Prediction of Football Match Results Using Virtual Data

1st Omar Rodrigo Muñoz Gómez
School of Engineering and Sciences
Tecnológico de Monterrey
Estado de México, México
A01169881@tec.mx

2nd Israel Cardenas Pimentel
School of Engineering and Sciences
Tecnológico de Monterrey
Estado de México, México
A01367005@tec.mx

3rd Leonardo Francisco Garrafa Pacheco
School of Engineering and Sciences
Tecnológico de Monterrey
Estado de México, México
A01222528@tec.mx

***Abstract*—With the increasing availability of data from multiple sources in football soccer, accurately predicting the outcome of a match has become a topic of great interest in the research community. However, this is a challenging task since the nature of the sport is associated with lower predictability compared to other sports. Accurately predicting the outcome of a match can help create successful betting strategies and aid the game's actors in better understanding the factors contributing to their success. In this study, we use player ratings from the EA FIFA video game to create a set of features that help predict the winner of a match. We investigate the extent to which videogame data can be used as a reliable source for creating relevant prediction models in football. Additionally, we include other non-videogame features and use a feature selection approach to identify how relevant the video game features can be compared to the non-videogame features considered. Our dataset consists of 917 matches played in 2018 across five different tournaments. We use player ratings at the beginning of 2018 to build and train a classification model that predicts whether local or visitor teams will draw or win the match.**

***Index Terms*—football soccer, match outcome prediction**

I. Introduction

Betting on sporting events, particularly football, has grown in recent years [1], [2]. Accurately predicting match outcomes is of great interest to teams, coaches, and bettors, as it can provide a competitive advantage [3]. However, this task remains challenging due to the complex nature of the game and the many factors influencing results, with prediction accuracies consistently below 60% in the literature [3].

The evolution of technology has increased the opportunities to gather relevant information about sports, creating public and private databases containing data on players, games, insights, teams, and historical records. However, significant costs associated with implementing technology often make these databases challenging to obtain. As a result, public data sets such as the FIFA video game player ratings have the potential to become valuable resources for researchers exploring ways to create better prediction models. Recent studies have demonstrated the potential of video game player rating data for predicting football match outcomes [4].

Through our research, we aim to expand the knowledge in this field by investigating the potential of using video game player ratings to predict football match outcomes. For this work, we use a public data set [5] that contains information about soccer matches from a time window from 2016 to 2018 (we focus our analysis on matches from 2018) and the player ratings in the EA FIFA video game scrapped from the SoFIFA website. Using these ratings, we create custom statistical features representing the skill level of teams involved in different soccer leagues such as La Liga, Serie A, Bundesliga, Premier League, and Ligue 1 for 917 matches occurring in the year 2018. The ratings will be considered for the different groups within a team (i.e., defenders, midfielders, forwards). We will also consider the influence of the number of players from each role to capture what could be understood as team formation. Predicting the outcome of a soccer match can be more challenging than predicting the outcome of other sports [3]. As a result, correct data and feature selection are crucial for better understanding the data and producing feasible predictions. In this work, we apply data exploratory methods to understand the data and select the best features for our predictive model.

Matches outcome prediction is an important application of machine learning techniques in this field, allowing researchers, companies, and teams to create betting strategies, learn team strategies and their ways to play, understand and formulate player compositions, and of course, measure the possible scenarios of a game given two teams. By studying the ability to improve prediction models from video game data, we aim to provide insights into the use of video game data as a reliable source of information and increase the understanding of the factors that influence the outcome of a match, leading to better match prediction models.

II. Related Work

When studying match prediction in football, studies can be mainly separated into those that try to predict the outcome of the match and those that attempt to predict the final scores [3]. Among those that attempt to predict the outcome of a match, we can distinguish between those that predict three classes (win, loss, draw) and those that choose to discard tied matches to simplify the prediction process. In [3], a study of the accuracy of multiple works was performed and it was found that when attempting to solve the three class problem, the accuracy results in the literature are consistently below 60%.

For example, the authors of [3] created different machine learning models to address the three class problem in 1,140 matches of the English Premier League for three seasons 2019-2021. The authors proposed to categorize matches according to their predictive difficulty based on the Kelly index. They found that when looking only at Type 1 matches, the best classifier achieved an accuracy of 70%, which was found to be higher than the accuracy of the classifier when all matches are considered (51.9%). Furthermore, they found that different attributes are useful for predicting matches in the different match groups. For Type 1 matches, they found that the most influential features are based on the team's historical match statistics. For Type 2 matches, historical data appeared less important and was replaced by odds data provided by bookmakers before the match. Finally, for Type 3 matches, they found that historical data appear inconsequential and that odds still have the advantage in predicting these matches. In this work, 52 features were generated, mostly consisting of historical statistical data for both teams.

The authors of [6] tackle prediction problem from a statistical overview, proposing a Bayesian dynamic generalized linear model to estimate the time-dependent skills of all teams in a league, and to predict the outcome of soccer matches in a time window, and using a simplification of the variables maintaining simpler as possible only describing the main events presented on matches a storing the in a subset of properties. As an outcome of this work, inference was performed and a comparison between bet and matches result was performed, simulating the betting process depending of the teams and matches, producing remarkable results.

In [7], it was proposed a soccer domain knowledge developing two feature engineering methods for match outcome prediction in order to construct two learning sets. The results of this work showed that the implementation of these two methods in the learning datasets outperform the metrics compared with others in the literature when it is applied in k-NN and XGBoost. This work is important because it shows how a goal can contain crucial information about a current game, and it is possible to construct a predictive model just using the goal information from both teams in a game alone producing good results and again, simplifying the data used to construct the predictive model.

Furthermore, in [4], the authors created two prediction models for 380 games of the Spanish first-division league. The first model used historical statistical data (e.g., red cards, home team shots on target). The second model was named a virtual predictor as it used player ratings from the EA FIFA 2015 video game to predict match outcomes. The authors showed that the real and virtual predictors had similar performance on the testing data, showing the potential to use data collected from video games to solve real-world problems. Although this work's limitations are the reduced amount of data and the that it all belongs to a single tournament, limiting the ability to validate player ratings at a broader scale. Therefore, validating this idea on a larger and more varied data set becomes interesting. Furthermore, different feature engineering approaches exist when exploiting the player rating data, and further exploration of these is interesting to determine the best ways to exploit this data but also to validate the data throughout different years.

III. Data and Methods

In order to build a model capable of predicting a match outcome given two teams, it will be important to follow a methodology based on the understanding of the problem and the data, the data preparation, the building of the desired model, the evaluation, and the deployment.

A. Business understanding

Football soccer games are complex because of the variables involved in a single game, and several analyses can be performed depending on the selected feature and the time window chosen. For this work, predicting of match outcome is the desired target, and therefore all the features that will be selected must be correlated with this goal.

B. Data understanding

We chose two datasets for our study: (i) a dataset from [5] containing information about soccer matches played in seven different competitions from 2016 to 2018; and (ii) the database scrapped from the SoFIFA website [8], which contains information on player ratings given at the beginning of 2018. We aim to combine these two datasets by engineering our features for the players involved in each match.

Since the player rating dataset only includes ratings from 2018, we removed the match information from 2017 and 2016. This is because we believe that for an accurate validation of the use of the player ratings for prediction, we need a match between the year the rating was given, and the year the game took place. Therefore, it would be inaccurate to use a player's perceived skill in 2018 to predict games from previous years.

C. Data preparation

We decided to create our data set containing the raking information of the teams and the result of matches during 2018. To that end, we obtained information on each player, their positions, and the result of each match. Then, we calculated each team's ranking information by averaging their players' overall ranking by their positions obtained from SoFIFA's database.

To capture the local team advantage, we created separate features for the local team and the visiting team. This design decision allows our model to exploit the locality condition, often considered an advantage. Furthermore, to extract locality information, we used the matches dataset. Next, we calculated each team's average rating of the goalkeeper, defensive, midfielder, and offensive players separately. This choice helps prevent smoothing out ratings of different team roles. For instance, if a team has a great offense but not such a strong defense, the average of the entire team could hide this relationship from the model. By separating the skill of specific groups of players within each team based on their local condition, we can better capture the skill differences between teams. We also

included the number of defensive, midfielder, and offensive players to see if the predefined team formations could affect the game's outcome. Additionally, these numbers give us an idea of how much information was used to compute the average ratings for each team group. The features generated for each team are shown in Table I.

TABLE I: Description of the set of features generated for both the local and visit teams separately.

Attribute	Range	Description
OVR_GK	[0-100]	Overall rating of the goalkeeper.
AVG_OVR_DF	[0-100]	Average overall rating of defensive team players.
AVG_OVR_MD	[0-100]	Average overall rating of midfielder team players.
AVG_OVR_FW	[0-100]	Average overall rating of offensive team players.
NUM_DF	[2-6]	Number of defensive players for the team at match start.
NUM_MD	[1-7]	Number of midfielder players for the team at match start.
NUM_FW	[0-4]	Number of offensive players for the team at match start.

In the next sections, we aim to understand the generated data set to get insights into the limitations of our data and to get an estimate of the effort that we would need in order to meet our prediction objective based on this data.

D. Data validation

Our first step is to determine the amount of available data, attributes, data types, and missing values. Our resulting data set comprises information about 917 matches, 14 attributes, and the target variable indicating the outcome of a given match. All attributes are numerical type, and the target variable is of text type with three possible classes in the data set: *LOCAL*, *VISIT*, and *DRAW*, which indicates the local team won, the visitor team won, and it was a draw, respectively. The attributes shown in Table I are included for both the local and visit teams.

We observe that three of the attributes contain missing values (23, 2, and 25 values for *L_AVG_OVR_FW*, *V_AVG_OVR_GK*, and *V_AVG_OVR_FW*, respectively); these missing values are likely due to the missing rating information for players participating in a given match. Assuming the missing data does not share the same observation, the total number of observations with missing values represents roughly 5.4% (i.e., (23+2+25)/917 $\approx$ 0.054) of the data, which is a small percentage. We drop observations with missing data to avoid guessing the missing values and because the number of missing data is small. This resulted in 50 observations being removed, meaning the missing values were from different observations, as assumed. Finally, after removing the missing values, our data set contains information on 867 matches.

E. Statistical analysis

In this section, we perform a study of the distribution of the different attributes and classes, and a study of the correlation between the attributes.

1) Attribute distribution: Figure 1 shows a comparison of the values for all the rating variables for both the local and visit teams using boxplots. We observe that GK and FW attributes have more dispersed data than DF and MD. Additionally, they also seem to have more ranking, however, the GK has a more average ranking than the others. Furthermore, 50% of the data of each feature is contained in a range of less than 10 of ranking (from Q1 and Q3 of *V_AVG_OVR_GK*).

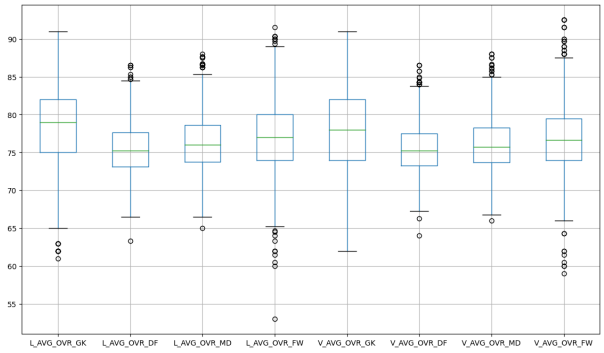


Fig. 1: Distribution of the average player ratings for different player groups for local and visit teams.

In Figure 2, we show a comparison of the number of players by defensive, midfielder, and offensive positions for both the local and visit teams using boxplots. In this case, we observe that the DF tends to be four players at least 50% of the time. On the other hand, MD and FW tend to be 3 or 4 players and 2 and 3 players, respectively, at least 50% of the time.

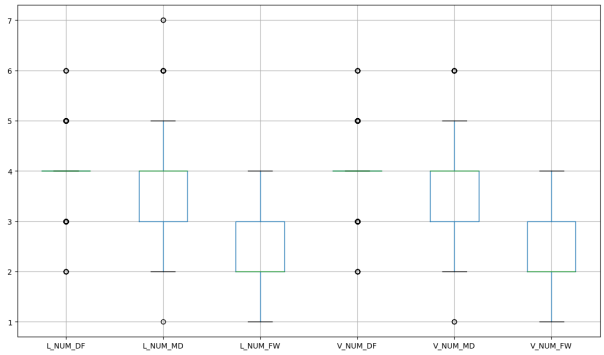


Fig. 2: Distribution of the number of players for local and visit teams.

2) Class distribution: Notice that since a football match has three possible outcomes (local team win, visit team win, draw), the probability of getting the right prediction by chance is 1/3 [2]. However, from the exploration of the 917 matches in our generated data set, we found that 25.41% of the matches in the data set end in a draw, 45.37% of the matches are won by the local team, and 29.22% are won by the visiting team. We observe that the classes are imbalanced. This observation is in accordance with what is known as local team advantage in the field of football. We also observe that a naive model that always predicts that the local team is the winner could be expected to have relatively good behavior (considering that accuracy results in the literature are usually below 60%). Because of this and considering how challenging the problem can be, we consider that a good performance target for our classifier would be to be above the naive classifier.

3) Correlation: Furthermore, we created the correlation plot from Figure 3 to better understand the relationships between our attributes. From here, we can observe that teams often have similar characteristics across their various groups. For example, we notice that teams with strong defense also tend to have strong midfield and offensive groups. Additionally, we found a negative correlation between the number of midfielders and the number of offensive players. This is to be expected since there is a maximum of 11 players allowed on the field at any given time.

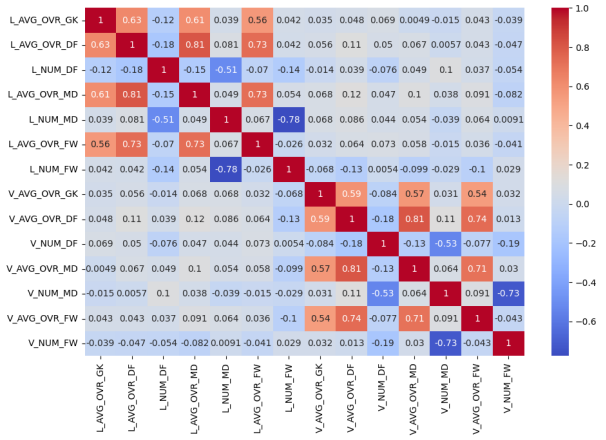


Fig. 3: Matrix correlation between attributes.

To illustrate the correlation between the variables, Figure 4 shows the positive correlation between the overall rating of defensive team players and the overall rating of midfielders for the local team.

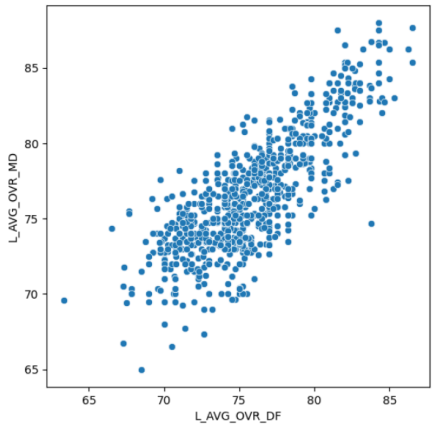
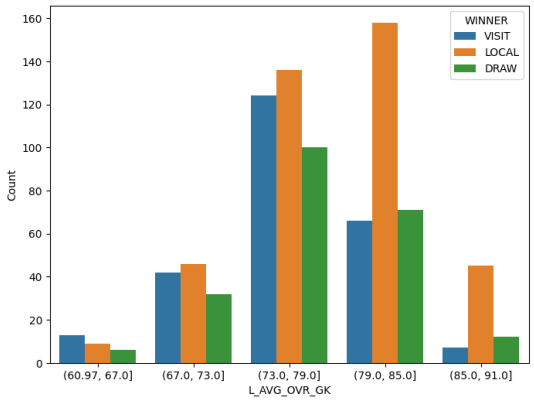
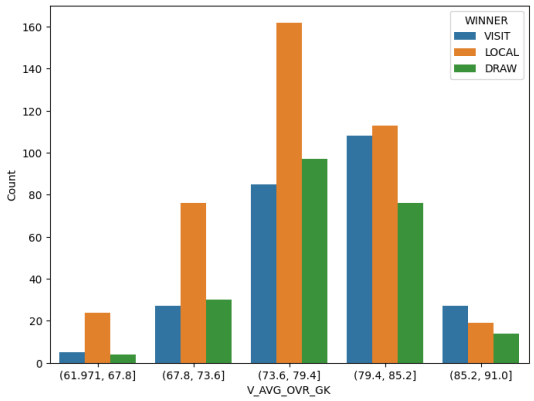


Fig. 4: Correlation between Average overall rating of defensive team players and Average overall rating of midfielder team players on local team.

4) Validating assumptions: During our exploration stage, we sought to validate our assumption that a high player rating for a given team would correlate with a higher number of matches won. To do so, we categorized the ratings into discrete bins and examined the distribution of classes in each bin. Our analysis suggests that high ratings for the local team serve as a better indicator of that team's success, whereas high ratings for the visiting team do not seem to be as strongly related to the team's success. To illustrate this behavior, we have included Figure 5, which presents a comparison of the goalkeeper ratings for both the local and visiting teams.



(a) Local team goalkeeper rating.



(b) Visit team goalkeeper rating.

Fig. 5: Distribution of goalkeeper player ratings for both the local and visiting teams, organized by class.

F. Feature selection

In order to identify which features have the greatest ability to differentiate between different classes, we employed a feature selection algorithm based on the Mean Decrease in Impurity (MDI) of a Random Forest model, which was designed to predict the target variable. This algorithm works by calculating the mean and standard deviation of the impurity decrease for each attribute used within the tree [9]. By ranking the variables by their MDI, we are able to determine which features are the most important. In Figure 6, we present the MDI for our features, revealing that the rating variables have a greater impact than the attributes indicating the number of players in each group. Among the attribute variables, the average rating of the visiting team's midfielders appears to be the most relevant.

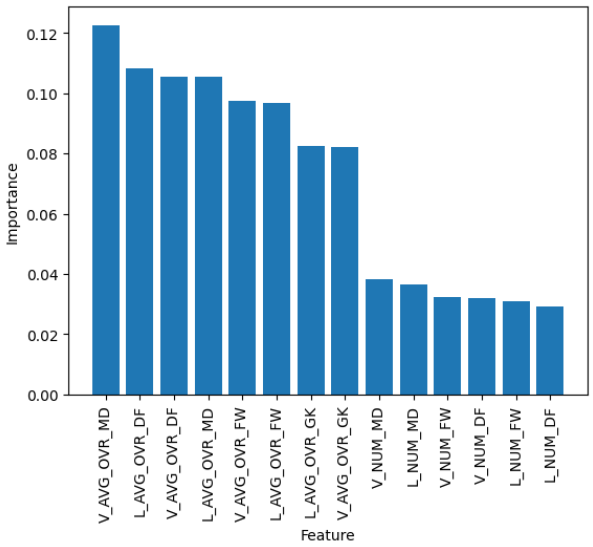


Fig. 6: Feature selection based on MDI.

G. Modeling

Our problem is that of predicting if the match will end in either a win for the local team, the visit team, or a tie. We propose to make a comparison between three different classification models. A Random Forest model, a Support Vector Machine and a Neural Network. We will begin by separating the complete data set into training and testing portions. Next, we will perform a comparison of the models using a 10-fold cross validation approach considering parameter tuning for each of the proposed models and a feature selection approach as discussed in the previous section.

H. Evaluation

To evaluate the model, metrics such as F1 score, ROC/AUC metrics, and confusion matrix, among others, will be used to understand the performance of the generated models. Furhtermore, to assess the potential of the created data set to improve match outcome predictions, the performance will be compared to a completely random guess of the outcome (with an accuracy of 1/3), and to the prediction of a naive classification approach that always predicts the local team will win. A complete explanation of this will be covered in the Results section.

IV. Results

V. Discussion

VI. Conclusion

References

[1] G. Gaub, "Prediction of english premier league soccer matches, based on player data using supervised learning," Master's Thesis, Leopold-Franzens-Universität Innsbruck, Innsbruck, Austria, December 2022.

[2] F. Rodrigues and A. Pinto, "Prediction of football match results with machine learning," *Procedia Computer Science*, vol. 204, pp. 463–470, 09 2022.

[3] Y. Ren and T. Susnjak, "Predicting football match outcomes with explainable machine learning and the kelly index," 2022.

[4] J. gon Shin and R. Gasparyan, "A novel way to soccer match prediction," 2014.

[5] L. Pappalardo, P. Cintia, A. Rossi, E. Massucco, P. Ferragina, D. Pedreschi, and F. Giannotti, "A public data set of spatio-temporal match events in soccer competitions," *Scientific Data*, vol. 6, no. 1, p. 236, 2019. [Online]. Available: https://doi.org/10.1038/s41597-019-0247-7

[6] H. Rue and O. Salvesen, "Predicting and retrospective analysis of soccer matches in a league," 01 1997.

[7] D. Berrar, P. Lopes, and W. Dubitzky, "Incorporating domain knowledge in machine learning for soccer outcome prediction," *Procedia Computer Science*, vol. 108, p. 126, 01 2019.

[8] "Players FIFA 23 Apr 6, 2023 SoFIFA." [Online]. Available: https://sofifa.com/

[9] Scikit-learn, "Feature importances with a forest of trees," https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html, 2023, accessed on: April 8, 2023.