcome of matches in the current World Cup. We compared the performance of each model and determined which model was the most accurate in predicting the outcome of the game.

4. Collecting Data

To start my work I had to learn how to get in an efficient way the data of the World Cup matches. At first I thought of doing it by hand, that is simply copying all the data I found in an excel table, however I soon realized that this method would take me much more time than expected and could lead me to get some error.

My project tutor suggested me the web scraping technique to speed up the process, which we had studied in one of the classes she was teaching, "Web scraping, also known as web extraction or harvesting, is a technique to extract data from the World Wide Web (WWW) and save it to a file system or database for later retrieval or analysis."[6]

Thanks to this method the process was more agile, although in some web pages it was complicated web scraping since the way the html was placed was tedious. The websites from which I obtained information about the matches of the different World Cups were Eurosport[7] for the 2010 and 2018 World Cups, Skysport[8] for the 2014 World Cup and BeSoccer[9] for the last World Cup that was held during the months of November and December last year in Qatar.

It was complicated to get useful data in the World Cups that were prior to 2010 so I decided to only use the data obtained from this World Cup, and even then I realized

---

[6] Zhao, Bo. "Web scraping." *Encyclopedia of big data* (2017): 1-3.

[7] "World Cup 2010 Table and Standings - Football Rankings | Eurosport." *Eurosport*, www.eurosport.com/football/world-cup/2010/standings.shtml.

[8] "Football (Sky Sports)." *SkySports*, www.skysports.com/football.

[9] Besoccer. "BeSoccer Livescore: All Today's Live Soccer Scores." *Besoccer*, www.besoccer.com.

that the last World Cups had much more information than the previous ones. However, in all the years there were data such as goals per team, ball possession, formation used, shots taken by each team, offsides, cards received, fouls, etc.

When I gathered all the data in csv format I had to add two more variables, since my goal was to know if a team had won or not and there was nothing to indicate this data, for this I used a function which compared the goals of the two teams that had faced each other and the one that had scored more goals would have the data 'Yes' indicating that it had won the match and the one that had scored fewer goals would have a 'Loose' indicating that it had lost the match, in case of being equal the data 'Tie' was taken. The goals conceded in a match did not appear either, but for this we only had to copy the goals scored by the opposing team and in this way we would have this data to be able to perform our analysis.

5. Descriptive analysis

To do this process Jupyter with the Python programming language was used, since this way it would be easier to obtain all the data I needed to be able to properly perform my research.

The Python library called "pandas" was used for the analysis and data manipulation since working with it made the work much easier as well as integrating well with other helpful libraries such as Numpy, Matplotlib and SciPy.

5.1 Correlations

As I began to observe the data I tried to look for correlations between the different variables and my objective variable, which was the last one I added to know whether or not a team had won the game. To do this I used Pearson's correlation, which follows the formula: $P = \frac{cov(XY)}{\sigma_x \sigma_y}$ [10], where cov(XY) is the covariance between X and Y and σ_x,σ_y are the standard deviation of each.

[10] Hahs-Vaughn, Debbie L. "Foundational Methods: Descriptive Statistics: Bivariate and Multivariate Data (Correlations, Associations)." *Elsevier eBooks*, Elsevier BV, 2023, pp. 734–50. https://doi.org/10.1016/b978-0-12-818630-5.10084-3.

The results obtained gave the variable goals scored a correlation of 0.51 being the variable with the highest correlation, another interesting data was the correlation with goals conceded which was -0.22. The variables that I also looked at since they were the object of our study were ball possession which only had 0.02 and the average age of the players which was 0.086.

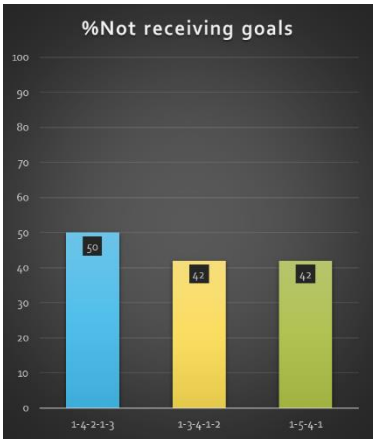
I decided to look for data that would give you a high chance of winning a match and I found that there was one piece of data that was clearly relevant, if you did not concede any goals you could not lose the match and the winning percentage was 72.43% and the rest were draws, except for one match that was lost on penalties. Also if you were able to score two or more goals you had a winning percentage close to 80%, which I also used as an object of study. And finally if you managed to shoot more than 12.5 times at the opponent's goal the winning percentage was 65%.

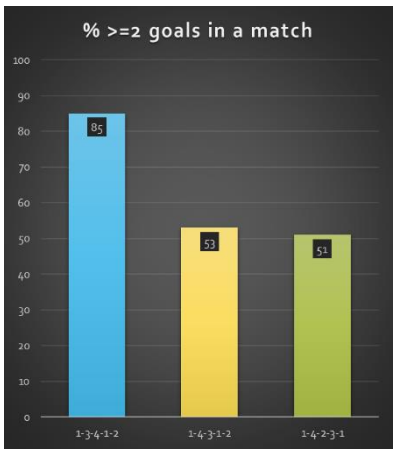
5.2 Differences Tactics Mourinho vs Guardiola

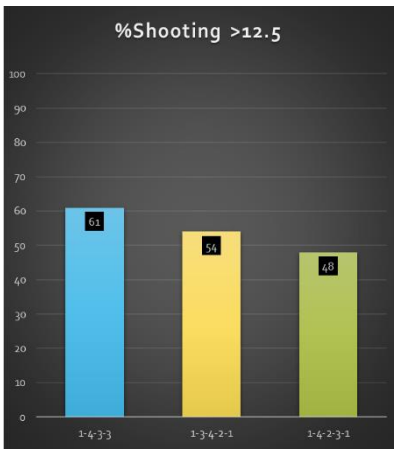
In this section I looked for which of the tactics used by each of the coaches gave you an advantage in each of the sections mentioned above such as not conceding goals, scoring two or more goals and getting more than 12.5 shots.

5.2.1 Formation

One of the formations used by Mourinho as 4-2-3-1 or its variation 4-2-1-3 was quite good as it had a good balance between the analysis objectives we had, standing out in the ability of not conceding goals with 42% being the second in this section, however the 3-4-1-2 formation that was used on several occasions clearly stood out in the section of scoring two or more goals with 85%. The graph below show the three most relevant formations in each section.

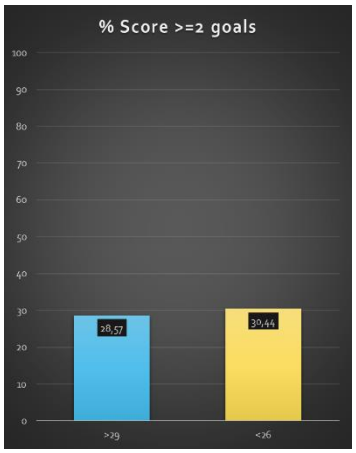


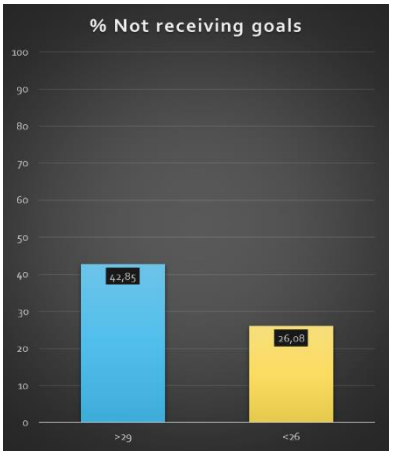


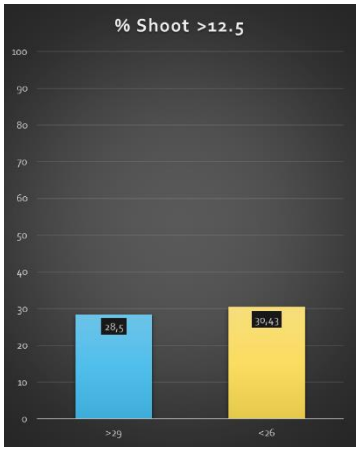


5.2.2 Average age of players

Regarding the age of the players, young players were taken as those under 26 years old and older players as those over 29 years old. The results were similar in all areas except in the section of not conceding goals where clearly the teams with more experienced players seemed to have a higher probability of not conceding goals.

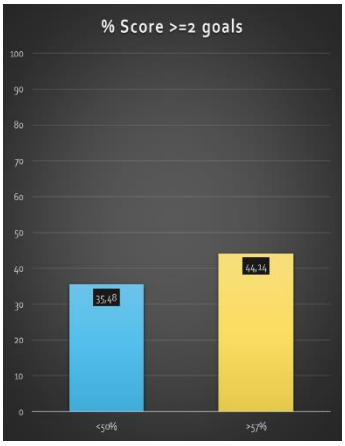


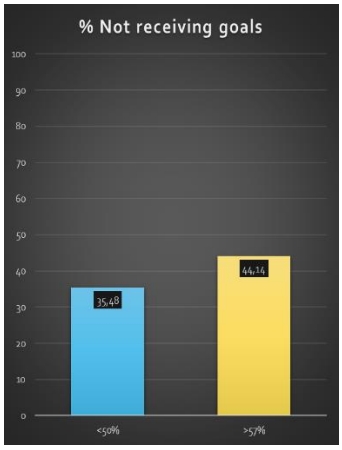


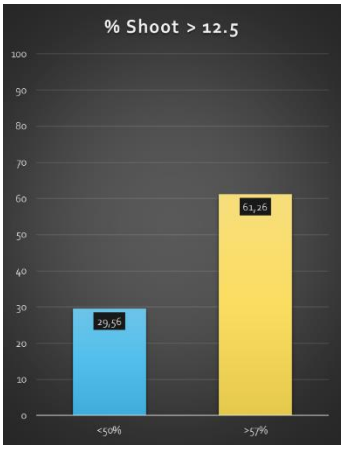


5.2.3 Ball possession

In terms of ball possession there was a clear difference in all sections in favor of having a possession greater than 57%, this difference was especially accentuated in the shooting section where it was observed that if possession was greater than 57% the probability of obtaining more than 12.5 was 65% while if possession was less than 50% this probability dropped to 28%.







6.Predicting models

After having analyzed how ball possession, player age and formation affected the soccer match, we started the process of searching for the best possible predictive

model for our dataset. Several models were tried and tested but most of them did not give results that were very good for our study.

The way to compare if the models were good was based on calculating the ROC curve, ("The ROC curve is obtained by plotting, for each possible choice of cutoff value, the sensitivity in ordinates and 1-specificity in abscissae"[11]) and the AUC (area under the curve, which is used to determine how good the model is at discriminating whether a match has been won or not, if the AUC is 0.5, our model is random, but the closer to the value 1 our model is better at guessing the possible winners of a match).

For the elaboration of our model a portion of 0.6 of our dataset was used as training since this was how the predictive models achieved a better AUC score. The games selected were chosen randomly so that they were not influenced in any way by the difference between how the game was played in the past and how it is played nowadays.

```
# Reparto train y test
# ==============================================================================
X = df_final.drop('Win',axis=1)
y=df_final['Win']
X_train, X_test, y_train, y_test = train_test_split(
                                        X,
                                        y,
                                        train_size   = 0.6,
                                        random_state = 121,
                                        shuffle      = True
                                    )
```

6.1 Logistic Regression

The best performing model was Logistic Regression. “This type of statistical model (also known as *logit model*) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit

[11] Ullibarri Galparsoro, López, and Pita Fernández. "Curvas roc." *Atención Primaria en la Red* 5.4 (1998): 229-35

transformation is applied on the odds—that is, the probability of success divided by the probability of failure."[12]

The formula for this model is:

- $log\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$
- $p = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}$

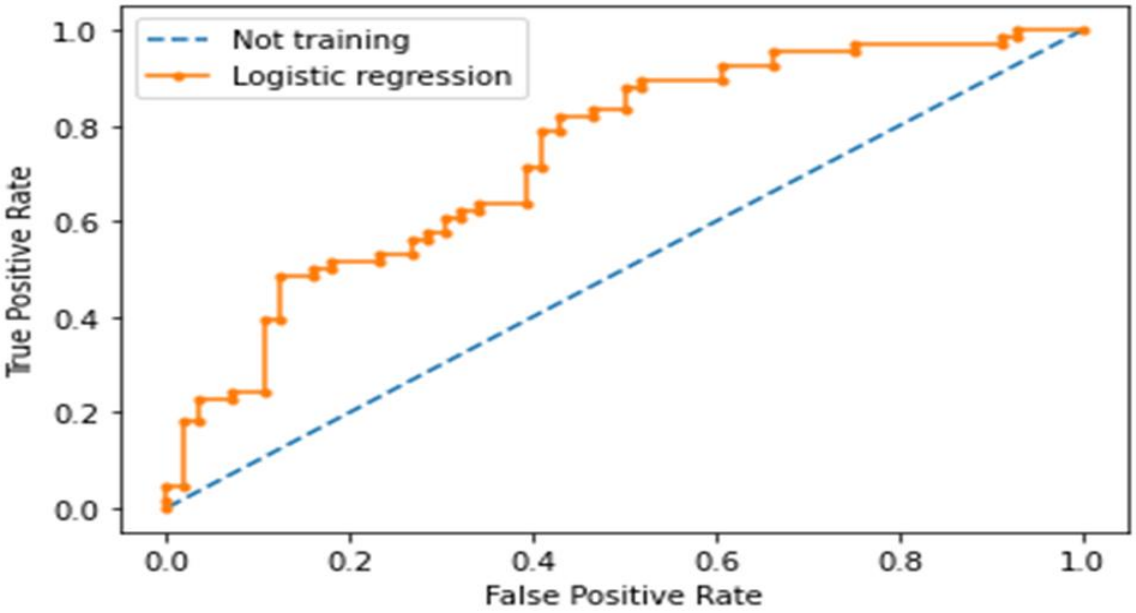
β_o: is the constant of the logistic regression model..

p: is the probability that the dependent variable is worth 1.

β_i: is the regression coefficient of variable i.

x_i: is the value of variable i.

The AUC obtained with this model was 0.70.



We tried to look if changing the hyperparameters of the model we were able to improve my model. To do this we use the function grid search, that is a tool for

[12] *What Is Logistic Regression? | IBM*. www.ibm.com/topics/logistic-regression#:~:text=This%20type%20of%20statistical%20model%20%28also%20known%20as,based%20on%20a%20given%20dataset%20of%20independent%20variables.

tweaking hyperparameters in machine learning. Grid Search offers a perfect collection of hyperparameters that enable us to get better results by doing several calculations on the available hyperparameters for each machine learning method.

In the image below you can observe the hyperparameters that we look to find which ones were the best parameters to increase the performance of our model.

```
# Define hyperparameters to tune using grid search
hyperparameters = {'penalty': ['l1', 'l2'],
                   'C': [0.01, 0.1, 1, 10],
                   'class_weight': [None, 'balanced'],
                   'solver': ['lbfgs', 'liblinear', 'newton-cg']}

# Perform grid search
grid_search = GridSearchCV(logreg, hyperparameters, cv=5, scoring='accuracy')
grid_search.fit(X_train, y_train)
```

These ones were the best parameters that our grid search found: {'C': 10, 'class_weight': 'balanced', 'penalty': 'l2', 'solver': 'lbfgs'}, but the AUC just increase 0.01 so there was not a big difference with the one that we used before.

6.2 Decision Trees (ID3 Algorithm)

One of the most famous predicting model is Decision Tree that is Tree-based classification and regression models are created using supervised machine learning algorithms.

In a decision tree, each branch:

- Node is a property or feature.
- A branch is a choice or regulation.
- Leaf: a result (continuous or categorical)

ID3 Algorithm: It is a classification algorithm that follows a greedy approach by selecting a best attribute that yields maximum Information Gain(IG) or minimum Entropy(H).

The equations to calculate each of the information needed for the ID3 algorithm are:

- $Entropy(S) = -\frac{p}{p+n} log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n} log_2\left(\frac{n}{p+n}\right)$
- $I(Attribute) = \sum \frac{p_i+n_i}{p+n} * Entropy(Attribute)$

- *Gain=Entropy(S) – I(Attribute)*

Where p are the positives (wins) and n are the negative (looses).

The steps in ID3 algorithm are as follows:

- Calculate entropy for dataset.
- For each attribute/feature.
- a) 2.1. Calculate entropy for all its categorical values.
- b) 2.2. Calculate information gain for the feature.
- Find the feature with maximum information gain.
- Repeat it until we get the desired tree.[13]

The AUC get with this model was not really good and it was not higher than 0.55.

6.3 XGBoost

Looking for ways to improve the Decision Trees model we found that there is a new model that is called XGBoost. This models trained on the training data set, by building a sequence of decision trees iteratively. In each iteration, the model tries to minimize the loss function by calculating the gradients and updating the parameters.

This model improved the decision tree model in a considerable way, the AUC obtained was 0.68. After this, I decided to use the grid search method to look for the parameters that best fit in order to get the maximum performance.

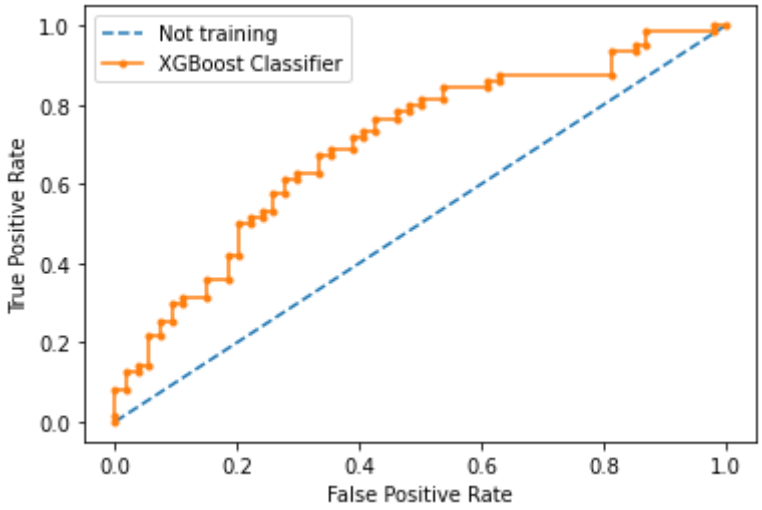
These were the parameters we compared:

[13] Mantri, Nidhi. "Using ID3 Algorithm to Build a Decision Tree to Predict the Weather." *OpenGenus IQ: Computing Expertise & Legacy*, June 2021, iq.opengenus.org/id3-algorithm.

```
# Define hyperparameters to tune using grid search
hyperparameters = {'max_depth': [2, 3, 4],
                   'learning_rate': [0.01, 0.1, 1],
                   'n_estimators': [100, 500, 1000],
                   'gamma': [0, 0.1, 1],
                   'subsample': [0.5, 0.7, 1.0],
                   'colsample_bytree': [0.5, 0.7, 1.0],
                   'reg_alpha': [0, 0.1, 1],
                   'reg_lambda': [0, 0.1, 1]}

# Perform grid search
grid_search = GridSearchCV(xgb_clf, hyperparameters, cv=5, scoring='accuracy')
grid_search.fit(X_train, y_train)
```

The parameters that we found that were the best to improve our model. It improves a few the model and we get a 0.71.



7. Conclusion

After analyzing the three variables we were targeting (formation, average age of players and ball possession) considering the soccer ideas of coaches Jose Mourinho and Pep Guardiola we realized that the decisions that gave us the best chance of winning in a world cup match were:

- **Formation**: 1-3-4-1-2, which was one of the formations Guardiola has used in his career.

- **Average age of players**: >29 years old, which as we said in our paper was the age that the Portuguese coach, Jose Mourinho, was betting on.

- **Ball possession**: >57%, one of the hallmarks of Guardiola's identity, who always bets on his team having more possession than the opponent in a clear way.

It should be noted that our analysis was made simply with data from the last four editions of the World Cup. The difference in stats and variables between the last competitions and the one held in 2010 was considerable.

We cannot extrapolate the results obtained in our project with other soccer competitions such as the Premier League or the Champions League, since the competitions are different and the teams that play in them are formed by different players of different nationalities, unlike in the World Cup where the players of the same team are all from the same nation. Besides as we have mentioned before the Premier League are 38 matches where the regularity of the teams is rewarded and however in the world cup are at most 7 matches where the pressure of the matches is different and losing affects in a more negative way than in the league.

In our goal to find the best predictive models we obtained that logistic regression and XGBoost were the best models for this purpose, however the AUC obtained was not higher than 0.73, which tells us that there is room for improvement for our model.

To improve my predictive models I would like to explore more about neural models and see if the results obtained with them improve in a significant way the ones we have obtained.

In conclusion, our project has allowed us to identify the variables that could influence the outcome of a world cup match and the decisions that could give more probability of success according to the soccer ideas of coaches Mourinho and Guardiola. In addition, we have identified that there is room for improvement in our predictive models and that we could explore neural models in the future. However, it is important to keep in mind the limitations of our analysis and that the results cannot be extrapolated to other soccer competitions.

Finally I leave here the link to my GitHub page (Jguti1201 (Jaime I) (github.com)) where you can find the programs I have used for this project, any student or person who needs to use any of the programs has my permission to use them.

8. REFERENCES

1. Días, Ricky García. “El Análisis De Datos Revoluciona El Fútbol: ¿Qué Es Y Cómo Se Aplica En España?” *Cadena SER*, 3 Feb. 2022, cadenaser.com/ser/2021/11/21/deportes/1637491611_014299.html.

2. “The Influence of Tactical Formation on Physical and Technical Match Performance in Male Soccer: A Systematic Review.” *International Journal of Sports Science & Coaching*, SAGE Publishing, May 2022, p. 174795412211013. https://doi.org/10.1177/17479541221101363.

3. Glasser, Hana. “The Problem With Possession: The Inside Story of Soccer’s Most Controversial Stat.” *Slate Magazine*, 27 June 2014, slate.com/culture/2014/06/soccer-possession-the-inside-story-of-the-games-most-controversial-stat.html.

4. “Effects of Playing Position, Pitch Location, Opposition Ability and Team Ability on the Technical Performance of Elite Soccer Players in Different Score Line States.” *PLOS ONE*, vol. 14, no. 2, Public Library of Science, Feb. 2019, p. e0211707. https://doi.org/10.1371/journal.pone.0211707.

5. Frankenfield, Jake. “Predictive Modeling Definition.” *Investopedia*, Feb. 2023, www.investopedia.com/terms/p/predictive-modeling.asp.

6. Zhao, Bo. "Web scraping." *Encyclopedia of big data* (2017): 1-3.

7. “World Cup 2010 Table and Standings - Football Rankings | Eurosport.” *Eurosport*, www.eurosport.com/football/world-cup/2010/standings.shtml.

8. “Football (Sky Sports).” *SkySports*, www.skysports.com/football.

9. Besoccer. “BeSoccer Livescore: All Today’s Live Soccer Scores.” *Besoccer*, www.besoccer.com.

10. Hahs-Vaughn, Debbie L. “Foundational Methods: Descriptive Statistics: Bivariate and Multivariate Data (Correlations, Associations).” *Elsevier eBooks*, Elsevier BV, 2023, pp. 734–50. https://doi.org/10.1016/b978-0-12-818630-5.10084-3.

11. Ullibarri Galparsoro, López, and Píta Fernández. "Curvas roc." *Atención Primaria en la Red* 5.4 (1998): 229-35

12. *What Is Logistic Regression? | IBM*. www.ibm.com/topics/logistic-regression#:~:text=This%20type%20of%20statistical%20model%20%28also%20known%20as,based%20on%20a%20given%20dataset%20of%20independent%20variables.

13. Mantri, Nidhi. “Using ID3 Algorithm to Build a Decision Tree to Predict the Weather.” *OpenGenus IQ: Computing Expertise & Legacy*, June 2021, iq.opengenus.org/id3-algorithm.