Predictive Value of Football Players Using Comparison of Decision Tree and Random Forest Algorithms

Kelly Mae
Major on Information Systems, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang, Indonesia
kelly.mae@student.umn.ac.id

***Abstract* — Classification is one of algorithms used in data mining which sets a variable into a group of variables into the target variable in order to do the prediction accurately on every case in the data. The research object is the value of the worldwide football players that is not always determined based on accurate supported data related to the players. The purpose of this research is to create a model to know the predictive football players value using the Decision Tree algorithm based on the football players facts. The data that will be used is secondary data that contains football players on the basis of a game called FIFA 21. The research outcome states that the data is appropriate to be analyzed using Decision Tree and Random Forest models. The result of research can hopefully be the reference in deciding the value of football players that is appropriate with their skills and abilities to prevent obtaining a lack of value in the market.**

Index Terms — ***Classification, Data Mining, Decision Tree, Random Forest, R, Football Players, Value***

I. INTRODUCTION

Football is the most famous sport that stands out among other sports and games which attracts millions of watchers worldwide. As a result, numerous news related to football have been published because of the high popularity and vast markets, including football players' transfer and their value. Football can be categorized as a low scoring game where the goal scoring event may occur rarely. Thus, the final scores for the match are unable to give an obvious view of the teams' technical and physical performances. Football has several chance components; however, this does not determine the luck rate of the successful teams compared to others [1].

There are few crucial factors in football that have to be looked at. The football players hold important roles in football. Every football team is required to have players with good performances in order to win a tournament that has been seen based on the market value. For the club, the market value of a player is a score that keeps increasing based on various components, such as talent, popularity, skill, playing style, efficiency, and so on. Market value can be explained as an indication for the comparison and economic valuation of various football players that shows their values and also performances [2]. It is also worth mentioning that all player positions like goalkeeper, defender, midfielder, and forward, have their own responsibilities that guide the indication of performances by position. In this case, players that play as forward are more visible for the watchers to be seen rather than other positions because of the characteristic of football which is goal-oriented [3]. To add, having the players acquired through transfer can be also stated as another important factor in the football business.

There may be several problems that some football clubs must face, involving recruiting and acquiring players in costly value. Consequently, most clubs have to waste funds without paying attention to the players' specifications and only depending on their popularity which results in getting defeated in the tournament. Football clubs could not take part in tournaments without having a proper team, the players hold a key value factor of a football clubs which lead to potential in forming and raising their value [4].

To take evidence, there are many factors including buying football players' value that may bring a downturn effect on some of European clubs who are in poor financial condition. There are just below 10% of European top division clubs that declared surplus losses of €10 million in the year 2017. According to the previous UEFA Club Licensing Benchmark Report, the clubs would possibly face bankruptcy in the foreseeable future if they do not have any booster in capital from outside. In 2018, the players' selling process was being combined with the transfer fee of €5.4 billion, however it was valued at just €10 billion at the time of their sale in the market. Therefore, the net book value and original transfer prices of the top 20 football clubs have gone up by more than 60% since 2015 that can be considered as huge increases in transfer costs [5]. This might result in some clubs with high wages to gain low revenue and transfer profits that may end in facing serious financial problems.

Hence, the problem statement of this research is creating prediction models about football players' value based on their specifications using the Decision Tree algorithm using Party model, followed by Random Forest algorithm. The research also focuses on comparing two models to consider which model is suitable for the real-life implementation.

II. Literature Review

A. Introduction to Decision Tree and Random Forest Algorithm

Decision Tree can be defined as a tree in which a test of an attribute and leaf node is being shown by every node in order to give classification. In this case, the classification of the test example begins at the root node that tests feature values per node and sorts down to the suitable branch until it reaches classification through the leaf node [6].

Random Forest is one of ensemble methods which uses a set of CARTs algorithm for prediction. Random Forest involves building a subset of training data through the replacement process by doing the "Bagging" method. In this case, the data with the same value can be chosen for multiple times, while another data may not be chosen at all [7].

B. History of Decision Tree and Random Tree Algorithm

There are several decision tree algorithms used for classification, such as The C4.5, ID3, and C5.0 decision trees. To begin with, ID3 or Iterative Dichotomiser 3 is a simple decision tree algorithm which was introduced in 1986 by Quinlan Ross. The ID3 algorithm has the purpose of building the decision tree by using a top-down, followed by greedy search mode through the given test sets for attributes in each tree node. There is also the C4.5 algorithm, an extension of Quinlan's previous ID3 algorithm. C4.5 algorithm is used to produce decision trees, applying categorical and numerical data for classification needs [8]. Lastly, the C5.0 algorithm is an improvement from the C4.5 algorithm which produces classifiers proved as decision trees or rule sets with improved features.

Random Forest was formulated by Leo Breiman and Adele Cutler whereas Random Forest had been stated as their hallmark. The term of the Random Forest algorithm was firstly formed by Tin Kam Ho of Bell Labs in the year 1995 which united with Breiman's "Bagging" method idea and the random selection features that were presented independently by Ho with help from Amit and German. As a result, Random Forest technique is able to build a group of decision trees with some alterations that are well-controlled [9].

C. Usages of Decision Tree and Random Forest Algorithm

There are few usages of Decision Tree algorithm. Decision Tree can divide some datasets into separated classes. Decision Tree is used for all kind of target variables; however, it is mainly used for ones that in form of categorical. Decision Tree involves using information gain method to do data splitting and Calculate Entropy or Gini Index method in data splitting to discover homogeneity in the dataset [10]. Additionally, Decision Tree is able to predict the result for future reports. As a result, Decision Tree is one of the most effective methods for data mining.

In comparison, there are several usages of the Random Forest algorithm. Random Forest has the "Bagging" method that is more generalized and has the ability to integrate diverse feature types. Random Forest has the majority voting approach that can minimize the existence of misclassifications effectively [11]. Random Forest is able to determine the rank of related significance for each predictor which is based on the regression predictor error at the Out-Of-Bag or OOB [12].

D. Advantages and Disadvantages of Decision Tree and Random Forest Algorithm

The Decision Tree algorithm has the advantage in classifying unknown records in a swift manner. Decision Tree is great in the presence of redundant attributes and slightly firm in the presence of noise if overfitting methods are given. However, Decision Tree has the disadvantage that not applicable data have a bad effect in the construction of the decision tree. On this occasion, any small changes in the data may affect the whole appearance of the decision tree. Moreover, a sub-tree in the decision tree can be produced in many amounts [13].

The Random Forest algorithm has some advantages in capturing elaborate interactions among features used by having knowledge of a nonlinear combination of them. Random Forest is able to operate evenly well with continuous, discrete, or missing values with slight or no modification. Random Forest is relatively robust to outliers, such as noisy examples and noisy labels. Nevertheless, Random Forest has the disadvantage in which an actionable scheme is indirectly evident given an example. Consequently, the model outcome is indirectly human-interpretable [14].

III. Methodology

A. Research Objects

This research uses worldwide football players as the objects who are still actively working in their occupations. Most football players come from

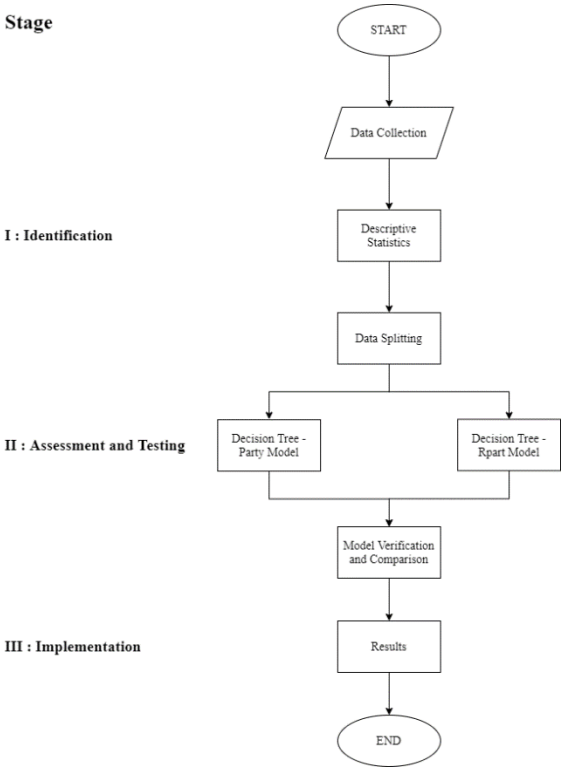
different countries and football clubs. Additionally, the football players are divided into several categories based on football skills.

B. Data Collection

This research involves quantitative data, which is a dataset of football player specifications. In order to collect the data, researcher use the secondary data which can be achieved from the data collection website, Kaggle.com.

The collected dataset is valid to be used because of its detail and complete data values. The dataset consists of 107 variables and 18.944 rows which makes it complex to be organized in a large volume.

C. Framework



D. Identification

The identification process requires an observation on the data structures. On this occasion, the data will be used in a form of Excel with the extension .xlsx. The data will be inputted and processed using R. During data processing, some packages will be used based on the analysis purpose for Decision Tree and Random Forest.

The first step of research consists of an identification process. Next, there will be an observation process whereas some independent data will be observed and then filtered for determining which data should be used. In this case, the numerical data from the dataset will be taken for the next process. The next process involves viewing data distribution that is suitable to be the target variable by looking at the target variable plot in numerical form. To add, the target variable will be a discrete variable for the chosen variables. Then, the chosen numerical variable will be changed into categorical for data splitting based on median value. For the further validation, there will be data splitting for the needs of training data and testing data for both models, which are Decision Tree and Random Forest.

E. Assessment and Testing

After going through data identification and validation, there will be a process of making prediction models with training data. By using a discrete variable as the target variable, this variable will be connected with several independent variables in the form of numeric. In this process, the decision tree algorithm will make nodes and their leaves based on the particular condition that is appropriate with the classification. On the other hand, the random forest algorithm will create a forest with several trees with the condition for the classification whereas the higher the number of trees in the forest, then greater is the accuracy of the results. Besides the level of accuracy, the level of sensitivity and specificity can be used for the comparison between two models.

F. Implementation

The implementation process contains comparison in the results of prediction with the testing data that have been made on the previous process. The comparison will be resulted in the form of confusion matrices and their statistics to see how accurately the models are predicted.

G. Results Validation

The results validation method comprehends inspecting the accuracy in each model from the confusion matrices to decide whether the models can be used for the prediction or not. In this case, accuracy is defined as the case percentage that is precisely identified. The level of accuracy above 50% marks that the model can be used with the higher accuracy rather than doing random guesses. The model with higher accuracy than the other determines which model is better to be applied in the real-life prediction.

IV. Results and Discussion

A. Identification

The data that will be used is the secondary data which can be achieved from the data collection website, Kaggle.com. The current data used is a complete data of worldwide player specifications.

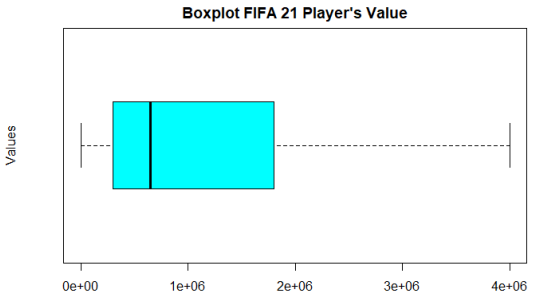
The data will be stored in the form .xlsx named D1_KellyMae_51428.xls. There are 18.994 rows and 107 variables, with thousand amounts of NA value.

The data comprises general data, such as id, name, age, photo, nationality, club name, value, wage, international reputation, and also specific skills like position, pace, shooting, passing, dribbling, defending, physic, and more variables. Then, the data will be inserted, observed, processed and at last resulted in visualization using R.

Here are variables that will be used for this research that entails data filtering, splitting, value removing, and target variable creation for the discrete variable and some diagrams for data visualization.

```
'data.frame':   16861 obs. of  7 variables:
 $ value    : Factor w/ 2 levels "< 650000",">= 650000": 2 2 2 2 2 2 2 2 2 2 ...
 $ pace     : num  85 89 78 91 76 96 76 94 93 78 ...
 $ shooting : num  92 93 91 85 86 86 60 85 86 90 ...
 $ passing  : num  91 81 78 86 93 78 71 80 81 77 ...
 $ dribbling: num  95 89 85 94 88 91 71 90 90 88 ...
 $ defending: num  38 35 43 36 64 39 91 44 45 33 ...
 $ physic   : num  65 77 82 59 78 76 86 76 75 73 ...
 - attr(*, "na.action")= 'omit' Named int [1:2083] 3 8 10 13 17 19 24 37 41 45 ...
  ..- attr(*, "names")= chr [1:2083] "3" "8" "10" "13" ...
```

Image 1 Data Filtering and Splitting




```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
      0   300000   650000  2224813  1800000 105500000
```

Image 2 Boxplot Distributions and Descriptive Statistics on Target Variable

On Image 1, there are 7 numerical variables that are going to be used, containing value, pace, shooting, passing, dribbling, defending, and physic. On Image 2, there is distribution of player's value in the form of boxplot and descriptive statistics in which the median of the target variable shows 650.000 so that the target variable is divided into 2 variables based on the median value.

B. *Assessment and Testing*

After the data identification and validation process, the prediction models will be made using training data. In this case, the target variable in the form of discrete will be combined with some independent numeric variables that have been chosen. The decision tree algorithm will make nodes and their leaves based on a particular condition which is appropriate for the classification. In contrast, the random forest algorithm will create a forest with several trees based on classification state whereas the higher the number of trees in the forest, then greater is the accuracy of the results. Additionally, the level of sensitivity and specificity can be used for the comparison between two models.

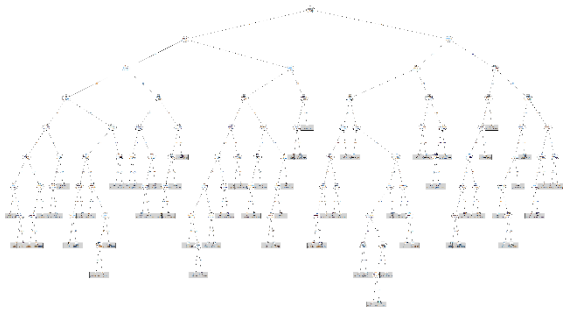


Image 3 Decision Tree with Party Model

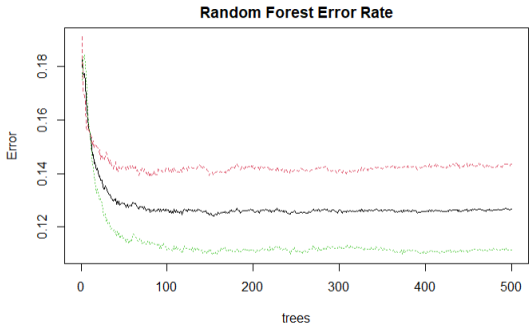


Image 4 Random Forest Error Rate

C. *Implementation*

The implementation process begins with the prediction at model decision trees based on training data using the testing data which results in confusion matrices.

```
Confusion Matrix and Statistics


predict_ctree    0    1
            0 1396  222
            1  229 1526

               Accuracy : 0.8663
                 95% CI : (0.8543, 0.8776)
    No Information Rate : 0.5182
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7322

 Mcnemar's Test P-Value : 0.7775

            Sensitivity : 0.8591
            Specificity : 0.8730
         Pos Pred Value : 0.8628
         Neg Pred Value : 0.8695
             Prevalence : 0.4818
         Detection Rate : 0.4139
   Detection Prevalence : 0.4797
      Balanced Accuracy : 0.8660

       'Positive' Class : 0
```

Image 5 Confusion Matrix and Statistics in Decision Tree

```
Confusion Matrix and Statistics

predict_rf    0    1
         0 1389  165
         1  236 1583

               Accuracy : 0.8811
                 95% CI : (0.8697, 0.8919)
    No Information Rate : 0.5182
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7615

 Mcnemar's Test P-Value : 0.0004729

            Sensitivity : 0.8548
            Specificity : 0.9056
         Pos Pred Value : 0.8938
         Neg Pred Value : 0.8703
             Prevalence : 0.4818
         Detection Rate : 0.4118
   Detection Prevalence : 0.4607
      Balanced Accuracy : 0.8802

       'Positive' Class : 0
```

Image 6 Confusion Matrix and Statistics in Random Forest

D. *Results Validation*

The last phase is the comparison between both models. In this step, the level of accuracy will be observed if both models are qualified for prediction. The level of accuracy above 50% marks that the model can be used with the higher accuracy rather than doing random guesses. The model with higher accuracy than the other determines which model is better to be applied in the real-life prediction.

```
predict_ctree    0    1
            0 1396  222
            1  229 1526
```

Image 7 Prediction with Decision Tree

```
predict_rf    0    1
         0 1389  165
         1  236 1583
```

Image 8 Prediction with Random Forest

E. *Discussion*

From all variables that have been used, it has been selected as numerical variables. After doing data cleansing and visualization, both prediction models with different methods have been made. Both models will use diverse approaches which result in dissimilar levels of accuracy. By using training and testing data, obtained the random forest algorithm model with higher level of accuracy that reaches 88.11%, than decision tree model algorithm which holds 86.63%. As a result, the random forest algorithm can be used to predict in upcoming time.

V. Conclusion

A. *Conclusion*

The two data mining methods will result in distinct levels of accuracy, depending on the approach for each method. Decision Tree uses nodes to illustrate the formed conditions, while Random Forest operates by building a multitude of decision trees during training time and bringing the mode of the classes as an output by the individual trees. Even though each model has their own advantages and disadvantages, the research requires using the same dataset with unlikely levels of accuracy from their predictions. To get an accurate prediction, the greater accuracy model is needed to avoid mistakes in predicting. This research has an outcome in which predictive value of football players with Random Forest model is more accurate so that the selected dataset and target variable are better to use for the prediction in the upcoming time.

B. *Suggestion*

For the purpose of reducing mistakes in recruiting a football player, the use of data analysis is required. By using a complete dataset, this can be observed target variables to determine the value of a football player. In this case, the Random Forest prediction model is the right model to be used in getting insights about the value of the player that will be recruited.

C. *Limitation*

This research requires a clear target variable where every case can use different target variables. As a consequence, the various types of target variable may highly affect the model.

Acknowledgements

The author would like to thank Ir. Raymond Sunardi Oetama, M.CIS, as the lecturer of Data Analysis, study program Information System, Universitas Multimedia Nusantara University for providing time to give advice, guidance, and knowledge during the study.

REFERENCES

[1] H. Lepschy, H. Wäsche and A. Woll, "How to be Successful in Football: A Systematic Review," *The Open Sports Science Journal*, vol. 11, pp. 3-23, 2018.

[2] P. Singh and P. S. Lamba, "Influence of Crowdsourcing, Popularity, and Previous Year Statistics in Market Value Estimation of Football Players," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 22, no. 2, pp. 113-126, 2019.

[3] A. Metelski, "Factors Affecting the Value of Football Players in the Transfer Market," *Journal of Physical Education and Sport*, vol. 21, no. 2, pp. 1150-1155, 2021.

[4] P. Igor, "Salaries to Revenue Ratio Efficiency in Football Clubs in Europe," in *Eurasian Economic Perspectives : Proceedings of the 22nd Eurasia Business and Economics Society Conference*, B. M. Huseyin, D. Hakan, D. Ender and C. Ugur, Eds., Springer, 2019, pp. 303-313.

[5] P. Igor, "Football Clubs Drowned by Players," *Polish Journal of Sport and Tourism*, vol. 27, no. 1, pp. 28-32, 2020.

[6] S. Gavankar and S. Sawarkar, "A Novel EagerDT Complexity Approach to Deal with Missing Values in Decision Trees," *International Journal of Simulation -- Systems, Science & Technolog*, vol. 19, no. 6, pp. 1-5, December 2018.

[7] M. Belgiu and L. Drăguţ, "Random Forest In Remote Sensing: A Review of Applications and Future Directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24-31, 2016.

[8] H. Sharma and S. Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining," *International Journal of Science and Research (IJSR)*, vol. 5, no. 4, pp. 2094-2097, April 2016.

[9] Y. Li, C. Yan, W. Liu and M. Li, "A Principle Component Analysis-based Random Forest with the Potential Nearest Neighbor Method for Automobile Insurance Fraud Identification," *Applied Soft Computing*, vol. 70, pp. 1000-1009, September 2018.

[10] P. Gulati, A. Sharma and M. Gupta, "Theoretical Study of Decision Tree Algorithms to Identify Pivotal Factors for Performance Improvement: A Review," *International Journal of Computer Applications*, vol. 141, no. 14, pp. 19-25, May 2016.

[11] X. Mei, R. Wang, W. Yang, F. Qian, X. Ye, L. Zhu, Q. Chen, B. Han, T. Deyer, J. Zeng, X. Dong, W. Gao and W. Fang, "Predicting Malignancy of Pulmonary Ground-Glass Nodules and their Invasiveness by Random Forest," *Journal of Thoracic Dicease*, vol. 10, no. 1, pp. 458-463, January 2018.

[12] Y. Everingham, J. Sexton, D. Skocaj and G. Inman-Bamber, "Accurate Prediction of Sugarcane Yield using a Random Forest Algorithm," *Agronomy for Sustainable Development*, vol. 36, no. 2, pp. 1-9, 2016.

[13] M. Somvanshi, P. Chavan, S. Tambade and S. V. Shinde, "A Review of Machine Learning Techniques using Decision Tree and Support Vector Machine," in *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, Pune, 2016.

[14] M. Ibrahim, "An Empirical Comparison of Random Forest-Based and Other Learning-To-Rank Algorithms," *Pattern Analysis and Applications*, vol. 23, pp. 1133-1155, 2020.