Predictive Value of Football Players Using Comparison of Linear Regression and Density-Based Clustering Algorithms

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

***Abstract* — Regression is a machine learning technique that is used to fit to an equation and estimates the value of the target variable as a function for the predictor variables for each situation in a dataset. Clustering can be defined as an unsupervised machine learning that divides data into groups of objects based on correspondence measurement. 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 Linear Regression and Density-Based Clustering 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 — ***Regression, Clustering, Machine Learning, Linear Regression, Density-Based Clustering, Prediction R, Football Players***

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 Linear Regression algorithm, followed by Density-Based Clustering 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 Linear Regression and Density-Based Clustering Algorithms*

Linear Regression can be defined as one of the easiest and most ordinary machine learning algorithms which is a mathematical method in achieving predictive analysis. In this case, linear regression is used for evaluations and measuring that determines the linear relationships between independent variables and dependent variables. Therefore, this method is able to model these linear relationships that have been examined from the analysis and finding out the current training results [6].

Density-Based Clustering is one of clustering algorithms that is able to find clusters with arbitrary shape and is insensitive to noise. Density-Based Clustering entails the process of forming clusters with the combination of separated dense areas by the field of spared regions. In this case, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) has the purpose of creating large spatial clustering with noise or outliers [7].

B. *History of Linear Regression and Density-Based Clustering Algorithms*

Linear Regression was born from Francis Galton investigations into the laws of heredity. To begin with, Galton's 1877 paper "Typical Laws of Heredity" met a problem that he had to deal with for a while since his 1865 book "Hereditary Genius". On this occasion, Galton discovered a phenomenon called reversion which is attracting to the process of an inheritance that resulted in parents' progeny who considerably deviated from the mean trait to revert back to the trait and experienced a less deviation from their parents. As a result, Galton came up with a new idea in explaining the outcome of the inheritance process from his 1877 presentation, which is regression towards mediocrity. Additionally, Galton accomplished in knowing that this phenomenon has come in a statistical way which is now recognized as "regression to the mean" [8].

The history of the Density-Based Clustering Algorithm begins in the year of 1996 when Martin Ester proposed DBSCAN to identify overall clusters by searching all core points and applying expansion on all density-reachable points in a systematic approach. However, DBSCAN has some drawbacks that bring out concerns in which one of them is having no ability to discover clusters of varying density. For this reason, some of the DBSCAN creators made OPTICS (Ordering Points to Identify Clustering Structure) which is an augmented ordering algorithm that derives the results of flat or hierarchical clustering. DBSCAN has become one of the most popular Density-Based Algorithms that gained the SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) 2014's Test of Time Award. To summarize, DBSCAN is the most Density-Based Clustering Algorithm used in the scientific community nowadays [9].

C. *Usages of Linear Regression and Density-Based Clustering Algorithms*

There are few usages of Linear Regression Algorithm. First, Linear Regression attempts to detail complex, deficient understanding real-life procedures in the simplest and most precise mathematical terms that are possible. Simultaneously, the construction in mathematical method is a model that tends to be used in a real-life relationship [10]. Second, Linear Regression is used to establish the relationship between predictor or independent variables and response or dependent variables in order to seek an estimation. Lastly, Linear Regression employs a straight-line model in which the mathematical equation is familiar to help separating indications from the noise occurring in the graphs.

On the contrary, Density-Based Clustering has some usages in data mining. Density-Based Clustering finds clusters with various granulite levels with the proper filtered noise or outliers, depending on the thought to seek the density of an area. As a result, Density-Based Clustering provides separation of dense regions from the noise. Besides, Density-Based Clustering uses two parameters, such as epsilon (ε) and Minpts in classifying each point as either core or border point [11].

D. *Advantages and Disadvantages of Linear Regression and Density-Based Clustering Algorithms*

Linear Regression has the advantage in carrying out a statistical model that shows relationships between independent variables and dependent variables that are almost linear are able to display optimal outcomes. Linear Regression is frequently applied to non-linear relationships; however, it is beneficial for data with linear implementations that

is an acceptable first-order estimation. On the other hand, Linear Regression has the disadvantage that is limitation in predicting numerical results. Linear Regression is insufficient in clarifying what has been discovered which can be a serious issue. Furthermore, Linear Regression does not fit properly with continuous or binary data [12].

There are several advantages while using the Density-Based Clustering algorithm. Density-Based is famously known because of the ability in detecting random shaped clusters without any initial group information presented in the dataset. Density-Based Clustering is capable of getting high density areas as the reasonable clusters to guarantee that the density is symbolized as the number of objects occurring in the neighborhood, surpassing particular thresholds. Nevertheless, the disadvantages of Density-Based Clustering are the lack of performance in the clustering process because of having limited parameters so that it might be difficult to have an estimated relevant value for diverse types of dataset without any sufficient preceding knowledge. Density-Based Clustering has a high level of complexity when working with large and high dimensional datasets which may result in scalability issues. Moreover, Density-Based Clustering is sensitive in the process of inputting data by order. In the end, having diverse data orderings may lead to various consequences [13].

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*

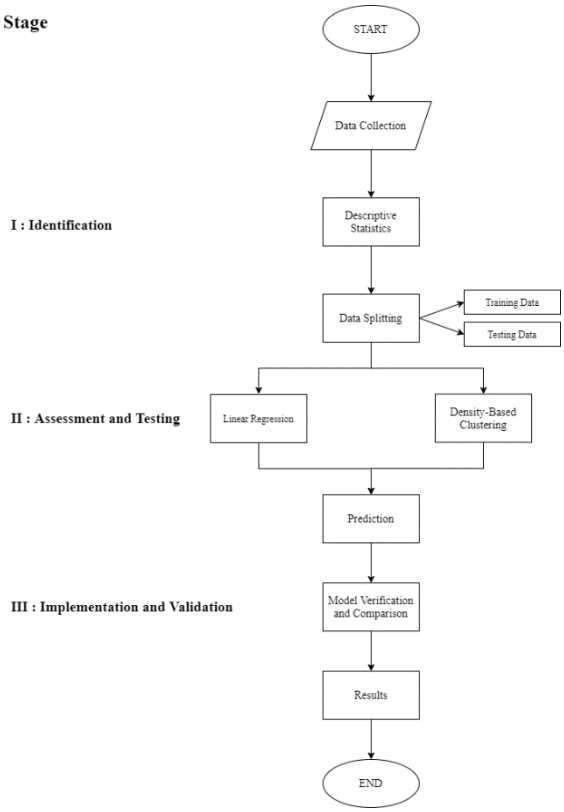

Image 1 Research Framework

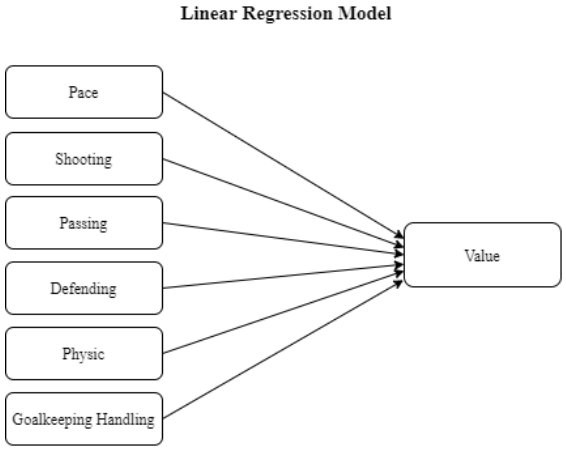

Image 2 Linear Regression Model

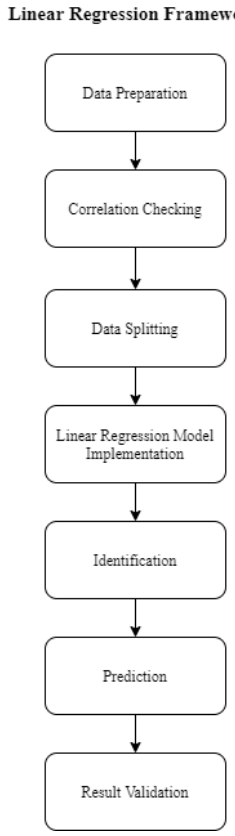

Image 3 Linear Regression Framework

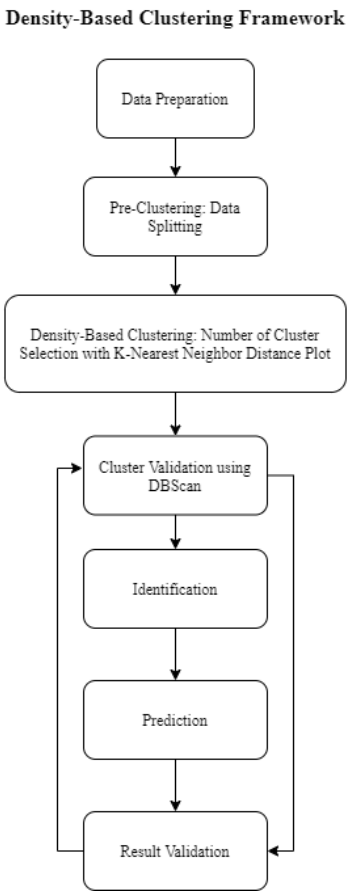

Image 4 Density-Based Clustering Framework

D. Identification

The identification process requires an observation on the data structures. On this occasion, the data will be used in a form of Microsoft 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 Linear Regression and Density-Based Clustering.

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 Linear Regression and Density-Based Clustering.

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. On this occasion, there will be a correlation checking process in the Linear Regression algorithm. To add, correlation is crucial in finding out relationships between two quantitative variables. In this process, Linear Regression model will be made with existing independent and dependent variables in order to understand how close the data has fitted into the model by looking from adjusted R-squared value. In addition, there will be model identification through regression diagnostic plot and other tests, such as Durbin-Watson Test, VIF (Variance Inflation Factor) Test, and Anderson-Darling Normality Test.

In contrast, the Density-Based Clustering algorithm will discover clusters in the form of arbitrary shape based on the calculation of the number of clusters with suitable epsilon (ε) and minPts in K-Nearest Neighbor Distances. Density-Based Clustering will validate its clusters by implementing DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method. This method will result in several significant clusters with their equivalent densities. By having minimum features in step with a cluster, this may result in significant clusters being divided into smaller clusters. Furthermore, Density-Based Clustering is able to detect outliers from the existing clusters.

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 RMSE (Rooted Mean Square Error) value for Linear Regression in order to know the difference between the actual and the predicted values, and confusion matrix for Density-Based Clustering with its statistics to see how accurately the models are predicted.

G. *Results Validation*

The results validation method comprehends inspecting the estimation of Linear Regression from the RMSE value and the confusion matrix from Density-Based Clustering to decide whether the models can be used for the prediction or not. In this case, Linear Regression model can be considered good if RMSE value is low and adjusted R-squared value is near to 1. Concurrently, accuracy is defined as the case percentage that is precisely identified. The level of accuracy above 50% from Density-Based Clustering model marks that it can be used with higher accuracy rather than doing random guesses. The model with higher accuracy with a good fit and slight error 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, defending, physic, goalkeeping handling, 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 graphs for the data visualizations.

```
'data.frame':   16861 obs. of  7 variables:
 $ value              : num  1 1 1 1 1 1 1 1 1 1 ...
 $ 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 ...
 $ defending          : num  38 35 43 36 64 39 91 44 45 33 ...
 $ physic             : num  65 77 82 59 78 76 86 76 75 73 ...
 $ goalkeeping_handling: num  11 11 6 9 13 5 10 10 14 15 ...
 - 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 5 Data Filtering and Splitting in Linear Regression

```
tibble [2,000 x 7] (S3: tbl_df/tbl/data.frame)
 $ value    : num [1:2000] 1 1 1 1 1 1 1 1 1 1 ...
 $ pace     : num [1:2000] 85 89 78 91 76 96 76 94 93 78 ...
 $ shooting : num [1:2000] 92 93 91 85 86 86 60 85 86 90 ...
 $ passing  : num [1:2000] 91 81 78 86 93 78 71 80 81 77 ...
 $ dribbling: num [1:2000] 95 89 85 94 88 91 71 90 90 88 ...
 $ defending: num [1:2000] 38 35 43 36 64 39 91 44 45 33 ...
 $ physic   : num [1:2000] 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 6 Data Filtering and Splitting in Density-Based Clustering

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

Image 7 Descriptive Statistics on Target Variable

On Image 5, there are 7 numerical variables that are going to be used in Linear Regression, containing value, pace, shooting, passing, defending, physic, and goalkeeping handling. On Image 6, there are 7 numerical variables that are going to be used in Density-Based Clustering. The variables used in Density-Based Clustering are almost similar as Linear Regression whereas Density-Based Clustering has dribbling, while Linear Regression uses goalkeeping handling as one of the independent variables. Image 7 shows 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.

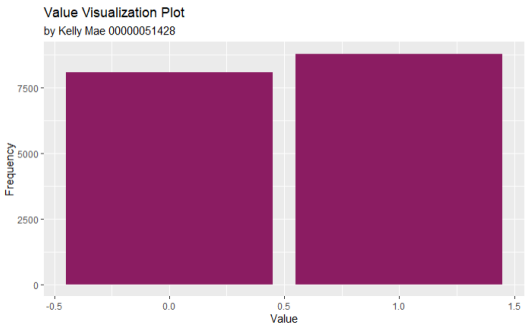

Image 8 Value Visualization Plot

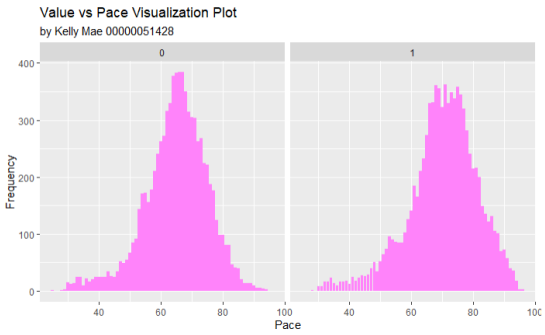

Image 9 Value and Pace Visualization Plot

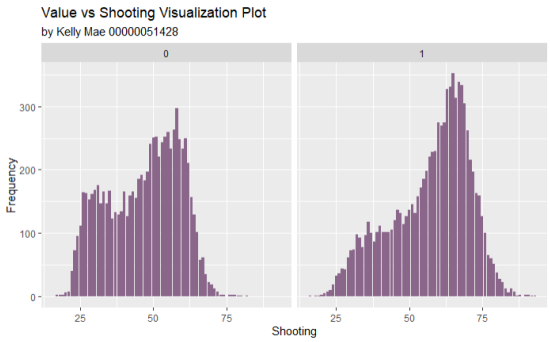

Image 10 Value and Shooting Visualization Plot

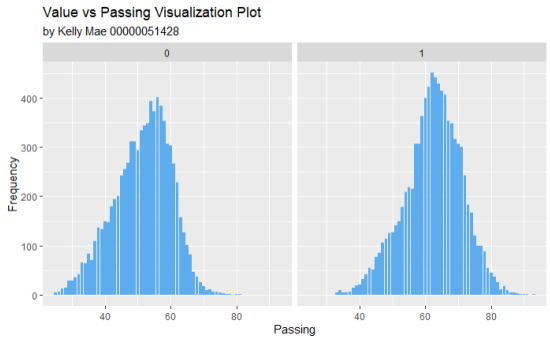

Image 11 Value and Passing Visualization Plot

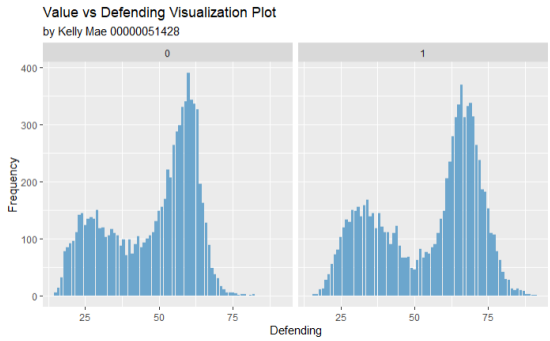

Image 12 Value and Defending Visualization Plot

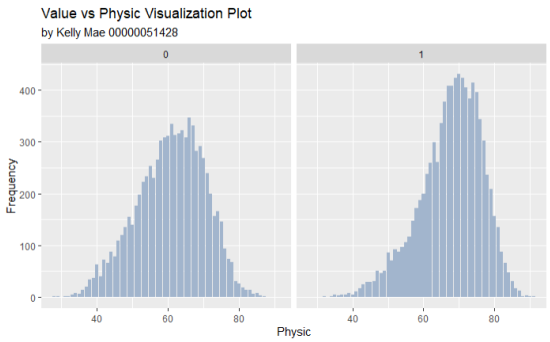

Image 13 Value and Physic Visualization Plot

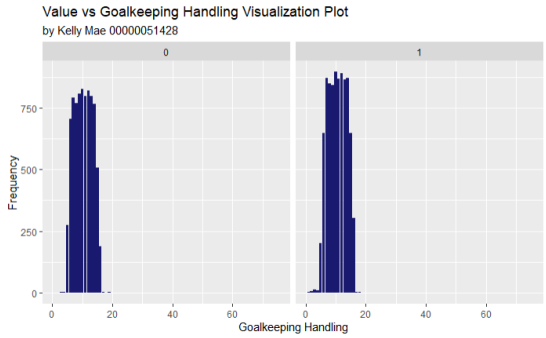

Image 14 Value and Goalkeeping Handling Visualization Plot

Image 8 until Image 14 shows frequencies between value as the predictor variable and other response variables through bar plots.

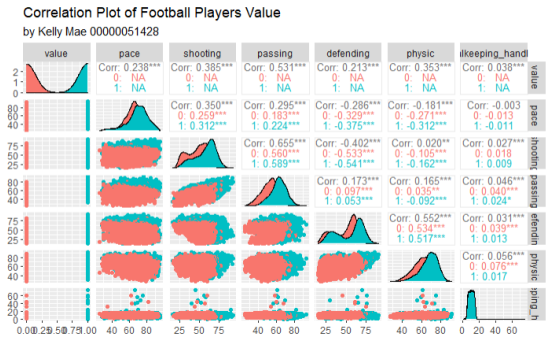

Image 15 Correlation Plot

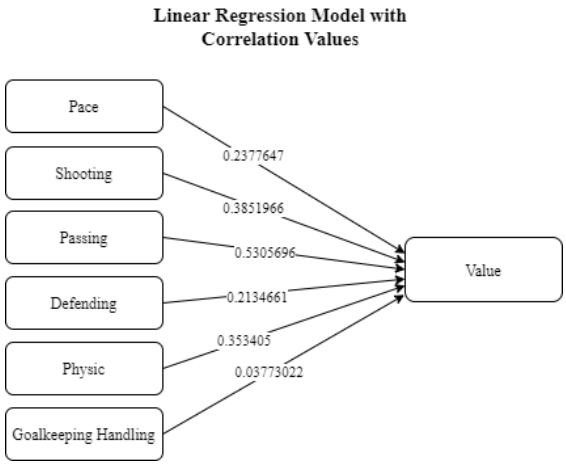

Image 16 Linear Regression Model with Correlation Values

Image 15 shows correlation plot which indicates relationships between response variable and other predictor variables used. Meanwhile, Image 16 illustrates the upcoming model that will be used with some correlation values between value and other predictor variables. In this case, there is no strong correlation between value and other dependent variables.

```
        Wilcoxon signed rank test with continuity correction

data:  numdata2$pace and numdata2$value
V = 2001000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Image 17 Wilcoxon Signed-Rank Test on Value and Pace

```
        Wilcoxon signed rank test with continuity correction

data:  numdata2$shooting and numdata2$value
V = 2001000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Image 18 Wilcoxon Signed-Rank Test on Value and Shooting

```
        Wilcoxon signed rank test with continuity correction

data:  numdata2$passing and numdata2$value
V = 2001000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Image 19 Wilcoxon Signed-Rank Test on Value and Passing

```
        Wilcoxon signed rank test with continuity correction

data:  numdata2$dribbling and numdata2$value
V = 2001000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Image 20 Wilcoxon Signed-Rank Test on Value and Dribbling

```
        Wilcoxon signed rank test with continuity correction

data:  numdata2$defending and numdata2$value
V = 2001000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Image 21 Wilcoxon Signed-Rank Test on Value and Defending

```
        Wilcoxon signed rank test with continuity correction

data:  numdata2$physic and numdata2$value
V = 2001000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Image 22 Wilcoxon Signed-Rank Test on Value and Physic

The Wilcoxon Signed-Rank Tests with continuity correction seen on the Image 17 until Image 22 are being used in the data exploration analysis. The tests show that there are statistically significant differences between value as the response variable with predictor variables. Moreover, the relationships between the response and predictor variables can be considered strong if p-value is greater than alpha ($\alpha = 0.05$).

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. During the assessment and testing stage, Linear Regression model will be created using independent and dependent variables that are existed in order to determined how close the data has fitted into the model by looking from adjusted R squared value. In addition, there will be model identification through regression diagnostic plot and other tests, such as Durbin-Watson Test, VIF (Variance Inflation Factor) Test, and Anderson-Darling Normality Test.

```
Call:
lm(formula = value ~ ., data = training)

Residuals:
     Min       1Q   Median       3Q      Max
-1.19950 -0.30841  0.01648  0.30659  1.10769

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2.230e+00  3.601e-02 -61.929   <2e-16 ***
pace               8.071e-03  3.455e-04  23.360   <2e-16 ***
shooting           7.026e-03  4.888e-04  14.372   <2e-16 ***
passing            1.358e-02  6.048e-04  22.459   <2e-16 ***
defending          4.819e-03  3.851e-04  12.514   <2e-16 ***
physic             1.267e-02  4.580e-04  27.653   <2e-16 ***
goalkeeping_handling  3.105e-05  1.071e-03   0.029    0.977
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3914 on 13481 degrees of freedom
Multiple R-squared:  0.3863,    Adjusted R-squared:  0.3861
F-statistic:  1415 on 6 and 13481 DF,  p-value: < 2.2e-16
```

Image 23 Linear Regression Model

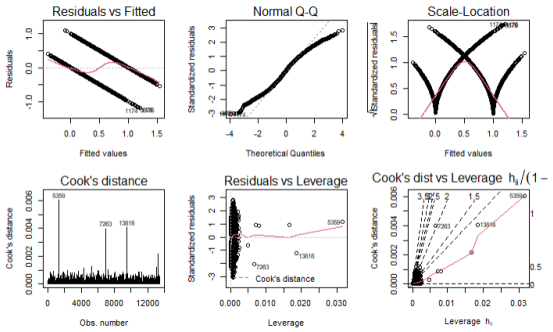

Image 24 Linear Regression Diagnostic Plot

```
        Durbin-Watson test

data:  model
DW = 2.0161, p-value = 0.8253
alternative hypothesis: true autocorrelation is greater than 0
```

Image 25 Linear Regression Model Identification using Durbin Watson Test

```
                pace             shooting              passing             defending
            1.264823             4.110562             3.390139             3.510749
              physic goalkeeping_handling
            1.774768             1.004257
```

Image 26 Linear Regression Model Identification using Variable Inflation Factor (VIF) Test

```
        Anderson-Darling normality test

data:  model$residuals
A = 54.23, p-value < 2.2e-16
```

Image 27 Linear Regression Model Identification using Anderson-Darling Normality Test

Image 23 shows the summary of the Linear Regression model in which the adjusted R-squared reached 39% (0.3861). On Image 24 until Image 27, the regression diagnostic has been performed in 3 tests, such as Durbin-Watson Test, VIF (Variance

Inflation Factors) Test, and Anderson-Darling Normality Test. To begin with, p-value (0.8253) in Durbin-Watson Test is greater than alpha ($\alpha = 0.05$) which concludes that the data has no autocorrelation. Based on VIF Test, there is a moderate correlation between the predictor variables. In this case, there is no multicollinearity in the data (VIF < 5), although there are several values like shooting, passing, and defending are almost close to 5. Furthermore, according to Anderson-Darling Test, p-value is lesser than alpha ($\alpha = 0.05$) so the data residuals are not normally distributed.

On the other hand, the Density-Based Clustering algorithm will reveal clusters in the form of arbitrary shape based on the number of clusters with suitable epsilon (ε) and minPts calculated in K-Nearest Neighbor (KNN) Distances. The next process involves validating the clusters with the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method. As a result, there will be several significant clusters with their corresponding densities. In addition, by having a minimum feature with the cluster developed, this cluster may be divided into considerable cluster size.

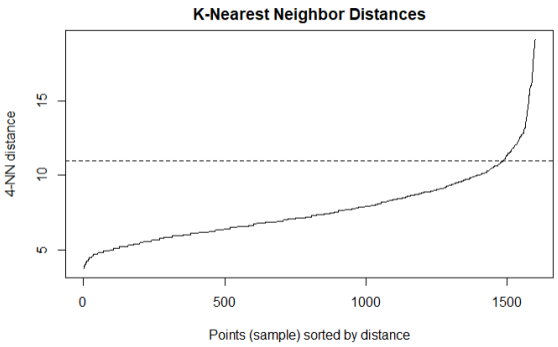

Image 28 K-Nearest Neighbor Distances Plot

```
DBSCAN clustering for 1600 objects.
Parameters: eps = 11, minPts = 5
The clustering contains 2 cluster(s) and 32 noise points.

   0    1    2
  32 1563    5

Available fields: cluster, eps, minPts
```

Image 29 Density-Based Spatial Clustering of Application Model

Image 28 shows the plot of K-Nearest Neighbor Distances using the applied epsilon (ε) and minPts, which are 11 and 5 respectively. On Image 29, the calculation of epsilon and minPts will be reimplemented in the DBSCAN method that results in clusters with the length of 2 factors which are made from the data filtering on target variables based on median in the previous part. In this circumstance, the minimum feature is selected in a cluster of 2 which might result in 2 clusters, however the clusters with less than 2 factors will be taken into consideration noise in order to fulfill its requirement.

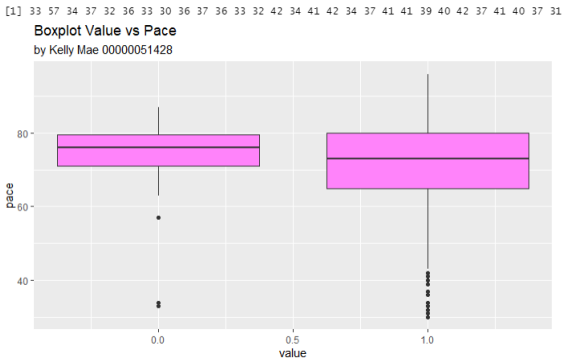

Image 30 Outliers at Value and Pace Boxplot

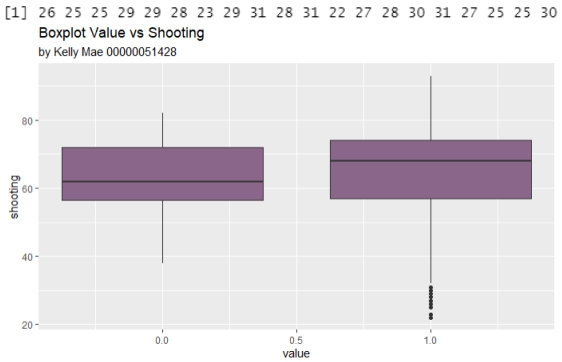

Image 31 Outliers at Value and Shooting Boxplot

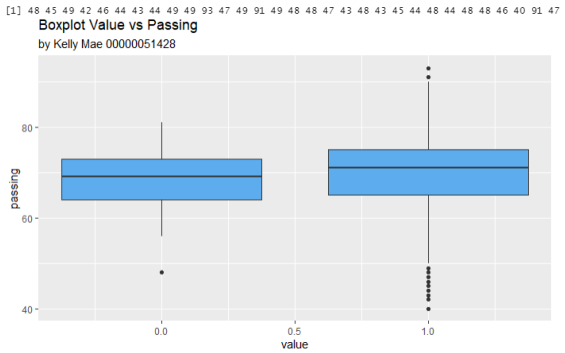

Image 32 Outliers at Value and Passing Boxplot

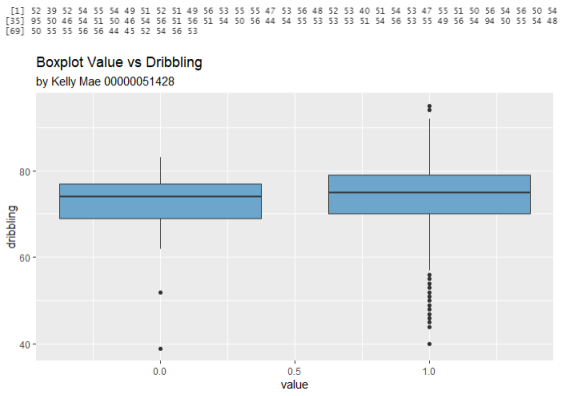

Image 33 Outliers at Value and Dribbling Boxplot

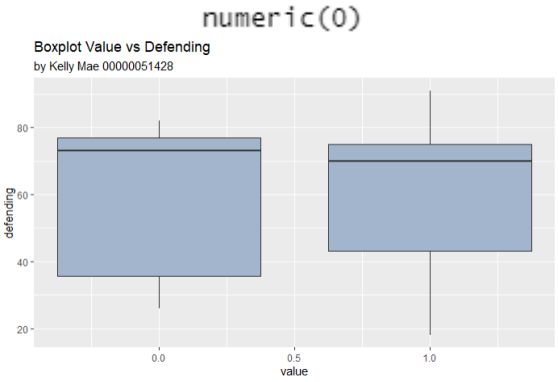

Image 34 Outliers at Value and Defending Boxplot

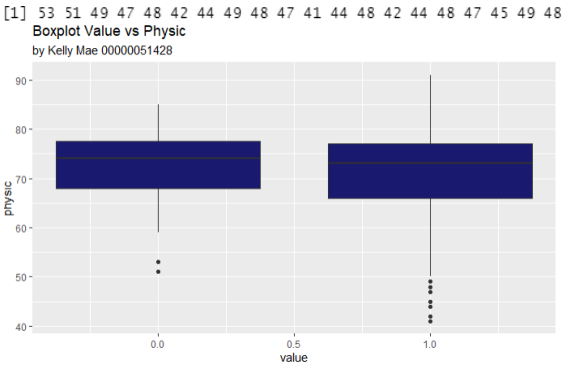

Image 35 Outliers at Value and Physic Boxplot

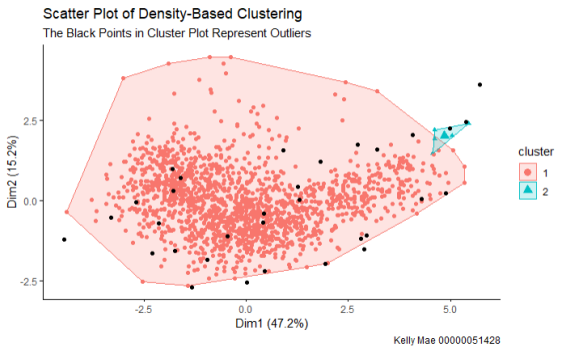

Image 36 Density-Based Clustering Scatter Plot

Image 30 until Image 35 shows outliers from the clusters that are separated based on their predictor variables in the form of boxplots. The scatter plot appeared in Image 36 shows the overall outliers that are represented with black dots from 2 clusters provided in the Density-Based Clustering model.

C. *Implementation*

The implementation process begins with the prediction at Linear Regression model based on training data using the testing data which results in RMSE (Rooted Mean Square Error) value in order to know the difference between the actual value and the predicted value, while Density-Based Clustering model will result in confusion matrix.

```
[1] 0.5907667
```

Image 37 Rooted Mean Square Error (RMSE) Value in Linear Regression Model

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1527   36
         1    5    0

               Accuracy : 0.9739
                 95% CI : (0.9647, 0.9812)
    No Information Rate : 0.977
    P-Value [Acc > NIR] : 0.8243

                  Kappa : -0.0056

 Mcnemar's Test P-Value : 2.797e-06

            Sensitivity : 0.9967
            Specificity : 0.0000
         Pos Pred Value : 0.9770
         Neg Pred Value : 0.0000
             Prevalence : 0.9770
         Detection Rate : 0.9739
   Detection Prevalence : 0.9968
      Balanced Accuracy : 0.4984

       'Positive' Class : 0
```

Image 38 Confusion Matrix and Statistics in Density-Based Clustering Algorithm

D. *Results Validation*

The last phase is the comparison between Linear Regression and Density-Based Clustering models. To begin, the estimation of Linear Regression can be seen from the RMSE value and adjusted R squared value whether the model is appropriate for the prediction. Linear Regression model can be considered good if RMSE value is low and adjusted R squared value is near to 1. Simultaneously, the level of accuracy from Density-Based Clustering will be observed in which the level of accuracy above 50% marks from another model can be used rather than doing random guesses. The model with

higher accuracy with a good fit and slight error than the other determines which model is better to be applied in the real-life prediction.

```
    15309      8071      3715     13274     10469
0.4080440 0.3921832 0.8917522 0.1538285 0.7269793
```

Image 39 First 5 Objects from Prediction with Linear Regression

```
          Reference
Prediction    0    1
         0 1527   36
         1    5    0
```

Image 40 Prediction with Density-Based Clustering

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 outcomes.

Based on the Linear Regression algorithm implementation that has been done, the model shows that the Adjusted R-Squared has reached 0.3861 (0.39) which is far from 1 and the RMSE (Root-Mean-Square Error) value reached 0.5 (0.5907667) which is really good. In this case, a linear regression model can be considered good if the Adjusted R-Squared is near to 1 and has lower RMSE value. As the result, this model is not acceptable to be used for further research because of the low Adjusted R-squared which is 0.38, in spite of good RMSE value. Meanwhile, the Density-Based Clustering algorithm model has higher accuracy that reaches 97% (0.9739). As a result, the Density-Based Clustering algorithm is acceptable to be predicted 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. Linear Regression involves correlation in independent and dependent variables to achieve adjusted R squared value, while Density-Based Clustering operates by using arbitrary-shaped clusters from the epsilon calculation in K-Nearest Neighbor Distances to be validated through the DBSCAN method which will bring out substantial clusters.

Although 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 Density-Based Clustering 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 Density-Based Clustering 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] D. H. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 140-147, 2020.

[7] K. M. Kumar and A. R. M. Reddy, "A Fast DBSCAN Clustering Algorithm by Accelerating Neighbor Searching using Groups Method," *Pattern Recognition*, vol. 58, pp. 39-48, 2016.

[8] A. Krashniak and E. Lamm, "Francis Galton's Regression towards Mediocrity and the Stability of Types," *Studies in History and Philosophy of Science*, vol. 86, pp. 6-19, 2021.

[9] M. Hahsler, M. Pickenbrock and D. Doran, "dbscan: Fast Density-Based Clustering with R," *Journal of Statistical Software*, vol. 91, no. 1, pp. 1-30, 2019.

[10] I. Boldina and P. G. Beninger, "Strengthening Statistical Usage in Marine Ecology: Linear Regression," *Journal of Experimental Marine Biology and Ecology*, vol. 474, pp. 81-91, 2016.

[11] P. Bhattacharjee and P. Mitra, "A Survey of Density Based Clustering Algorithms," *Frontiers of Computer Science*, vol. 15, no. 1, pp. 1-27, 2020.

[12] S. R. Joseph, H. Hlomani and K. Letsholo, "Data Mining Algorithms: An Overview," *International Journal of Computers and Texhnology*, vol. 15, no. 6, pp. 6807-6813, 2016.

[13] Y. Lv, T. Ma, M. Tang, J. Cao, Y. Tian, A. Al-Dhelaan and M. Al-Rodhaan, "An Efficient and Scalable Density-Based Clustering Algorithm for Datasets with Complex Structures," *Neurocomputing*, vol. 171, pp. 9-22, 2016.