A Study on the Validity of Home Field Advantages in Football Matches using Machine Learning

Jose Lorenzo M. Cansana [1], Gabriel Rasheed H. Jamalul [2],
Leanne Marie C. Loyola [3] and Giann Jericho Mari F. Marasigan [4]

De La Salle University, 2401 Taft Ave, Malate, Manila, 1004 Metro Manila
[1] jose_lorenzo_cansana@dlsu.edu.ph
[2] gabriel_jamalul@dlsu.edu.ph
[3] leanne_loyola@dlsu.edu.ph
[4] giann_marasigan@dlsu.edu.ph

Abstract. In competitive sports, teams playing at home are usually considered to have an advantage over teams playing away from home. It is a well-known phenomenon that playing at home field is an advantageous condition for professional short teams more specifically in football which is known to be the most popular sport in the world with over 250 million players across 200 countries. For this reason, the home field advantage in team sports is an important issue to be explored. The aim of this project is to focus on the current state of knowledge with regards to home advantage in football, whether it is the real thing or just a phenomenon. Using machine learning to analyze the chosen dataset will be able to determine the validity of home advantage based from the scores from both teams and location of the event. For this reason, 41,540 football matches from 1872 to 2019 were analyzed. The machine learning pipeline used was similar to Andrew Ng's and Ian Goodfellow. After getting the best model by hyperparameter tuning, the best accuracy score the researchers were confident on was 66%.

Keywords: Machine Learning, Home field advantage, Football, Decision Trees, Data Splitting

1 Introduction

1.1 Data Science for Football

As one of the most popular sports on the planet, football has always been followed very closely by a large number of people. In recent years, new types of data have been collected for many games in various countries, such as play-by-play data including information on each shot or pass made in a match. The collection of this data has placed Data Science on the forefront of the football industry with many possible uses and applications:

- Match strategy, tactics, and analysis
- Identifying players' playing styles
- Player acquisition, player valuation, and team spending
- Training regimens and focus
- Injury prediction and prevention using test results and workloads
- Performance management and prediction
- Match outcome and league table prediction
- Tournament design and scheduling
- Betting odds calculation

In particular, the betting market has grown very rapidly in the last decade, thanks to increased coverage of live football matches as well as higher accessibility to betting websites due to the development of mobile and tablet devices.

1.2 Background and Motivation

In competitive sports, teams playing at home are usually considered to have an advantage over teams playing away from home [1]. Many studies and explanations have been proposed to explicate this competition reality phenomena.

However, the precise cause and the way in which they affect performance are still not clear. In team sports, a team playing at its own stadium or arena is known as the home team, the other team is known as the visitors. Teams typically play their home games in or near their home region; they will generally have half of their total games at home in a season based on the structure of the competition and fixtures. It was indicated that the number of home game wins usually exceeds the number of away game wins over a balanced home and away competition [2].

Football is known to be the most popular sport in the world with over 250 million players across 200 countries. For over decades, or even centuries, countries gather in a single place and compete with each other. Sometimes it is only for a friendly match, to display exhibition and entertainment for international fans, but mostly countries are competing for a trophy. Although the home field advantage is an important factor in determining the results of football matches, the exact reasons for this are not yet clear. The first paper to consider the concept of home advantage applied exclusively to soccer was by Dowie in 1982 in which he commented on the success of countries hosting the World Cup and considered three possible causes of the advantage in soccer which he labeled fatigue, familiarity and fans [3]. Data from the Football League in England were used, but no clear-cut conclusions were reached. A more detailed study by Pollard International Journal of Soccer and Science Journal Vol. 3 No 1 2005 29 in 1986 soon followed and this still serves as the starting point for a general review of the way in which home advantage applies to football [4]. Data from various competitions in England and Europe were used to assess the effects on home advantage of crowd support, travel fatigue, familiarity, referee bias, tactics and psychological factors. The home field advantage has also become a subject, about whose existence the football players, the technical team, the fans and the media organizations talk from time to time, but they cannot exactly be sure of its existence.

1.3 Machine Learning Approach

Decision Trees are a popular Machine Learning technique to link input variables, represented in the tree's branches and nodes, with an output value represented in the tree's leaves. Trees can both be used in classification problems, by outputting a category label, or in regression problems, by outputting a real number. With that said, this project will be using decision tree for classification purposes.

Decision Trees have the advantage that they scale very well with additional data, they are quite robust to irrelevant features and they are interpretable: the choices at each node allows us to understand the impact of each predictor variable towards the outcome [5].

1.4 Objective

The aim of this project is to focus on the current state of knowledge with regards to home advantage in football, whether it is the real thing or just a phenomenon. Using machine learning to analyze the chosen dataset will be able to determine the validity of home advantage based from the scores from both teams and location of the event.

2 International Football Results Dataset

2.1 Data origin

The researchers have obtained their dataset from the Kaggle Data Science website called 'Kaggle International football results from 1872 to 2019' [6]. This database has been made publicly available and This dataset includes 41,540 results of international football matches starting from the very first official match in 1972 up to 2019. The matches range from FIFA World Cup to FIFI Wild Cup to regular friendly matches. The matches are strictly men's

full internationals and the data does not include Olympic Games or matches where at least one of the teams was the nation's B-team, U-23 or a league select team.

All of the data from the dataset have been used. It includes the following columns:

- **date** - date of the match
- **home_team** - the name of the home team
- **away_team** - the name of the away team
- **home_score** - full-time home team score including extra time, not including penalty-shootouts
- **away_score** - full-time away team score including extra time, not including penalty-shootouts
- **tournament** - the name of the tournament
- **city** - the name of the city/town/administrative unit where the match was played
- **country** - the name of the country where the match was played
- **neutral** - TRUE/FALSE column indicating whether the match was played at a neutral venue

2.2 Data pre-processing

An important step before building our model is to analyse and pre-process the data to make sure that it is in a usable format for us to use when training and testing different models.The researchers built a script in python to extract the chosen data and store it in new tables and some columns of the data have been dropped because it does not affect the results. The following are snippets of the code:

```
import pandas as pd

# reading csv file
data = pd.read_csv('results.csv')
data = data.drop('date', 1)
data = data.drop('tournament', 1)
data = data.drop('city', 1)
data = data.drop('country', 1)
```

```
import numpy as np

def handle_non_numerical_data(data):
    columns = data.columns.values
    for column in columns:
        text_digit_vals = {}
        def convert_to_int(val):
            return text_digit_vals[val]

        if data[column].dtype != np.int64 and data[column].dtype != np.float64:
            column_contents = data[column].values.tolist()
            unique_elements = set(column_contents)
            x = 0
            for unique in unique_elements:
                if unique not in text_digit_vals:
                    text_digit_vals[unique] = x
                    x+=1

            data[column] = list(map(convert_to_int, data[column]))

    return data

data2 = handle_non_numerical_data(data)
X = data[data.columns[:-1]]
y = data[data.columns[-1]]
```

2.3 Data Features

Data features are the measurable properties or characteristics of a phenomenon being observed. The following columns are the ones used as features.

- **home_team** - the name of the home team
- **away_team** - the name of the away team
- **home_score** - full-time home team score including extra time, not including penalty-shootouts
- **away_score** - full-time away team score including extra time, not including penalty-shootouts

2.4 Data Labels

Data labels are the final choice, meaning they are the output. Once we have trained our model, this will give it sets of new input containing those features. For this project, the researchers have added **home_winner**, a TRUE/FALSE column indicating whether the match was won by the home team.

3 Methodology

3.1 General pipeline

In this section, we will present the general pipeline that we have set up in order to easily be able to train and test different models and compare their performances. The pipeline the researchers have created is illustrated in Fig.3.1

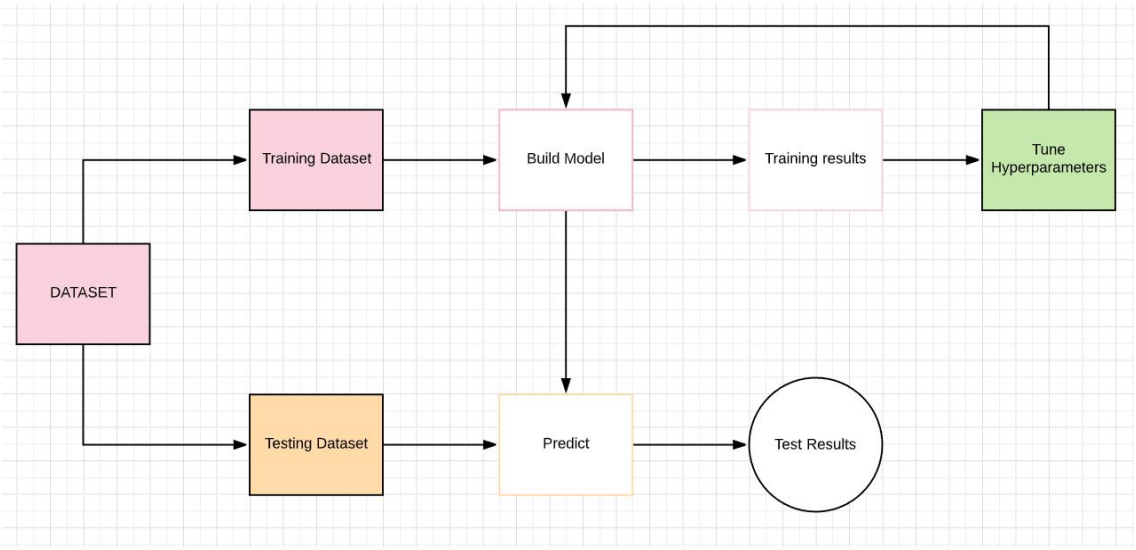

Fig. 3.1. General project pipeline

The machine learning adapted this pipeline which will be explained below:

3.2. Process of the Pipeline

Data Splitting. In this project, the football dataset was divided into two parts: test data and training data. The split was 80-20, which favors the testing data because there are numerous data on the dataset, which is 40,000. The split

of the data is done randomly but the percentage of how it is split can be tuned by the researchers. They can lessen it to 60-40, 50-50, etc., but the data within the splits are random.

Initial Model. From the dataset, it would be split into training data and testing data. As mentioned above, the researchers decided to split the data which favored the testing data (more data goes to testing data). Once the dataset is now split, the training data will build its initial model given the data it has and its initial understanding of the dataset. From that, it will output the training results.

Tuning of Hyperparameters. The researchers will now tune the hyperparameters of the program if they are not confident with the initial results. This process would iterate until the researchers are confident enough to test it with the testing data.

Final Model. Once the final model is built, it will now predict based from the testing dataset and this will now generate the result of the program.

3.4 Purpose of the Process

The researchers chose to split the data which favored the testing dataset because they are confident that the accuracy of the program will be more significant based from the researchers trial and errors. Also, with the large dataset, it allows the researchers to tinker with the data and accuracy scores. The model will be changed throughout the program until the researchers are confident enough to test it with the testing data. With the tuning of the hyperparameters, they can determine the best fit or model to accurately represent the program, which will lead to their desired goal.

3.5 Input and Output

In the dataset, the inputs included the home and away team, home and away scores, type of tournament, country and city, date, and the label which is home winner. The researchers decided to drop the date, country and city, and tournament because the program felt inaccurate. With these number of features being used, the program got a high accuracy score, which is unreliable. Therefore, the researchers only used the home and away team, home and away scores, and the label as the inputs. The output of the program is the accuracy score of the final model that was tested with the testing datasets.

3.6 Hyperparameters

The ultimate hyperparameter of this program is the split of the data. The researchers used this function call train_test_split = (X,Y, test_size = 0.6, random_state = 5). X and Y are the inputs from the dataset, while the test size and random state will be the one that will be tuned depending on the accuracy score. If the accuracy score is low, the researchers will try to tune the test_size and random_state. The test_size determines the split of the data. In this case, 80% will go to the testing data and 20% will go to training data. On the other hand, the random_state determines the overall randomness distribution of the data.

4 Results and Analysis

4.1 Decision Tree

In this model, the hyperparameters most relevant is the size of the sample set used for testing and the randomness of the distribution of this sample set on the entire set of data. In gathering of results, the researchers used the following range of values {.45,.50,.60,.70,.80,.90,.95} as the size of the sample set. In the results, the researchers refer to the correct assumptions whether it is true or false as positives, while the mistaken assumptions as negatives. First, the precision was measured.

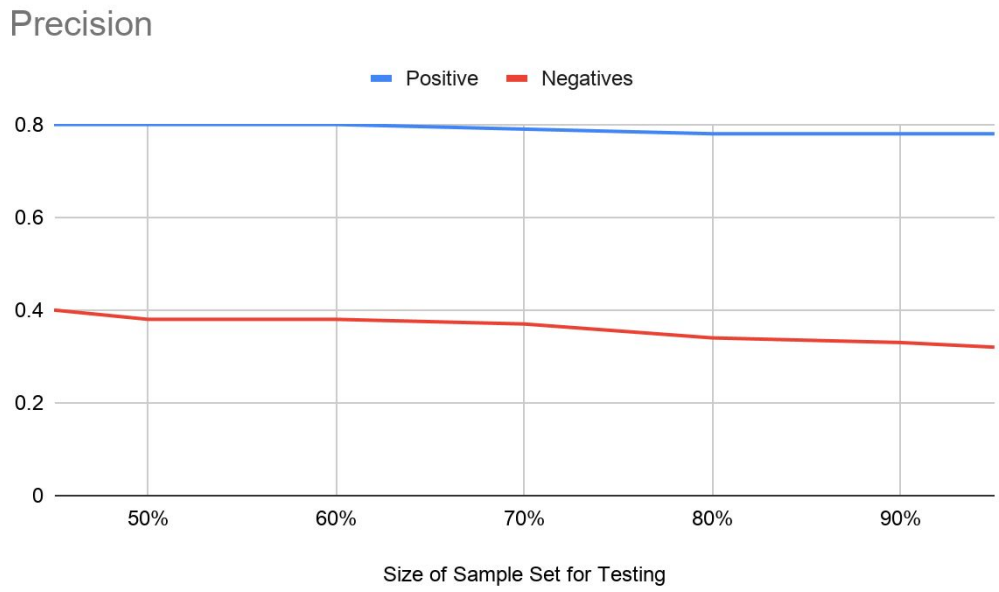

Fig.4.1.1 Precision scores on the varying sample set testings

As seen in the diagram above, there is a slight variation of precision values throughout the range. It is also noticed that there is a slight decrease of precision value as the size of sample set increases. This is because as the sample set increases, there will be a greater chance of more selected items to be irrelevant to the training model. Therefore, in precision, it is suggested to not overestimate the required size for the sample set.

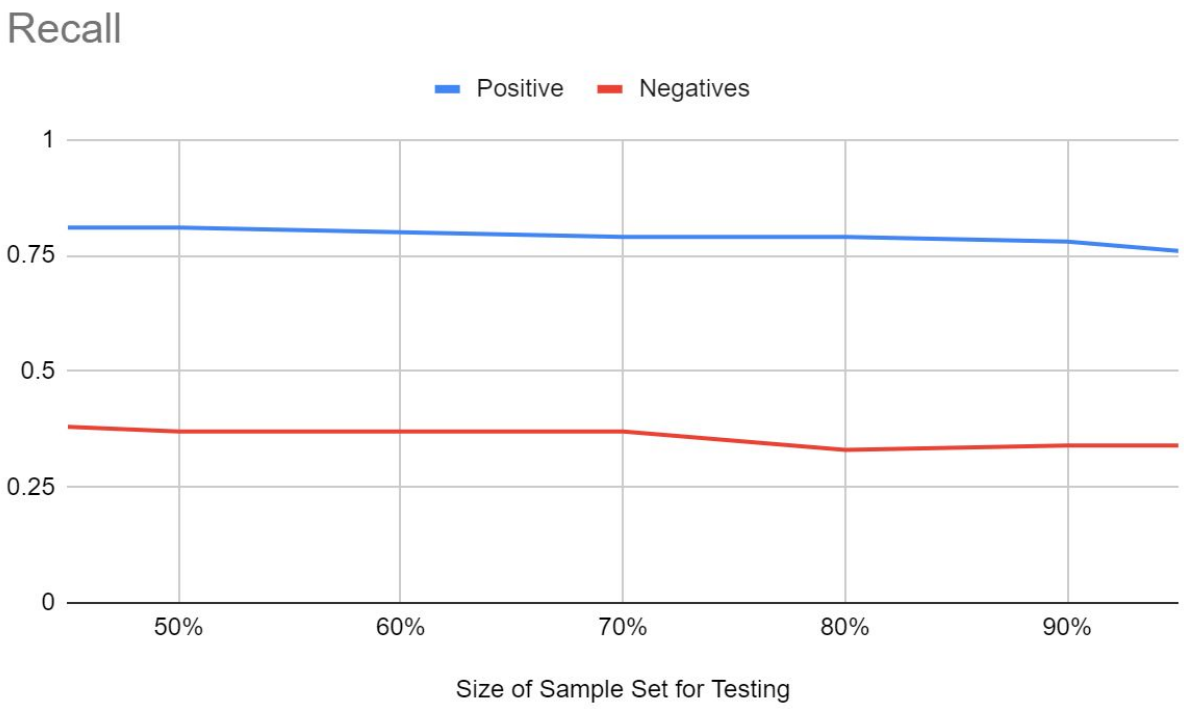

Fig.4.1.2 Recall scores on the varying sample set testings

In the recall scores gathered, also known as sensitivity of the data, it is noticeable that there is hardly any changes to the score as the size of increases. In the recall score, we get to identify how many of the relevant items are selected and based on the data gathered, we can conclude that regardless of how large sample size is, the recall scores still remains in the same area.

We can deduce that given Fig.4.1.1 and Fig.4.1.2. that the second hyperparameter is weak. It is evident in both figures that the scores does not change all too much as the sample size of the testing set increase. In other words, no matter how random the distribution of the selection of testing data is, it will not make that big of a difference in the resulting scores. Therefore, this hyperparameter, is no longer required to be tuned. From the recall and precision scores, the mean relation of both can be found, or the F-Measure (F1 score).

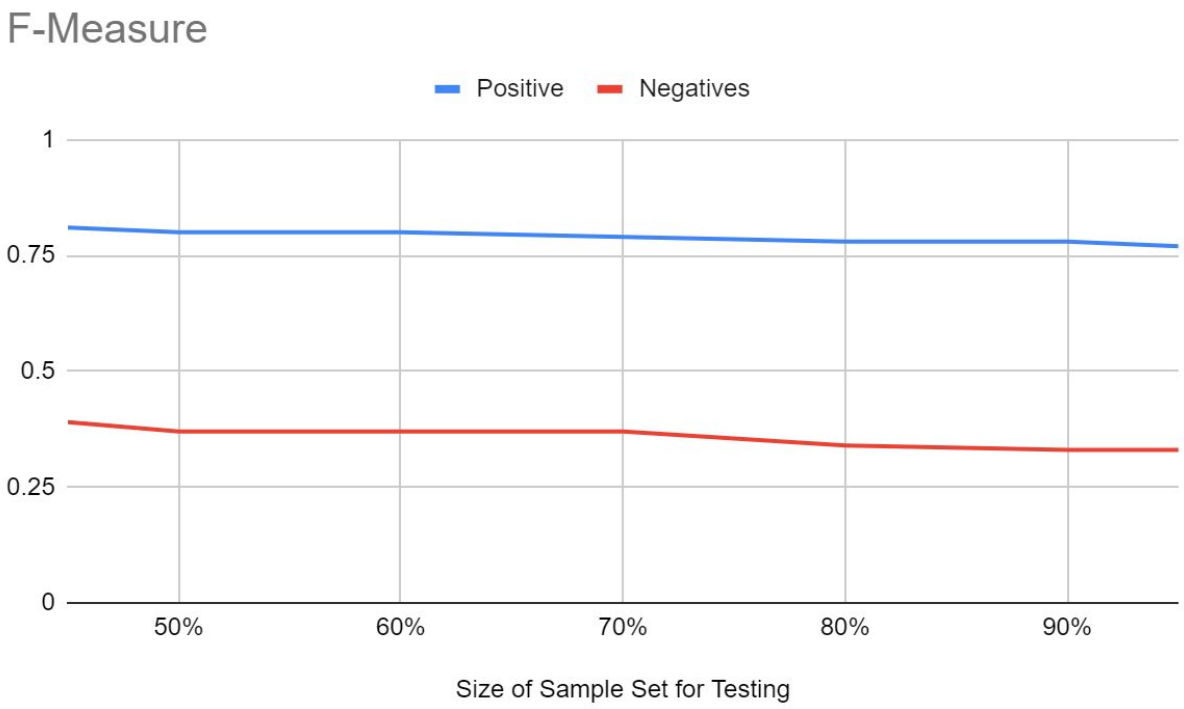

Fig.4.1.3 F1 Scores on the varying sample set testings

Given both recall and precision scores, the f-measure was computed. As seen in the figure above that as the size of the sample set increases, the gap difference between the positive and negative decreases. This change starts to increase greatly from 80% onwards. This is due to the fact that the sample data set has already reached the point of where it almost encapsulated the entire dataset. This not desired for training the model for as the testing set increases, the lower the size of data used to train the model. This is well reflected with the accuracy of the model using this range of sizes.

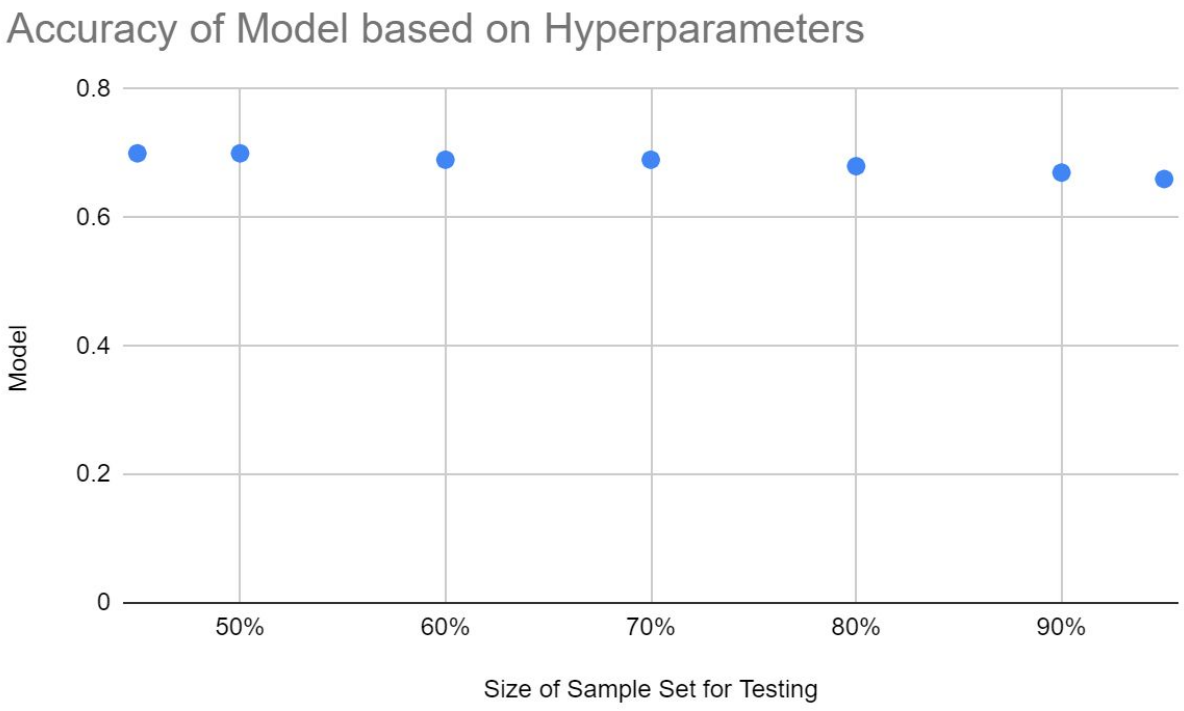

Fig.4.1.4 Accuracy of the model based on hyperparameters with the varying size of the sample set.

Similarly, deducing from the accuracy scores of the model, that having a bigger size of sample set will result in a less accurate model. Though, having a smaller size of sample set may increase the accuracy of the model by a fraction, it is still optimal to find the mean value where the sample size of the testing set is still high enough without suffering much on the accuracy. This yields the optimal hyperparameters of this model to be sample size of 80% or .80 and randomness distribution set to any value as this should not affect the measures in noticeably.

5 Conclusions and Recommendations

Using machine learning technique such as Decision trees to analyze the chosen dataset to determine and generate the validity of home advantage based from the scores from both teams and location of the event has been accomplished. Also, a model training and testing pipeline was built to quickly and easily tweak the model, and using Scikit-Learn to link the different model components together.

5.1 Model and Hyperparameters

After numerous testing, training and tuning, the most accurate model with its hyperparameter the researchers concluded was train_test_split = (X,Y, test_size = 0.8, random_state = 5). With test data size receiving 80% of the dataset and randomness of the data with 5. The accuracy score the researchers got with this hyperparameter was at least 66%.

5.2 Recommendations

- **Use more classifiers and graph the difference with decision tree classifier** to increase accurateness of the program. Since the program only used decision tree classifier, future programs related to this dataset can use other methods like Neural Networks, Regression, etc. By using other methods together with the decision tree classifier, the researchers would be more confident with the result and the accuracy of it.

- **Test with more hyperparameters.** This program's main hyperparameter was the data split of the set. Future researchers should find more hyperparameter and tinker with it in order for the accuracy score to increase. With more hyperparameters, the accuracy score can range from 70-80% which is a huge improvement compared to this current program.
- **Dataset with more in-depth features.** All the team data such as the teams' possession, number of crosses, number of tackles, number of headers, etc., could be used to classify each team into different categories of playing styles. Models could then be built to understand the interaction between the playing styles of two different teams and help predict the outcome of a match where these two teams face each other.
- **Implement Machine Learning with other labels.** Instead of just focusing on the main label or home winner, the program should consider other labels of the dataset too such as the percentage of winning the game.
- **Retrieve more datasets connected with this current dataset.** This dataset did not consider the rankings of the teams or how good they were. For example, Germany is one of the top five teams in the world and they are competing against Iran, which is not even in the top 30. Statistically, Germany would win the match regardless of the home court advantage. Rankings and other features should be considered to improve the accuracy of the Machine Learning.

References

1. Kerr, J. H., & van Schaik, P. (1995). Effects of game venue and outcome on psychological mood states in rugby. Personality and Individual Differences, 19(3), 407–410. https://doi.org/10.1016/0191-8869(95)00044-7
2. Courneya, K. S., & Carron, A. V. (1992). The home advantage in sport competitions: a literature review. Journal of Sport & Exercise Psychology, 14(1).
3. Dowie, J. (1982). Why Spain should win the world cup. New Scientist, 94(1309), 693-695.
4. Pollard, R. (1986). Home advantage in soccer: A retrospective analysis. Journal of sports sciences, 4(3), 237-248.
5. Herbinet, C. (2018). Predicting Football Results Using Machine Learning Techniques. London: Department of Computing Imperial College of Science, Technology and Medicine. Retrieved from https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1718-ug-projects/Corentin-Herbinet-Using-Machine-Learning-techniques-to-predict-the-outcome-of-profressional-football-matches.pdf
6. Jürisoo, M. (2019, November 21). International football results from 1872 to 2019. Retrieved from https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017.

Appendix A. Contribution of Members

Table 1. Contribution of Members.

Name	Contributions
Cansana, Jose Lorenzo M.	Documentation
Jamalul, Gabriel Rasheed H.	Documentation
Loyola, Leanne Marie C.	Documentation
Marasigan, Giann Jericho Mari F.	Algorithm, Source Code