VIETNAM NATIONAL UNIVERSITY OF HOCHIMINH CITY

THE INTERNATIONAL UNIVERSITY

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING



**FIFA 22 WORLD CUP PREDICTIONS**

By

NGUYEN VIET ANH

PRE-THESIS

Ho Chi Minh City, Vietnam

2022 – 2023

**FIFA 22 WORLD CUP PREDICTIONS**

APPROVED BY: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_,

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

PRE-THESIS COMMITTEE

**TABLE OF CONTENTS**

**LIST OF FIGURES**. i

**LIST OF TABLES**. ii

**ABBREVIATIONS**. iii

**TOPIC MOTIVATION**. iv

**ABSTRACT**. v

INTRODUCTION

1.1. Scope and Objectives. 12

1.2. Structure of the Report 12

LITERATURE REVIEW / RELATED WORK.. 12

2.1. Machine Learning: Machine. Learns 13

2.1.1. Anatomy. 13

2.1.2. Training & Testing Data. 14

2.1.3. Supervised Learning. 15

2.2. WC Dataset 16

2.3. Random Forests algorithm.. 17

2.3.1. Overview.. 17

2.3.2. Specialties. 17

2.4. Decision Tree algorithm.. 19

2.4.1. Overview.. 19

2.4.2. Specialties.. 19

2.5. Logistic Regression.. 21

2.5.1. Overview..21

2.5.2. Specialties..21

2.6. Applying the Logistic Regression for Model 22

2.6.1. Strengths 21

2.6.2. Build Logistic Regression model.. 22

IMPLEMENTING.. 23

3.1. WC Dataset Structure. 23

3.2. Exploratory Data Analysis. 25

3.2.1 Find the correlation, and unnecessary columns, and change the column names. 31

3.2.2 Train Test Split. 32

3.2.3 Scaling. 32

3.2.4 Confusion Matrix. 32

3.2.5 Classifiers. 33

3.3. Simulation. 39

3.4. Comparison. 43

CONCLUSION AND FUTURE WORK.. 44

4.1. Conclusion. 45

4.2. Future work. 45

REFERENCES. 45

**LIST OF FIGURES**

*Figure 1 – Anatomy of machine learning. 12*

*Figure 2 – WC Dataset. 15*

*Figure 3 – Example of growing random trees to produce a prediction. 17*

*Figure 4 – Diagram of decision tree.. 18*

*Figure 5 – Occam’s Razor 19*

*Figure 6 – Logistic Regression 20*

*Figure 7 – Five last rows of the dataset. 23*

*Figure 8 – Datatypes of the dataset. 24*

*Figure 9 – Number of rows of data. 25*

*Figure 10 – Dataset before encoding. 25*

*Figure 11 – Dataset after encoding 26*

*Figure 12 – 32 Teams participating in WC 2022. 27*

*Figure 13 – Mean ranking of 32 countries in WC 2022 (following the FIFA ranking). 27*

*Figure 14 – Top 10 ranking calculated by average win percentage. 28*

*Figure 15 – Find the correlation of home team result.. 29*

*Figure 16 – Dropping unnecessary columns.. 29*

*Figure 17 – Change the column names to the short name of the variable.. 30*

*Figure 18 – Train Test Split.. 30*

*Figure 19 – Scaling.. 30*

*Figure 20 – Confusion Matrix.. 31*

*Figure 21 – Random Forest Classifiers.. 32*

*Figure 22 – Decision Tree Classifiers.. 33*

*Figure 23 – Ada Boost Classifier.. 34*

*Figure 24 – XGB Boost Classifier.. 35*

*Figure 25 – Random Forest classifier when removing draw.. 36*

*Figure 26 – Testing holdout and finalizing the model. 37*

*Figure 27 – Create a pipeline for the XGB model. 38*

*Figure 28 – Feature engineering. 38*

*Figure 29 – Training again for the model. 39*

*Figure 30 – Logistic Regression Model. 40*

*Figure 31 – Round 16 WC 2022. 40*

*Figure 32 – Quarter-Final WC 2022. 41*

*Figure 33 – Semi-Final WC 2022. 41*

*Figure 34 – Winner WC 2022. 42*

*Figure 35 – Comparison between Rand, DT, and Logreg. 42*

**LIST OF TABLES**

Table 1 – Comparison between techniques based on the practicality of input/output data.

**ABBREVIATIONS**

(Google)Colab: Colaboratory

EDA: Exploratory Data Analysis

Rand: Random Forest

DT: Decision Tree

Logreg: Logistic Regression

NA: Not Available

NaN: Not a number

Corr: Correlation

WC: World Cup

Euro: Europe

ML: Machine Learning

**TOPIC MOTIVATION**

Football is known as the king of sports! First and foremost, football is easy. This is a

fantastic illustration of the saying "simplicity is beautiful" in action. One goal for all 22 players: to put a ball into the net without touching it by hand. Although there are countless approaches a team might use to accomplish that goal, the fundamentals are still the same.

Second, football offers enormous scope for personal expression and originality. Each particular move, whether it be a lovely pass, a deft dribble, or a potent hit, may be distinguished by a specific element of flair. One merely needs to contrast Messi's and Neymar's dribbling techniques. While Messi and Neymar are both fantastic footballers, Neymar favors extremely artistic touches and plays, while Messi excels at rapid, short dribbles. There are various ways that each player may use their unique abilities and influence the game. Every second on the field may be customized to your liking, whether it is through tackling, heading, or simple movement.

Thirdly, football is a mental as well as a physical struggle. Compromises, playing methods, tactical formations—the list is endless. Managers are essential to a team's performance because of how they execute tactics, make replacements, and coach their players. A team can have some highly talented players, but if there is no overarching plan that unifies them and inspires them to achieve, your team's efficacy will be constrained.

Last but not least, football is adored worldwide. Its widespread use is evidence of people and our remarkable skills. The game is played by billions of people, and it has brought joy and hope to countless individuals all over the world. (Except, of course, if your team consistently loses).

**ABSTRACT**

During World Cup 2022, people predict the results of each match to get gifts from the company or betting website. So, I also would like to predict it by science (Machine Learning algorithms) to explore many aspects of the prediction step by step. From EDA to split, train, and test the data. The accuracy of the machine learning algorithm is a prerequisite factor to building models. The simulation is the best option to guide the viewers to observe the final results visually. This paper focuses on predicting the World Cup by Machine Learning(Rand, DT, Logreg).

[1] “With the existing data as a [machine-learning training mechanism](https://theconversation.com/what-is-machine-learning-76759), we employed [Random Forest algorithm](https://www.ibm.com/cloud/learn/random-forest) to predict results for every World Cup fixture.”

Keywords: Machine Learning, EDA, Rand, DT, Logreg, split, train, test, predict.

# **CHAPTER 1 - INTRODUCTION**

Did you know that 3.57 billion viewers watched World Cup 2018? Approximately 50% of the aged more than four population globally. The final match had 1.12 billion viewers around the world. These above reasons prove why World Cup is a fascinating sports event in the world. Although betting is forbidden in our country and these predictions will be had bad aspects, however, I just created scientific research and my target is to guide everyone to the results positively.

Machine Learning is the main method to perform the results through the chart, and accuracy. I need to explore data and clean it before using them for processing. The role of the data is extremely important because it will be affected the accuracy of the models. In addition, I have to find the correlation between the data and the model. From that, the report will be more tightly.

Last but not least is the surprises, every team can create a wonderful performance. Maybe the big teams will have big advantages to win the trophies. Nevertheless, the ball is in someone’s court.

## **1.1.** **Scope and Objectives**

The goal of this paper is to implement a suitable machine-learning algorithm to predict the WC 2022 and compare the attribute of each team. A successful prediction will help fans have the fairest view of soccer and cheer for all the teams regardless of the results.

## **1.2.** **Structure of The Report**

The main ideas of the report are split into a Literature Review, Implementing, and a Conclusion.

- Literature Review: Provides information about what machine learning is, what type of machine learning method will be used, and which dataset will prove beneficial to prediction.

- Implementing: A specific machine learning progress.

- Conclusion: To conclude that the chosen method can overwhelm others in this field.

# **CHAPTER 2 - LITERATURE REVIEW / RELATED WORK**

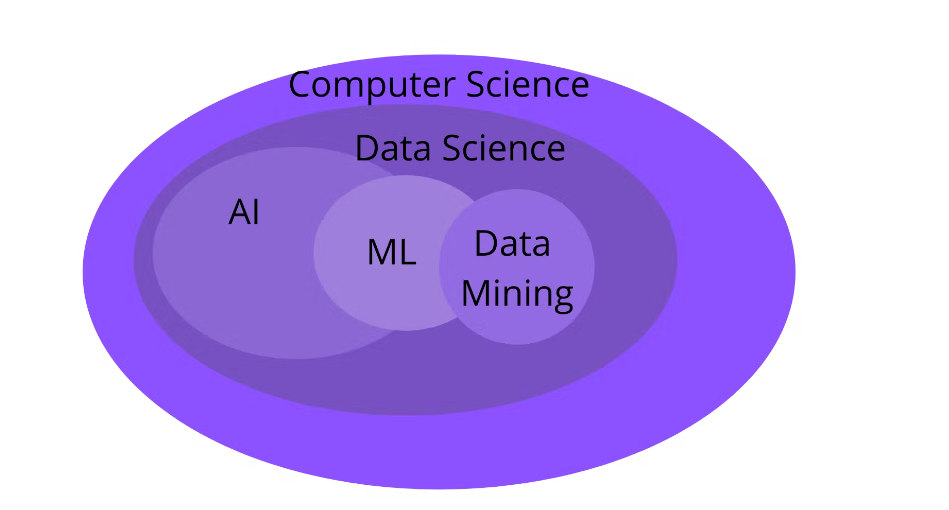
The implementation of the Random Forests, Decision Tree, and Logistic Regression algorithm on the prediction and mitigation method will be observed throughout this study.

## **2.1.** **Machine Learning: Machine. Learns**

Machine learning is a subfield of artificial intelligence (AI) and computer science that focuses on using data and algorithms to mimic how humans learn, gradually improving its accuracy. To be specific, machine learning is assisting in developing better forecasting algorithms for bookies, teams, and professional punters, as well as providing fresh insights into more accurate predictive models.

## **2.1.1. Anatomy**

Machine learning, data mining, artificial intelligence, and computer programming are all subsets of computer science, which encompasses all aspects of computer design and operation. Data science is the next broad field within the all-encompassing space of computer science. Data science, which is more limited than computer science, consists of methods and systems for extracting knowledge and insights from data using computers.



*Figure 1: Anatomy of Machine Learning[4]*

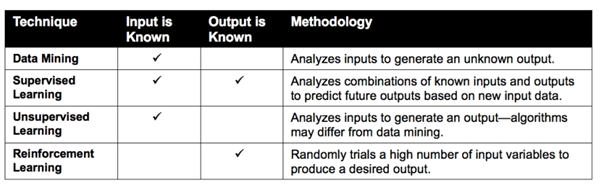
Machine learning is an excellent starting point for anyone interested in AI because it provides a narrower and more practical lens of study (in comparison to AI). Machine learning algorithms can also be used in other fields such as perception and natural language processing.

## **2.1.2.** **Training & Testing Data**

In machine learning, input data is typically divided into two categories: training data and test data. The initial reserve of data used to develop the model is the training data, which is the first split of data. In the WC data, the false positives such as win, lose, and draw are shown in the model. So, the model must then be modified, for example, the accuracy of the model is changed depending on the numbers of the match scores.

When you have developed a model based on patterns extracted from the training data and are satisfied with the accuracy of its predictions, you can test the model on the remaining data, known as the test data. If you are satisfied with the model's performance using the test data, the model is ready to filter normal and anomalies in a live setting and generate decisions on how to categorize those accuracies.

# **2.1.3.** **Supervised Learning**

*Table 1: Comparison between techniques based on the practicality of input/output data[5]*

Whereas data mining focuses on analyzing input variables to predict a new output, machine learning extends to analyzing both input and output variables. This includes supervised learning techniques that compare known combinations of input and output variables to discern patterns and make predictions and reinforcement learning which randomly trials a massive number of input variables to produce the desired output.

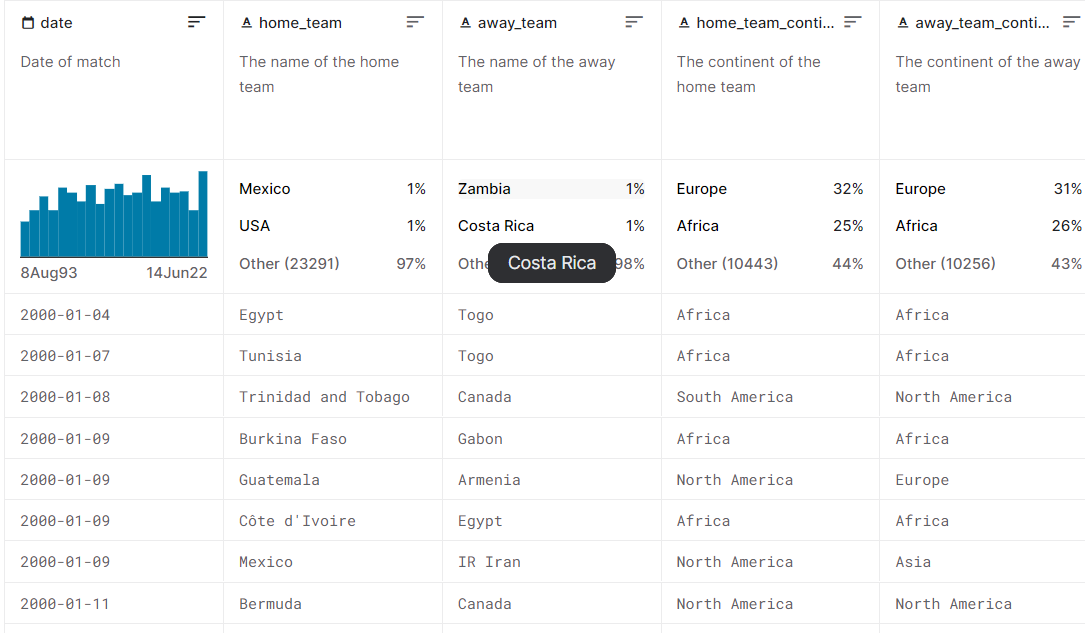
Supervised learning resembles our ability to extract patterns from known examples and apply that insight to engineer a repeatable outcome.

There are various algorithms in the name of Supervised Learning such as K-Nearest Neighbor (kNN), Support Vector Machine (SVM), Linear Regression, Logistic Regression, Decision Tree, Random Forests, Gradient Boosting, and so on[2]. In this paper, Random Forest, DT, and Logreg will be primarily covered as the main method for WC Prediction.

## **2.2.** **WC Dataset**

The dataset is generated by Data Scientist Brenda Netherlands. [6] “This dataset provides a complete overview of all international soccer matches played since the 90s. On top of that, the strength of each team is provided by incorporating actual FIFA rankings as well as player strengths based on the EA Sports FIFA video game”. However, I only use the data from 2000 because the previous year is so far and the prediction has been less accurate.

The dataset contains many columns but the mains are date, home team, away team, home team continent, away team continent, home team FIFA rank, away team FIFA rank, away team total points, and home team score. A few columns need values so I have to find them and the correlation from the data to build models more reasonably.

*Figure 2: WC dataset [6]*

In the dataset, we can sort ascending and descending belonging to the attribute of each column. Therefore, we can analyze any aspects that we want and help viewers easy to zoning the range we do.

## **2.3.** **Random Forests algorithm[5]**

# **2.3.1. Overview**

Random forests are a popular machine-learning model. They are a variant of the bagging algorithm. Any classifier or regressor can be used in bagging. The base classifier or regressor in random forests is always a decision tree.

The main idea of this algorithm is to take the average value of all multiple different trees prediction to predict the final result.

## **2.3.2.** **Specialities**

In the circumstance of bagging, different variations of the training data are run through each tree. This does not finish the problem of overfitting, the dominant patterns in the dataset will appear in a higher number of trees and emerge in the final class or prediction. It became a great solution to deal with outliers and lower the degree of variance typically found with a single decision tree.

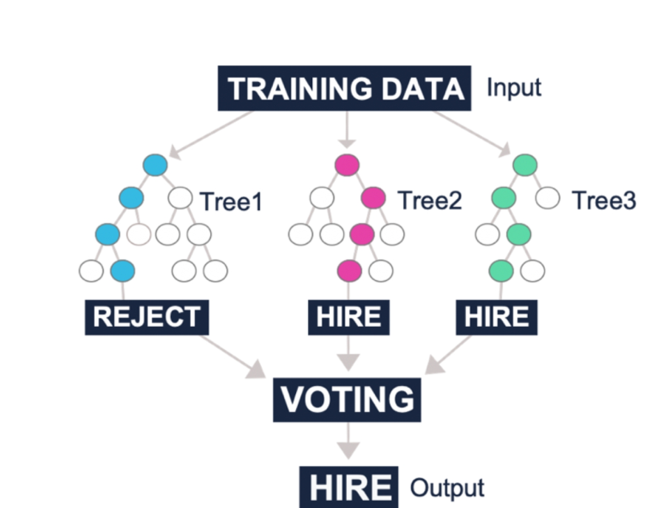
While both methods grow multiple trees and use bootstrap sampling to randomize the data, random forests artificially limit the number of variables considered for each split. In other words, the algorithm is not permitted to take into account all n variables at each partition.

Another feature of random forests is that when training a tree, the search for the best split is limited to a subset of the original features chosen at random. Each split node has a different set of random subsets. The goal is to introduce more randomness into the learning procedure in order to try to decorrelate the prediction errors of the individual trees.

As an outcome, random forests employ randomization on both axes of the data matrix:

· by bootstrapping samples from every tree in the forest

· At each node of the tree, a subset of features is randomly chosen.



*Figure 3: Example of growing random trees to produce a prediction[5]*

**2.4. Decision Tree algorithm**

**2.4.1. Overview**

A decision tree is a set of tests that delivers a suitable categorization at each stage of a study.

Decision trees categorize instances by sorting them along the tree from the root node to some leaf node, which offers the classification of the instance. Each node in the tree represents a test of some property of the instance, and each branch descending from that node represents one of the possible values for that attribute.

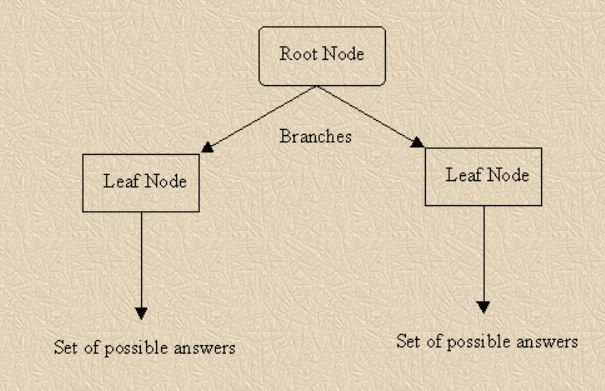
Starting at the root node of the decision tree, an instance is categorized by testing the attribute indicated by this node, then progressing along the tree branch according to the value of the attribute. This process is then repeated at the next node on the branch, and so on until a leaf node is reached.

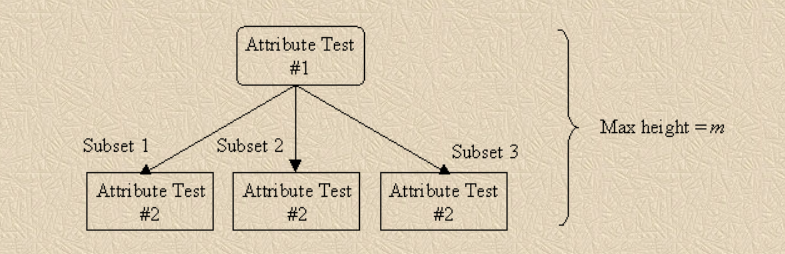
## **2.4.2.** **Specialities**

Each non-leaf node is linked to a test, which divides the available responses into subgroups according to various test outcomes.

Each branch transports a portion of a given test result to another node.

Each node is linked to a collection of possible replies.

*****Figure 4: Diagram of decision tree[7]*

*****Figure 5: Occam’s Razor[7]*

**2.5. Logistic Regression algorithm**

**2.5.1. Overview**

Logistic regression is a statistical approach for determining the relationship between one or more categorical or continuous independent (explanatory) variables and the categorical dependent (response) variable.

## **2.4.2.** **Specialities**

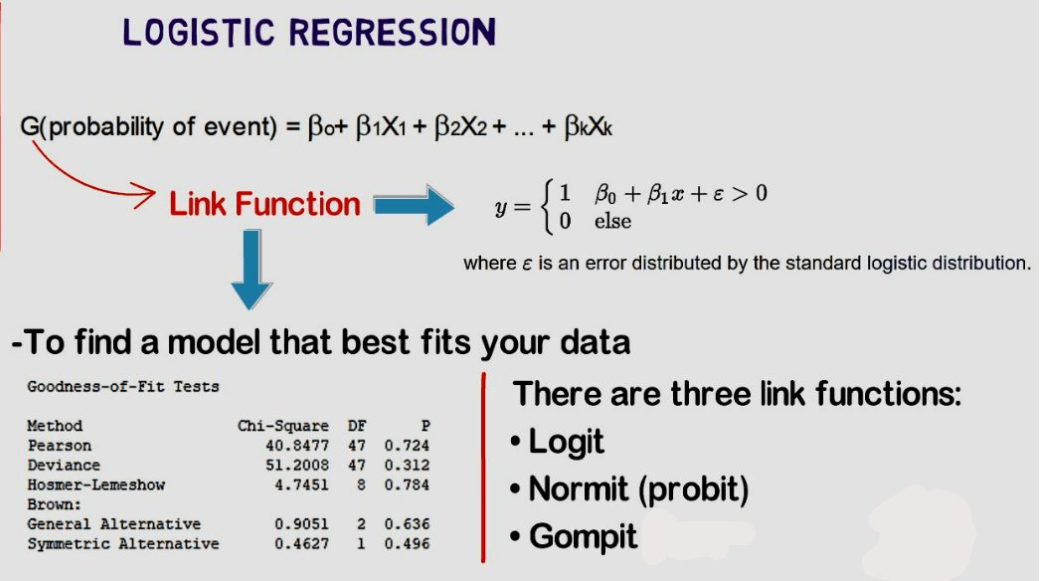
The dependent variable should be categorical (binary, ordinal, nominal, or count occurrences).

The predictor or independent variable should be categorical or continuous.

The correlation between predictors or independent variables (multi-collinearity) should be minimal, although independent variable linearity and log odds exist.

The data should be a representative sample of the population, and it should be recorded in the order in which it was obtained.

The model should produce an accurate match to the data.



*Figure 6: Logistic Regression [8]*

**2.6.** **Applying Logistic Regression for model**

## **2.6.1.** **Strengths**

One of the common machine learning algorithms with the highest accuracy available.

Easily well implemented.

Many sectors of science and technology can benefit from categorization.

Classification issues are frequently used to describe image recognition tasks.

# **2.6.2.** **Build Logistic Regression model**

Main steps:

Removing missing data columns

Using Dummy Variables to Handle Categorical Data

Removing unnecessary columns of the dataset

Creating training and testing

Train Logreg model

Making Predictions from our Model

Measure the performance of our model

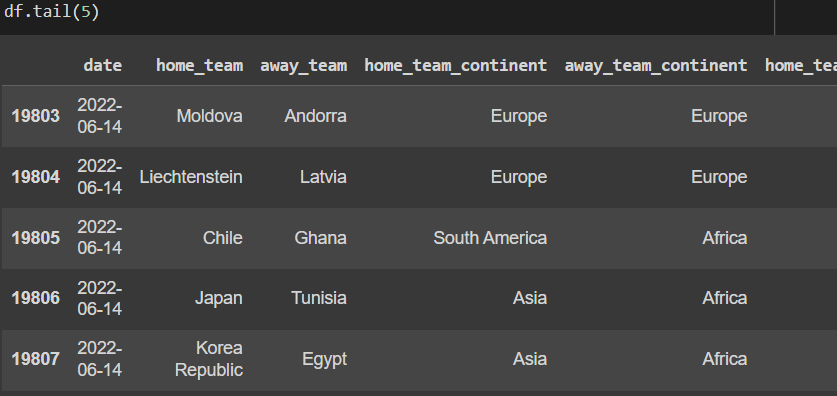
# **CHAPTER 3 - IMPLEMENTING**

The implementation will be executed using python language on Google Colab. It is a strong open-source platform that provides many useful libraries for machine learning which is recommended.

# **3.1.** **WC Dataset structure**

The dataset Data\_WorldCup2022.csv is shorted by the original. I only selected the year 2000 up to now. The WC data has 25 columns.

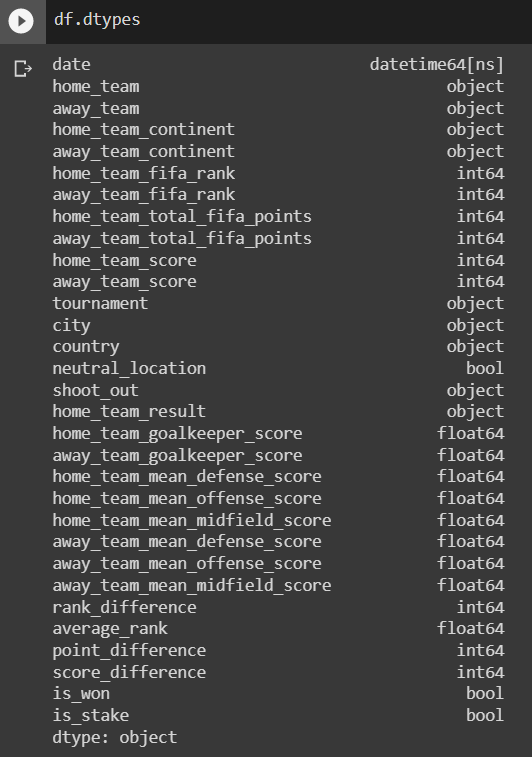
1. Date: Date of the match, YY-DD-MM
2. Home\_team: name of the teams that play at home.
3. Away\_team: name of the teams that play away.
4. Home\_team\_continent: the home teams which have their continent
5. Away\_team\_continent: the away teams which have their continent
6. Home\_team\_fifa\_rank: ranking of the home teams
7. Away\_team\_fifa\_rank: ranking of the away teams
8. Home\_team\_total\_fifa\_points: total points in the home teams, NA
9. Away\_team\_total\_fifa\_points: total points in the away teams, NA
10. Home\_team\_score: number of goals in one game of home teams
11. Away\_team\_score: number of goals in one game of away teams
12. Tournament: the name of the tournament that all teams join in.
13. City: the city held the match.
14. Country: the country held the match.
15. Neutral\_location: the third country that held the match, however, they don’t participate in it.
16. Shoot\_out: numbers of the ball shot by the football players, NA.
17. Home\_team\_result: the results of the home teams(win, lose, and draw).
18. Home\_team\_goalkeeper\_score: The score of the goalkeeper of the home team, NA.
19. Away\_team\_goalkeeper\_score: The score of the goalkeeper of the away team, NA.
20. Home\_team\_mean\_defense\_score: The mean score defense of the home team, NA.
21. Away\_team\_mean\_defense\_score: The mean score defense of the away team, NA.
22. Home\_team\_mean\_offense\_score: The mean score offense of the home team, NA.
23. Away\_team\_mean\_offense\_score: The mean score offense of the away team, NA.
24. Home\_team\_mean\_midfield\_score: The mean score midfield of the home team, NA.
25. Away\_team\_mean\_midfield\_score: The mean score midfield of the away team, NA.



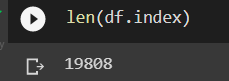
*Figure 7: Five last rows of the dataset.*

# **3.2.** **Exploratory Data Analysis**

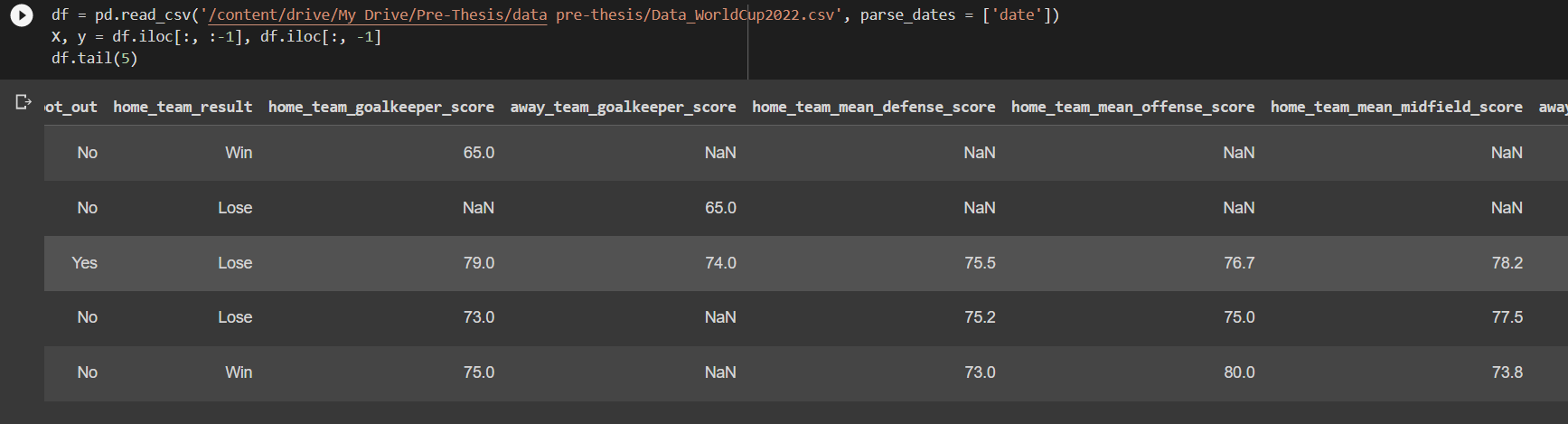
In machine learning, there is an import step called encode. Even if columns have the same type name, there might be differences that cause data conflict. Encode is the step that puts columns into states that the machine can figure out how to learn.



*Figure 8: Datatypes of the dataset*

19808

*Figure 9: Number of rows of data*



*Figure 10: Dataset before encoding*

After sorting the columns needed, there are about 5 rows. The last column is considered an output, others are inputs. In the case of this dataset, the main six columns are needed: date, home\_team, away\_team, home\_team\_fifa\_rank, away\_team\_fifa\_rank, and home\_team\_score.

→As seen in the figure above, the values are numbers and NaNs. We have to edit it in a way that the machine can understand and study it.

→ Encoding is an essential step.

- Filling means of NA’s for the missing value of offense, defense, and midfield score of home/away teams.

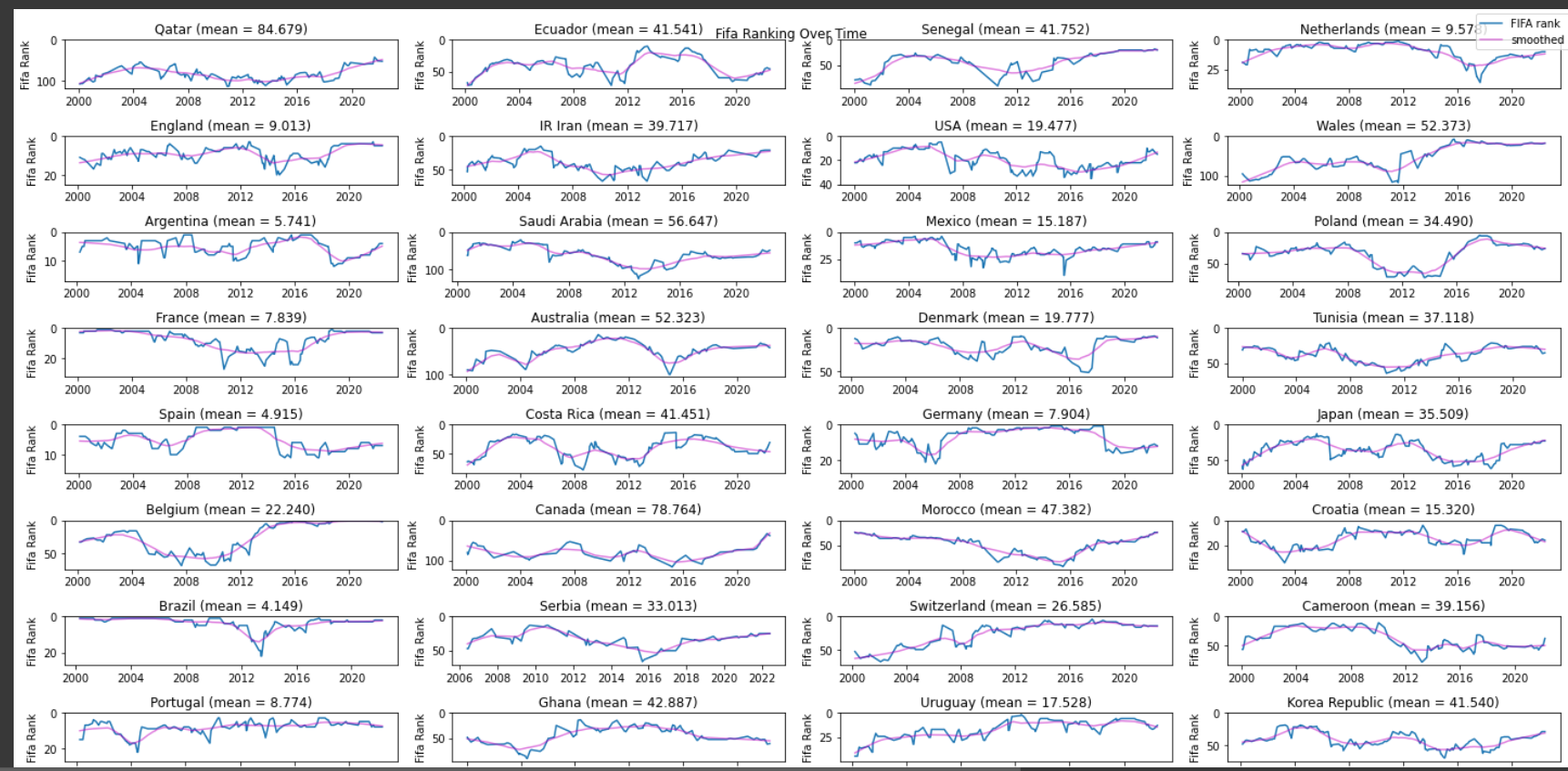
- With teams not available in FIFA, so they are not less than average performing teams, so giving an average score of 50 for all.



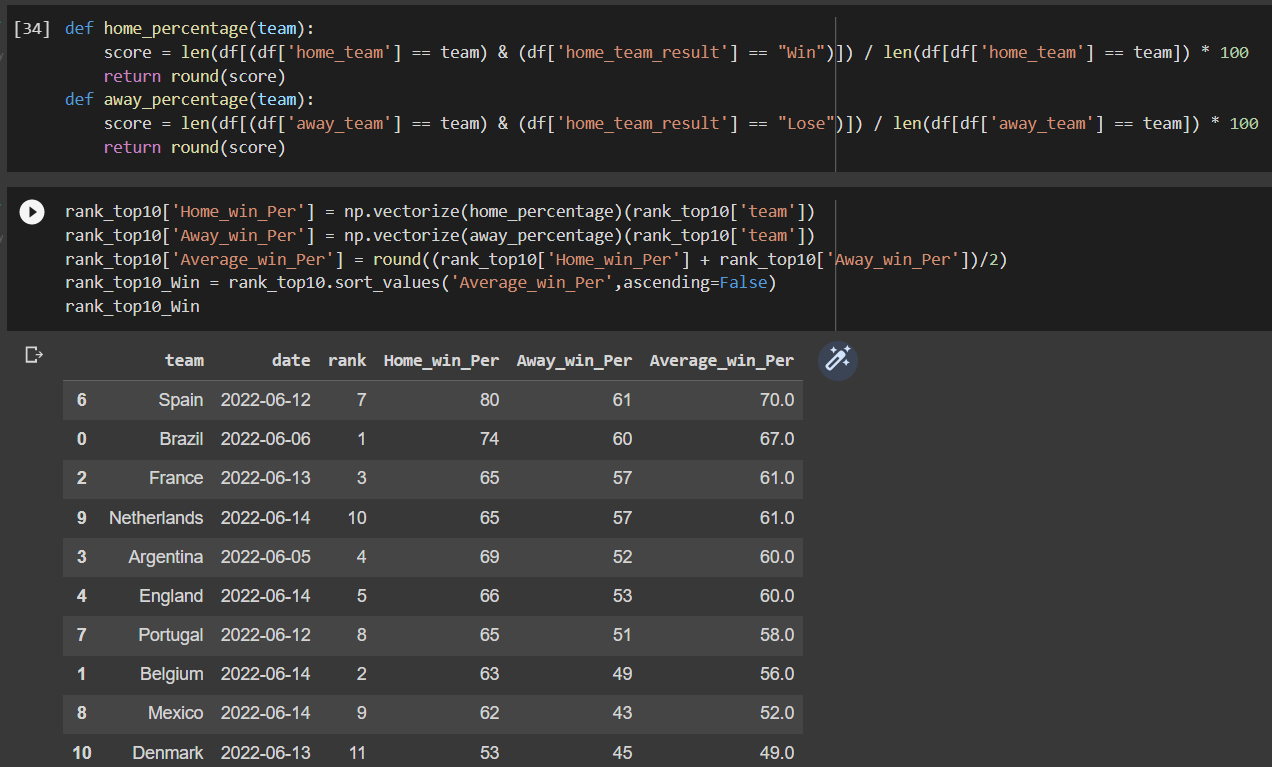
*Figure 11: Dataset after encoding*



*Figure 12: 32 Teams Participating in WC 2022*

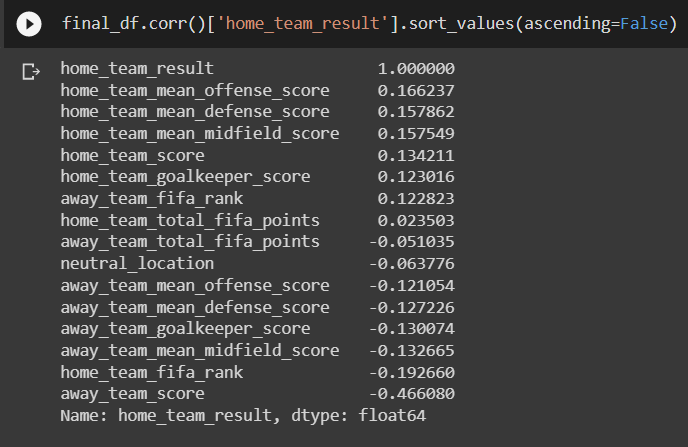


*Figure 13: Mean ranking of 32 countries in WC 2022 (following the FIFA ranking)*

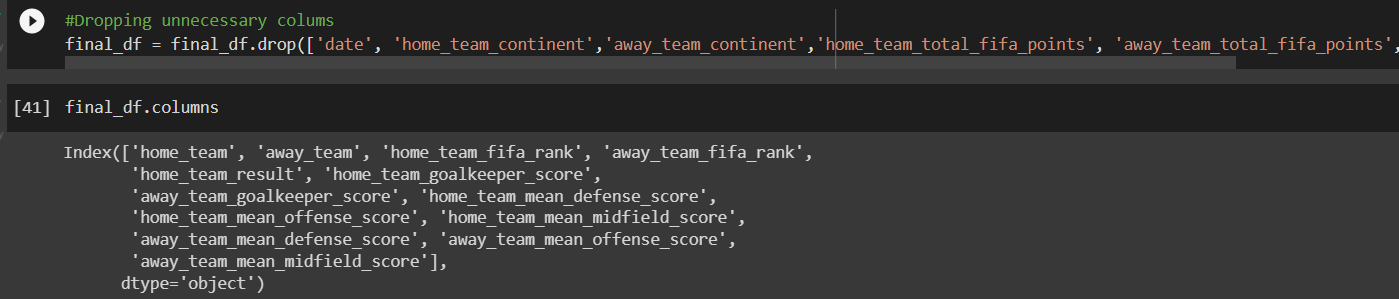


*Figure 14: Top 10 ranking calculated by average win percentage*

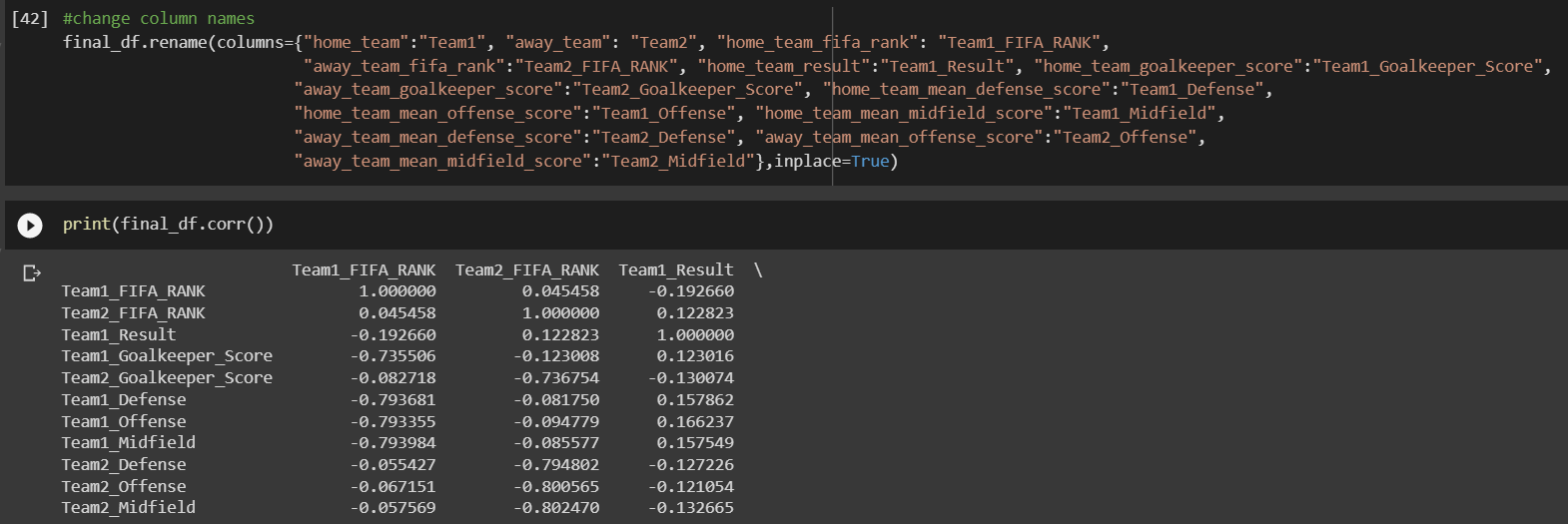
**3.2.1** **Find the correlation, drop unnecessary columns, and change the column names**



*Figure 15: Find the correlation of home team results*

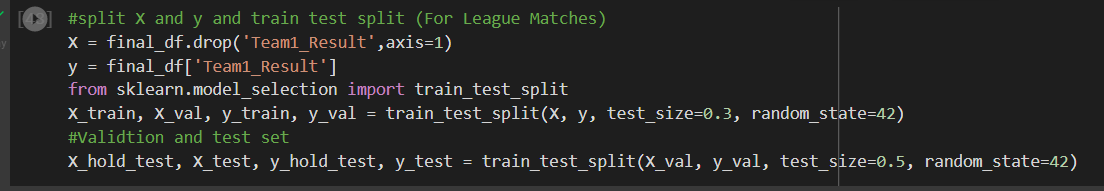


*Figure 16: Dropping unnecessary columns*



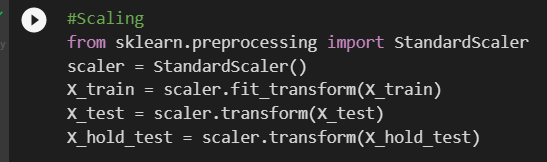
*Figure 17: Change the column names to the short name of the variable*

**3.2.2** **Train Test Split**



*Figure 18: Train Test Split*

**3.2.3** **Scaling**

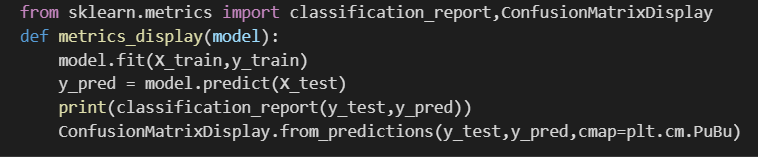
****

*Figure 19: Scaling*

**3.2.4 Confusion Matrix**

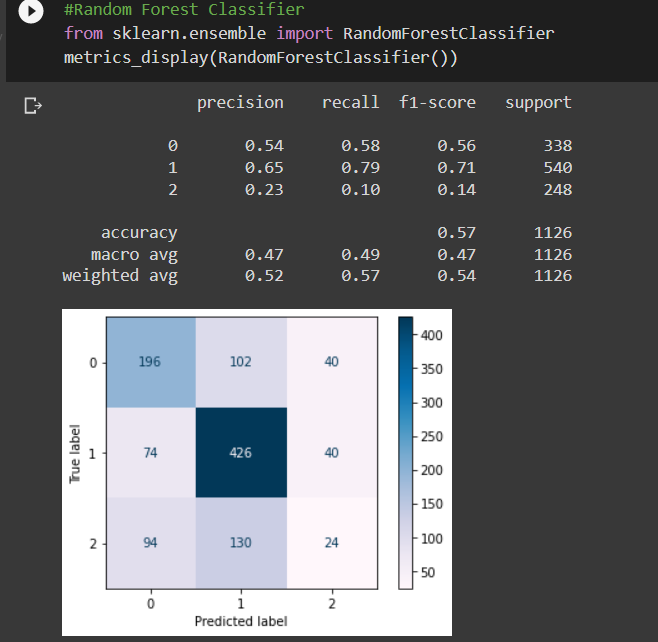
A confusion matrix is a summary of classification problem prediction outcomes.

The confusion matrix displays the many ways in which your classification model is perplexed when making predictions.

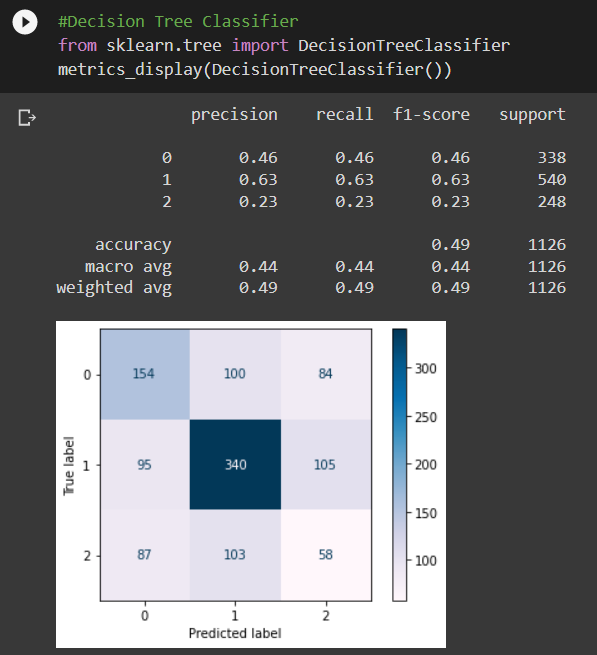


*Figure 20: Confusion Matrix*

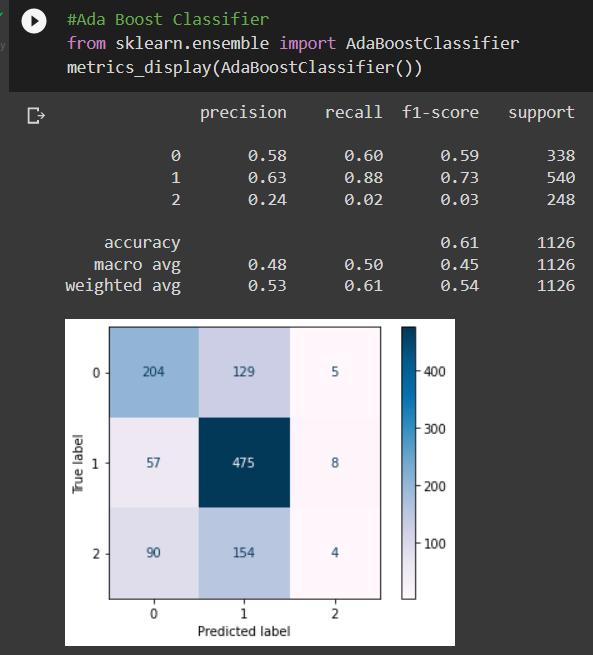
**3.2.5 Classifiers**

****

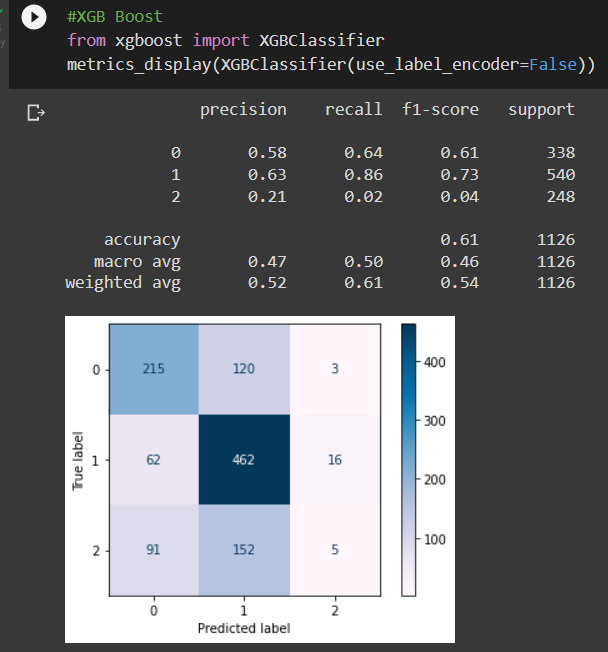
*Figure 21: Random Forest Classifiers*



*Figure 22: Decision Tree Classifiers*



*Figure 23: Ada Boost Classifier*



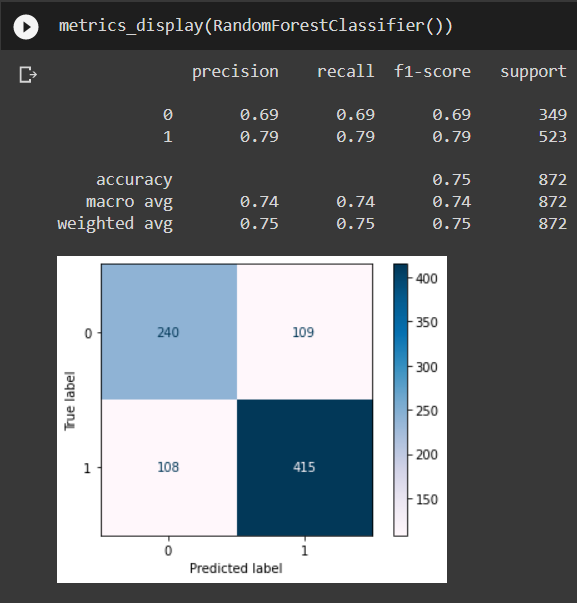
*Figure 24: XGB Boost Classifier*

Both of these algorithms are designed to turn weak learners into strong learners by updating based on residuals (XGBoost) or misclassifications (AdaBoost).

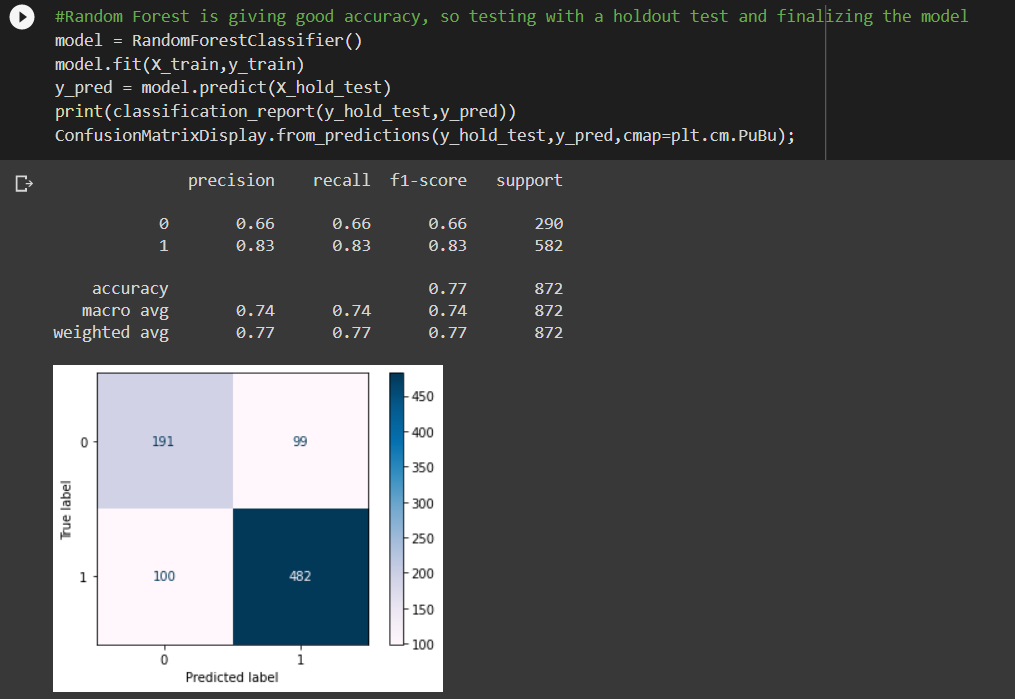
In low datasets, AdaBoost is generally resistant to overfitting and has just a few hyperparameters that must be modified to optimize model performance. However, given noisy data, its efficacy is debatable, resulting in poor performance as a result of the algorithm spending too much time learning extreme instances and skewing outcomes. Furthermore, AdaBoost is not tuned for speed, therefore it is substantially slower than XGBoost.

XGBoost was created to improve speed and performance by incorporating regularization settings to minimize overfitting and thereby successfully reducing variance. However, XGBoost is more difficult to comprehend, visualize, and modify than AdaBoost, which has a \*plethora of hyperparameters \*that may be modified to improve performance.

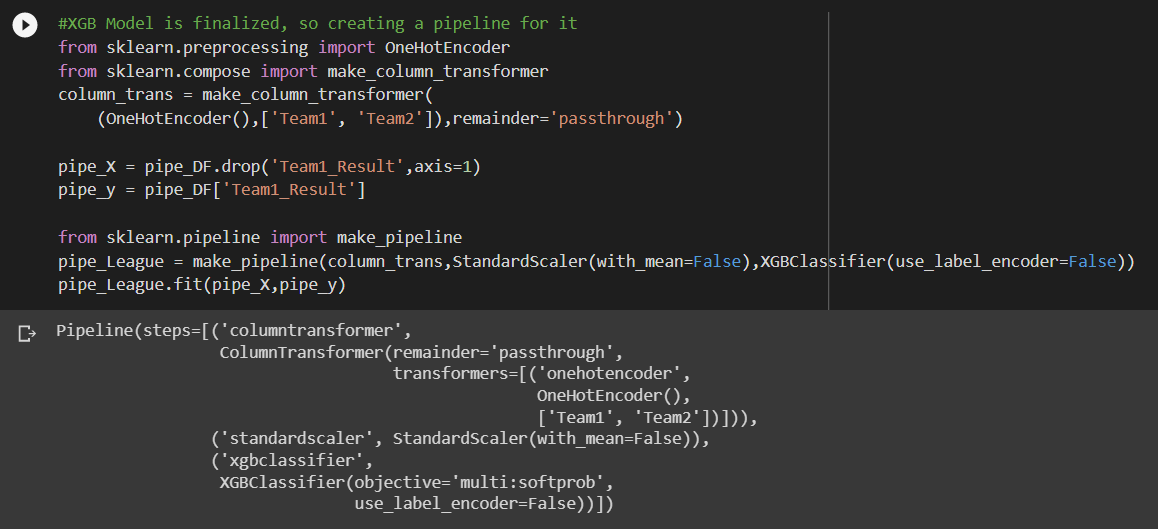
As a result, one of the key advantages of XGBoost is its lightning-fast performance when compared to competing algorithms, such as AdaBoost.



*Figure 25: Random Forest classifier when removing the drawing*

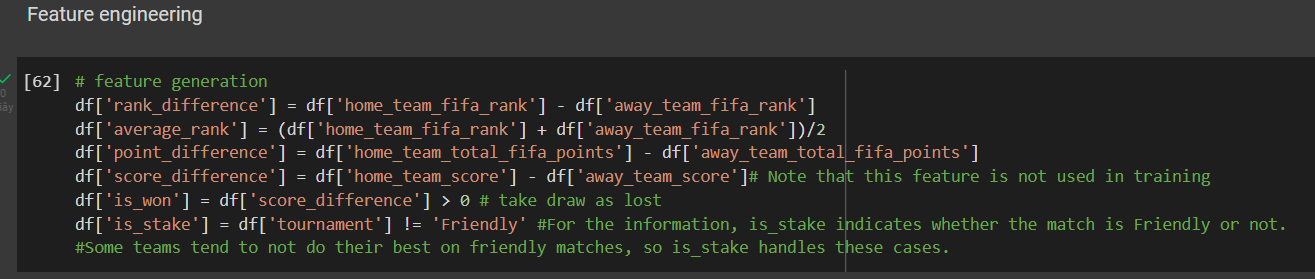


*Figure 26: Testing holdout and finalizing the model*

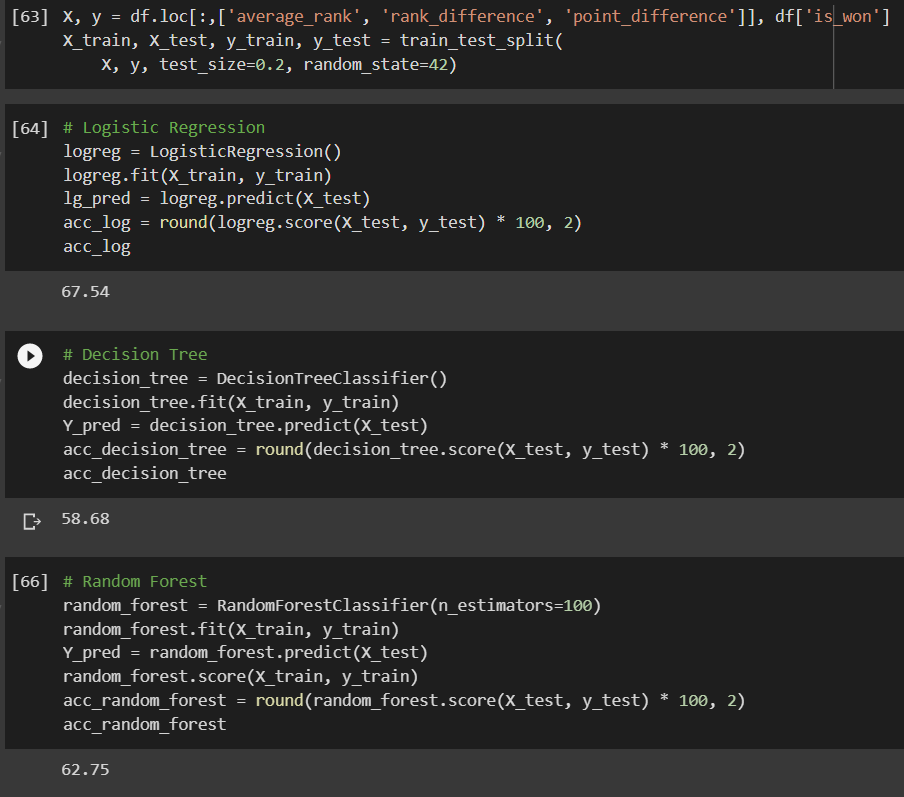


*Figure 27: Create a pipeline for the XGB model*

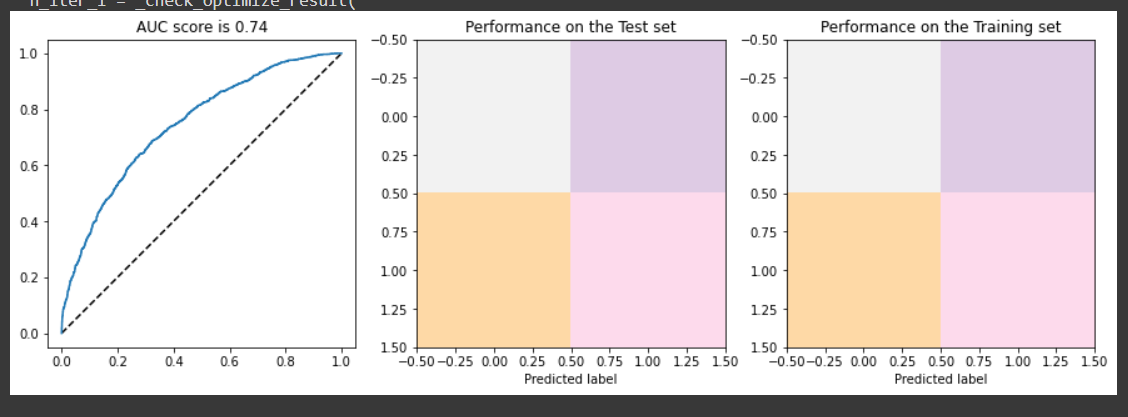
## **3.3.** **Simulation**



*Figure 28: Feature engineering*

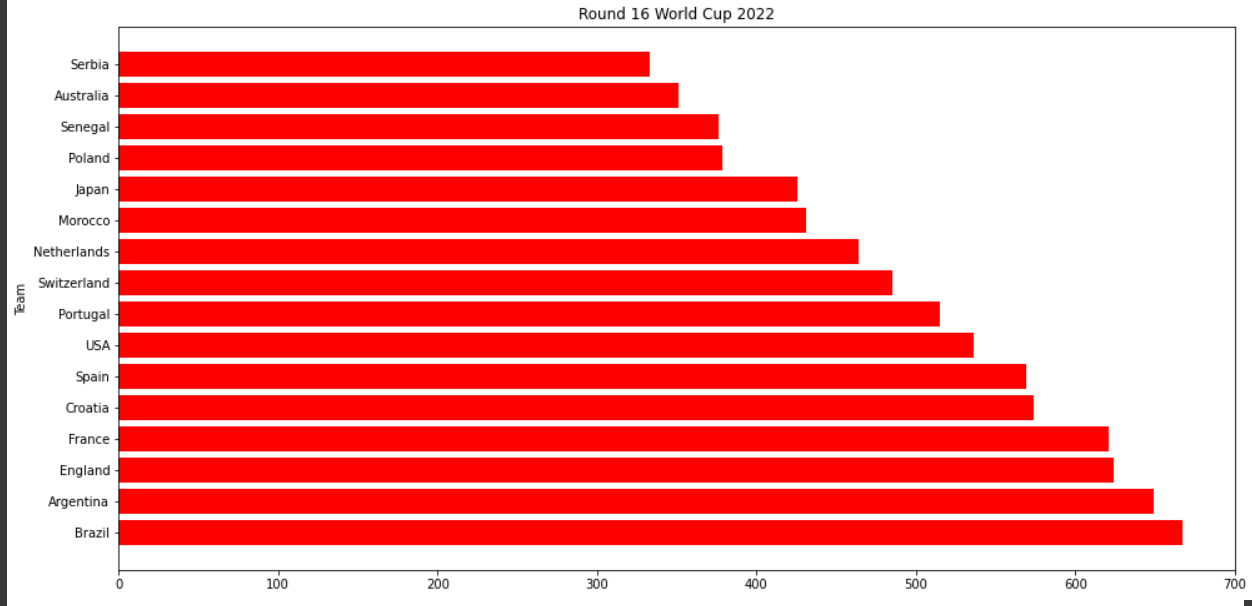


*Figure 29: Training again for the model*

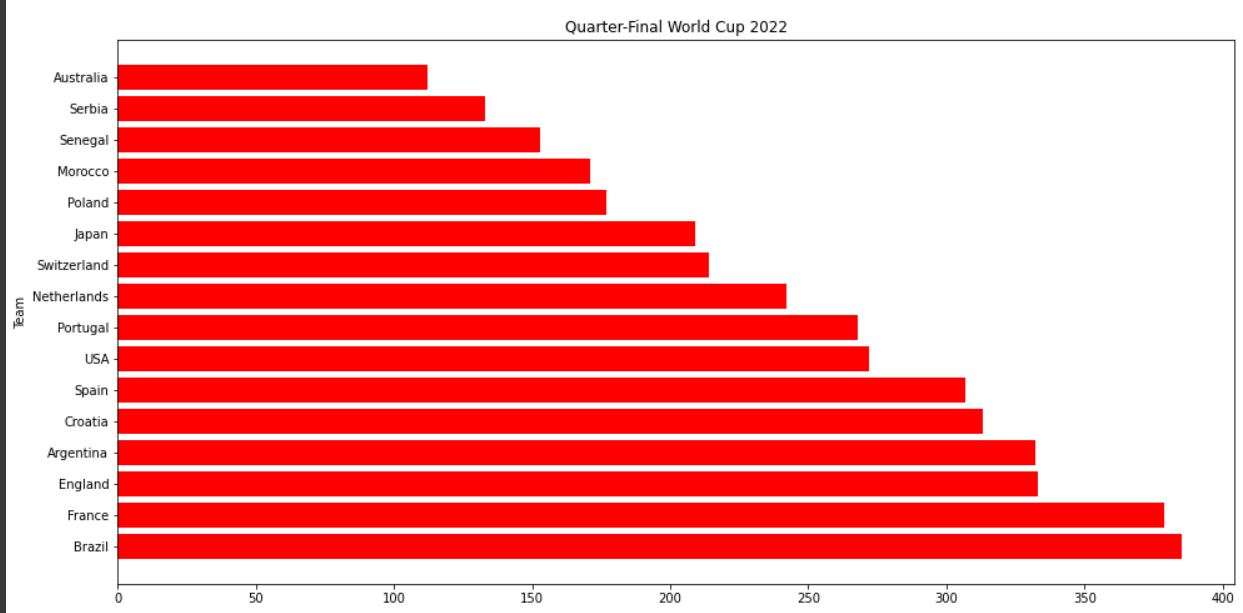


*Figure 30: Logistic Regression Model*

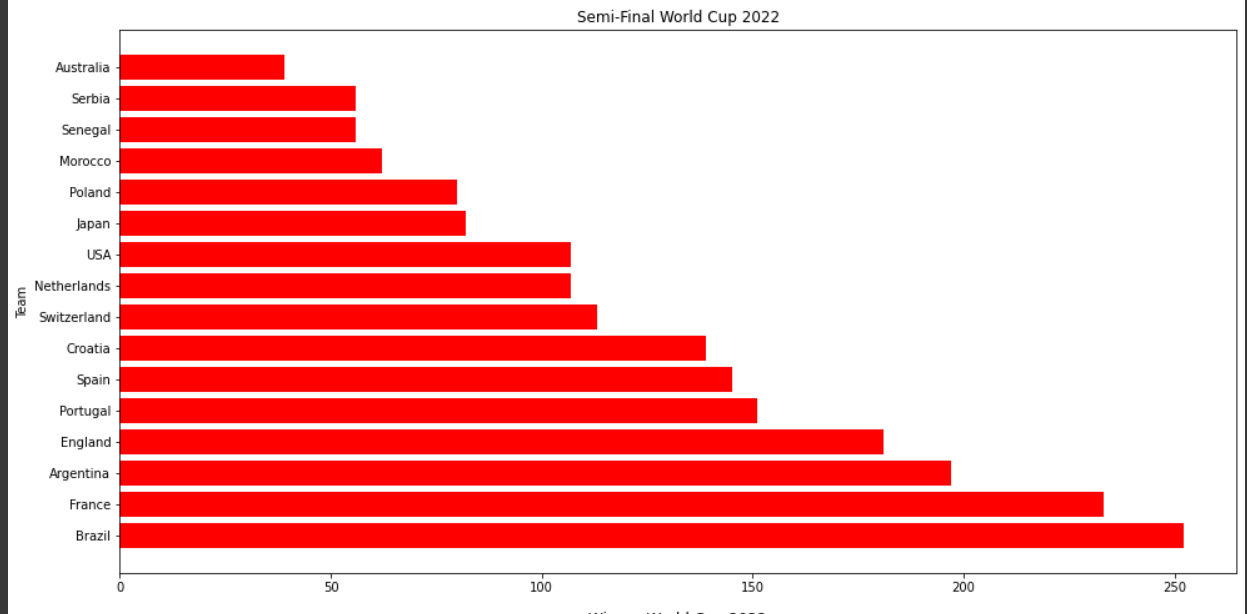
The model of simulation on the vertical is the true positive rate (TPR), and on the horizontal is the false positive rate (FPR). The AUC score is 0.74, so the model is quite good.



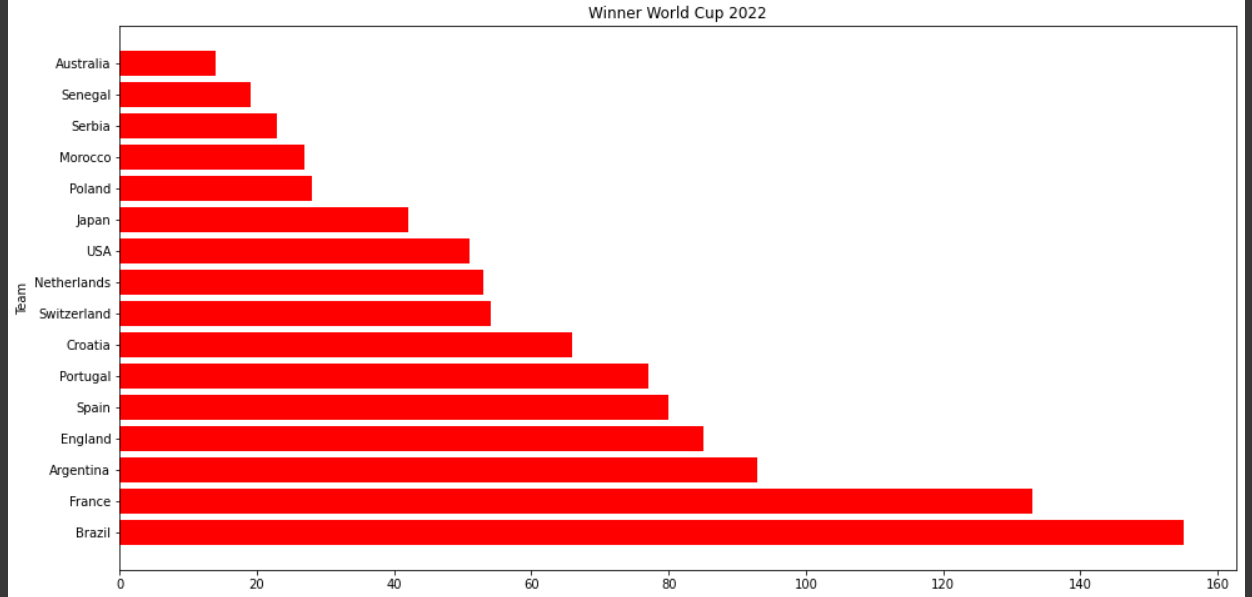
*Figure 31: Round 16 WC 2022*



*Figure 32: Quarter-Final WC 2022*

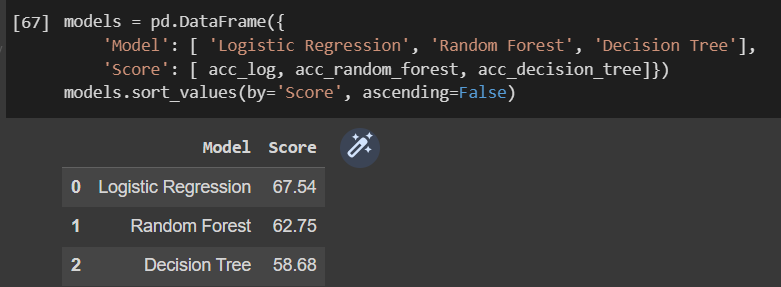


*Figure 33: Semi-Final WC 2022*



*Figure 34: Winner WC 2022*

## **3.4.** **Comparison**



*Figure 35: Comparison between Rand, DT, and Logreg*

The above figure shows that the Logistic Regression algorithm has the highest accuracy after being trained, up to 67.54%. On the other hand, the Random Forest is 62.75% and the Decision Tree is 58.68%.

Additionally, the point here is that Logreg only takes the machine about 1.30 minutes to finish the training session. The others disappointedly take up to nearly an hour each.

# **CHAPTER 4 - CONCLUSION AND FUTURE WORK**

# **4.1.** **Conclusion**

Random Forest, Decision Tree, and Logistic Regression have shown dominance in terms of training and testing an enormous dataset in comparison to other machine learning methods under the name of supervised learning. Not only do these algorithms have a faster learning procedure, but it is also more accurate. This is undeniable proof that these algorithms are among the bests of ML for large datasets since they provide speed and accuracy to the overwhelming state. Therefore, Rand, DT, and Logreg is the correct choice that the machine needs to use to study how to predict WC 2022.

# **4.2.** **Future work**

With my passion, I hope I will build web betting to create the finished product not only to improve my knowledge but also to earn money from that. Moreover, machine learning is an extremely important role in that I unlock many new things. I will try my best to research the others algorithms to implement for my future models.

# **CHAPTER 5 - References**

*[1] Das, R., & Ahmed, W. (2022, December 12). World Cup 2022: crunching 150 years of big data to predict the winner. The Conversation.*

*https://theconversation.com/world-cup-2022-crunching-150-years-of-big-data-to-predict-the-winner-196418*

*[2] Introduction to Machine Learning with Python. (n.d.). Google Books.*

*https://books.google.com.vn/books?id=1-4lDQAAQBAJ*

*[3] TGM RESEARCH: MARKET RESEARCH AND ONLINE PANELS. (2022, November 4).*

*TGM World Cup Global Survey 2022 | Research Report | Football Survey. TGM Research.*

[*https://tgmresearch.com/tgm-global-football-world-cup-survey-2022.html/*](https://tgmresearch.com/tgm-global-football-world-cup-survey-2022.html/)

*[4] S, I. (2022, May 15). Introduction to Machine Learning. Quasar Community Blog.*

*https://blog.quasarcommunity.org/introduction-to-machine-learning*

*[5]Machine learning for absolute beginners. (n.d.). Google Books.*

[*https://books.google.com.vn/books?id=Ex8\_tAEACAAJ*](https://books.google.com.vn/books?id=Ex8_tAEACAAJ)

*[6] FIFA World Cup 2022 ⚽️🏆. (2022, August 28). Kaggle.*

*the* [*https://www.kaggle.com/datasets/brenda89/fifa-world-cup-2022*](https://www.kaggle.com/datasets/brenda89/fifa-world-cup-2022)

*[7] The University of Regina DBD. (n.d.). Machine Learning/Inductive Inference/Decision*

*Trees/Overview.*

*http://www2.cs.uregina.ca/%7Edbd/cs831/notes/ml/dtrees/4\_dtrees1.html*

*[8] https://www.scribbr.com/citation/generator/folders/5ssPBTT1N6Ua5U4qHvVG14/lists/46Zp3CWTvshbLnsF6wo4pc/*