Comparison Between Decision Tree Prediction Model and Naïve Bayes in Predicting the Rating of Rental Car

Classification Algorithm

Maria Darlene Kusnadi[1]
[1] Information System Study Program, Faculty of Engineering and Informatics, Multimedia Nusantara University, South Tangerang, Indonesia

Accepted on 9 October 2021
Approved on 9 October 2021

***Abstract*—Classification is a data mining algorithm that will assign a variable in a group of variables into a target variable , therefore accurately predicting that target variable for every data cases[1]. The research object of this report is the car rating of Car Rental. The purpose of this research is to form a model that will predict that rating based on the supporting facts using Decision Tree and Naïve Bayes algorithms and then compare both of the results to decide which model is the better model for this dataset. Hopefully, this analysis result could become a reference in predicting the rating of a car in car rental business to prevent procuring unprofitable cars.**

Index Terms— Decision Tree; Naïve Bayes; Data Mining; Classification; Car Rental; Rating **(;)**

I. INTRODUCTION

In recent years, Big Data has Big Data is a term used for a massive and complex collection of data that is hard to parse or process using traditional or simple processing means. Lately, this term is also often used to refer the usage and processing of the data itself in Information Technology fields. The defining characteristics of big data could be concluded in 4V, which are volume, velocity, variety, and veracity. Volume is the size of the data that needs to be processed. As the name suggest, Big Data processes a very large amount of data, it could even reach petabytes of data each day. Velocity refers to the speed of the produced data. The data processed by big data are often of high speed, therefore needing a specialized form of data processing. Variety refers to how Big Data is sourced from many sources, and contain different types of data, which are structured data, semi-structured data, and unstructured data. Finally, veracity refers to the accuracy or the consistency of the data. A highly accurate data would produce good quality analysis, in contrast, data with low accuracy would also produce low quality analysis, therefore data processing is highly important in Big Data usage[2].

In recent years, Big Data is getting more and more popular among businesses. This is caused by the potential benefits that Big Data brought for businesses. Big Data gives immense benefit in decision making process. Big Data could process the numerous trivial-looking data into vital information for the business. Business owner could make more impactful and accurate decisions based on data, therefore reducing the risks involved in decision making and improve business' growth. Big Data could also help in making predictions through market trend and customer's input, helping decision makers in forecasting future steps that the business would take. Better decision making leads to cost reduction and profit increase, all of which are very beneficial for businesses[3]. Every

business benefits from Big Data analysis, not excluding the car rental business.

Car rental is a business where the customer would borrow a car, according to their needs and circumstances, from the vendor for a price. Their target customers are usually need based customers and leisure travelers. The success of the car rental business is dependent on how well received it is by their customers. Therefore, selecting suitable cars for rent is a vital process to gain favorability from the customers. In this case, Big Data could play a big role in helping the car-rental business' decision making. It could help in predicting which car is more likable amongst the customers, resulting in a less risky investments from the car rental vendor[4].

II. Literature Review

A. *Classification Algorithms*

Classification is a fundamental learning method in data mining. Most of the classification algorithms are supervised methods. Classification algorithms work by learning a set of examples that are already classified, and then it will learn to assign unseen examples based on the prior examples that were presented. The primary task for this algorithm is to assign class labels to new observations. It is a very useful algorithm as it is used to deal with problems that occur often, such as detecting fraudulent in transactions, determining medical conditions based on symptoms, detecting spam mails, and many more[5].

B. *Decision Tree Algorithm*

Decision Tree Algorithm is one of the basic classification methodology. It is a flowchart-like tree structure, modeled using a set of hierarchical decisions on the feature variables, with nodes that represents a test on an attribute, branches that represents an outcome of a test, and class label that is represented each leaf node. In Decision Tree, the first node is called the "root node", and the terminal node is called the "leaf node". As a learning model, Decision Tree is used as a predictive model that maps observations about an item's target value[6] .

C. *Naïve Bayes Algorithm*

Naïve Bayes algorithm is one of the basic classification methodology. It is an algorithm that estimates the probability of observing outcome when given an outcome by using predictor values. It is an algorithm that operates based on maximum-likelihood. The Naïve Bayes is called naïve because of the naïve assumption that the covariates are independent within each class. Even though it is called 'naïve', this algorithm still works well in complex situations, and usable in many conditions[7].

D. *Benefit and Weakness of both algorithms*

Decision Tree is an algorithm that could be implemented in a simple way, and also easy to interpret. With sufficient training data, it could even model complex decision boundaries accurately. However, to make a good Decision Tree model with complex boundaries, it needs a very large amount of training data. Without sufficient training data, it would result in coarse approximation of the boundaries rather than a proper approximation, resulting in an overfitted model. Another drawback in using Decision Tree is the possibility of overlapping decision, especially when it needs to process a dataset with large amount of classes. The large amount of data also causes difficulty in arranging the most optimal tree, and requires a large amount of memory to store it[8].

Naïve Bayes is a simple and flexible algorithm that fits many types of cases. It could be implemented in a wide range of settings. Unlike the Decision Tree, Naïve Bayes doesn't need as much training data to make a good model. Naïve Bayes could process missing data by ignoring it in the calculation, however, Naïve Bayes can't process 0 probability value, as the predicted probability would also become 0. The

accuracy of Naïve Bayes algorithm often times are not very optimal, due to the strong naïve assumption[9].

E. *Confusion Matrix*

Confusion matrix is one of the methods used in calculating the accuracy of data mining concepts. Confusion matrix shows the predicted and the actual classifications of the resulting predictions of classification algorithms. Values that are predicted as the positive values and are the actual positive values, are called true positives, while values that are predicted as positive values but are actually negative values are called false positive. Values that are predicted as negative values and are the actual negative values are called true negatives, while values that are predicted as negative values but actually are positive values are called false negatives[10].

There are several things that could be determined from a confusion matrix. For example, the accuracy, sensitivity, and specificity of the model. Sensitivity measures the number of true positive statements in comparison to all positive statements, meaning it shows how good a model is in detecting abnormal condition. Specificity measures the number of true negative statements in comparison to all negative statements, meaning it shows how good a model is in detecting normal condition. Both specificity and sensitivity contributes to the accuracy of the model. Accuracy measures the number of correct assessments in comparison to all assessments. It indicates the correctness of a diagnostic test in identifying and excluding a given condition[11].

III. METHODOLOGY

A. *Research Object*

The research object for this research is the Car Rental data in major US cities. The dataset was taken from a data science learning community website called Kaggle. This dataset includes data related to the car rental business in US, such as the car type, the city which it operates, the typical rental fare in US, the amount of trips taken by the renter, the amount of review submitted by different renters on a car, the year of the vehicle, and the cumulative rating done by the customer.

B. *Data Gathering*

The data used in this research are quantitative data, which includes daily rate, the amount of renter trips taken, review count, vehicle year, and it's rating. In total, the data consists of 15 variables, 5 and 5851 rows. However, since the analysis only includes quantitative data, it will analyze 5 of the 15 variables.

C. *Framework*

These are the steps done to conduct this research analysis.

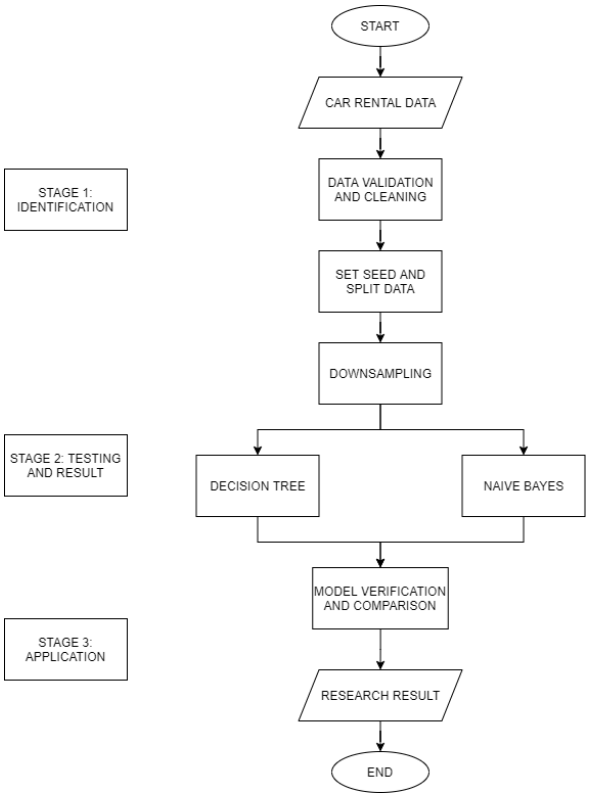

Figure 3. 1 Framework work-flow

D. *Identification*

Identification is the first stage. In this stage the data, which is in the form of comma-separated values (.csv) would be inputted into

R before it could be analyzed using the same program. In preparation of processing the data, it is also necessary to install or call all the packages needed. After the data and the packages has been imported, then it is time to decide on the target variable. In this research case, the target variable is the rating variable. The rating variable is a numerical variable, therefore to utilize all the chosen classification algorithms, it needs to be transformed into a categorical variable. A new variable was made by the name of 'class', where ratings below 4.8 are considered as "bad", and from 4.9 and above are considered as "good".

As stated before, the data used in this research are qualitative data, therefore the dataset needs to be filtered. To make sure the algorithm would function well, the data needs to be cleansed. It is necessary to check whether the data has missing values or not. In R, it could be done by using the missmap function. The missing data was then omitted, making the formerly 5851 rows of data into 5350 rows. The data was then split into testing data, which is 25% of the whole data, and training data, which is 75% of the whole data, in preparation to make classification algorithm models.

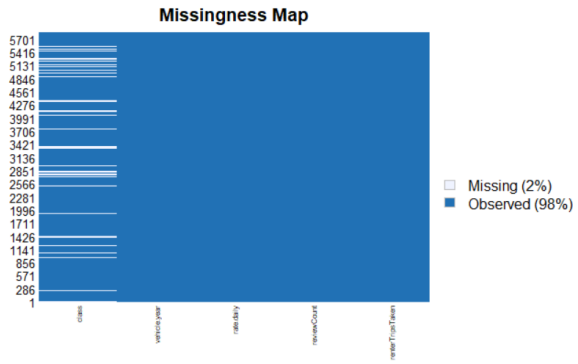

Figure 3. 2 Missingness Map before data cleansing

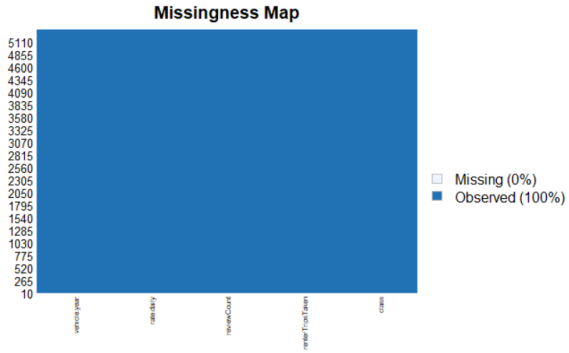

Figure 3. 3 Missingness Map after data cleansing

After cleaning the data, descriptive statistics were applied to see the distributions of the data. Unfortunately, it was found that the distributions are very imbalanced, therefore down-sampling was used. Descriptive statistics were applied again to confirm the renewed distributions. This step concludes the identification stage of the analysis process.

E. Testing and Predicting

The second stage is the testing and predicting stage. In this stage, the prediction models using classification algorithms are made. The models are made using training data. Discrete variable value as the target variable, connected to independent variables which are numeric variables from the dataset. The Decision Tree algorithm model would create nodes according to the conditions formed by the classification, whereas Naïve Bayes algorithm model would use probabilities from every predicted possibilities. From these models, it is possible to see the accuracy, sensitivity, and specificity to compare the models.

F. Application

The final stage is the application stage. In the application stage, the two models that were made in prior stage would be validated using the accuracy of each model. If the accuracy exceed 50%, then the model is usable because it means that using the model is better than wild guessing. However, if the model's accuracy is below 50%, than it is not usable. After validating, the two models

would be compared to each other. The comparison could be seen from each model's confusion matrix.

IV. Results and Discussion

A. *Identification*

The data used in this research is a secondary data retrieved from the data science learning community website, Kaggle. It is a car rental data from United States, with 5350 rows of data and 5 variables. Below is the data used in this research after filtering and cleansing.

```
'data.frame':   5350 obs. of  5 variables:
 $ class          : Factor w/ 2 levels "Bad","Good": 2 2 2 2 2 2 1 2 2 1 ...
 $ renterTripsTaken: int  13 2 28 21 3 13 13 12 1 22 ...
 $ reviewCount    : int  12 1 24 20 1 12 12 10 1 17 ...
 $ rate.daily     : int  135 190 35 75 47 58 42 117 102 49 ...
 $ vehicle.year   : int  2019 2018 2012 2018 2010 2012 2005 2018 2016 2018 ...
 - attr(*, "na.action")= 'omit' Named int [1:501] 17 30 45 88 107 113 122 123 145 152
...
  ..- attr(*, "names")= chr [1:501] "17" "30" "45" "88" ...
```

Figure 4. 1 data structure

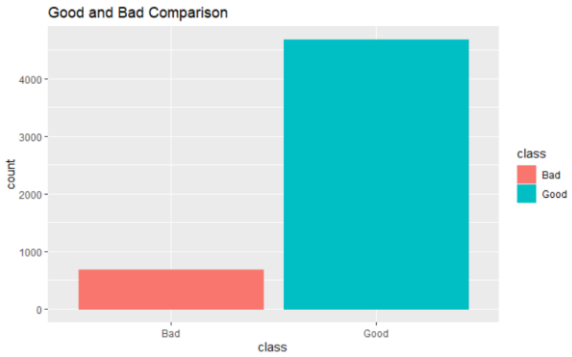

Figure 4. 2 Good and bad comparison before down-sampling

Because of the unbalanced data, down-sampling was done to improve the model. Below are the data and it's visualization after down-sampling was done.

```
'data.frame':   1018 obs. of  5 variables:
 $ renterTripsTaken: int  14 41 7 144 11 34 49 42 79 109 ...
 $ reviewCount     : int  10 32 6 116 11 28 42 29 62 88 ...
 $ rate.daily      : int  125 33 74 48 29 62 47 35 29 30 ...
 $ vehicle.year    : int  2016 2014 2015 2015 2011 2015 2010 2011 2013 2016 ...
 $ Class           : Factor w/ 2 levels "Bad","Good": 1 1 1 1 1 1 1 1 1 1 ...
[1] 5350
```

Figure 4. 3 Down-sampled data structure

Figure 4. 4 RenterTripsTaken Distribution by class

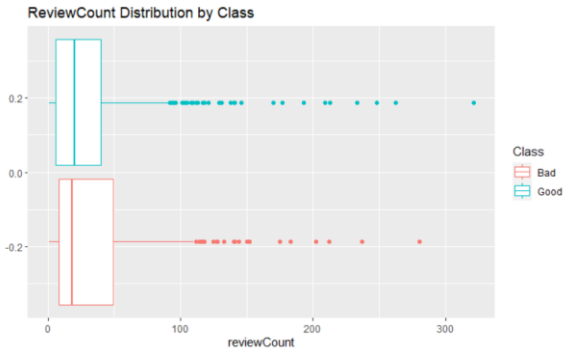

Figure 4. 5 ReviewCount Distribution by class

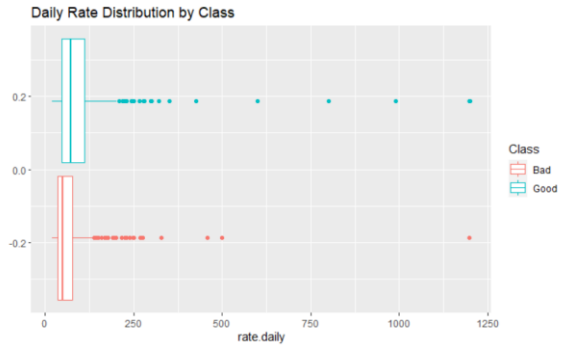

Figure 4. 6 Daily Rate Distribution by Class

B. *Testing and Predicting*

After the identification stage, comes the testing and predicting stage. The classification algorithms models made in this stage are Decision Tree and Naïve Bayes.

There are two versions of Decision Tree algorithm model made for this research. The first one uses party package, and the other one uses rpart package.

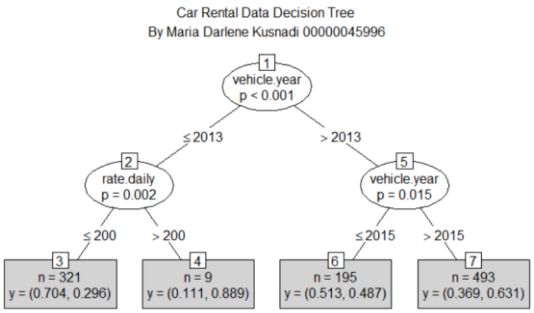

Figure 4. 7 Party package decision tree model visualization

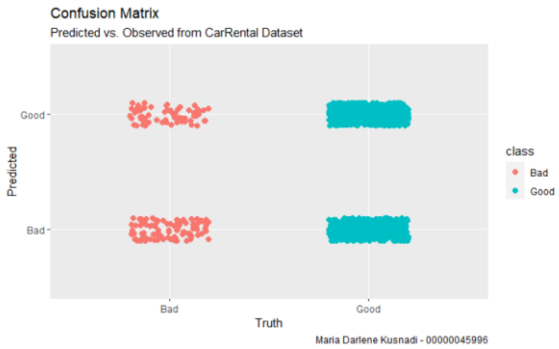

Figure 4. 8 Party package confusion matrix visualization

```
predict_data Bad Good
        Bad   98  498
        Good  67  675

               Accuracy : 0.5777
                 95% CI : (0.5507, 0.6044)
    No Information Rate : 0.8767
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0798

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.59394
            Specificity : 0.57545
         Pos Pred Value : 0.16443
         Neg Pred Value : 0.90970
             Prevalence : 0.12332
         Detection Rate : 0.07324
   Detection Prevalence : 0.44544
      Balanced Accuracy : 0.58469

       'Positive' Class : Bad
```

Figure 4. 9 Party package confusion matrix

The confusion matrix of party package decision tree shows that the accuracy of this tree is 57.77%, the sensitivity is 59.39%, the specificity is 57.54%, and the precision is 16.44%. This model is usable, because the accuracy is more than 50%, making it a better solution than pure wild guesses.

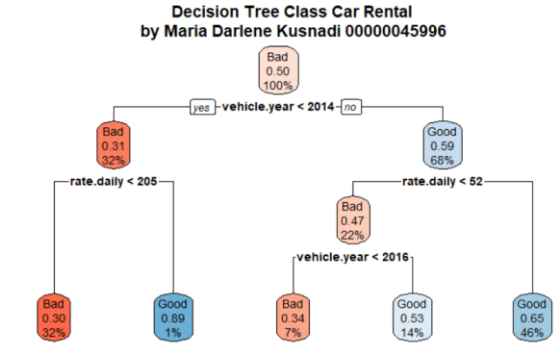

Figure 4. 10 Rpart package decision tree model visualization

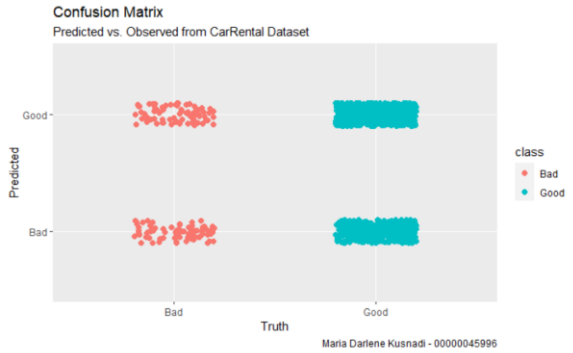

Figure 4. 11 Rpart package confusion matrix visualization

```
predict_rpart_car Bad Good
             Bad   84  344
             Good  81  829

               Accuracy : 0.6824
                 95% CI : (0.6567, 0.7073)
    No Information Rate : 0.8767
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1281

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.50909
            Specificity : 0.70673
         Pos Pred Value : 0.19626
         Neg Pred Value : 0.91099
             Prevalence : 0.12332
         Detection Rate : 0.06278
   Detection Prevalence : 0.31988
      Balanced Accuracy : 0.60791

       'Positive' Class : Bad
```

Figure 4. 12 Rpart package confusion matrix

The confusion matrix of rpart package decision tree shows that the accuracy of this tree is 68.24%, the sensitivity is 50.90%, and the specificity is 70.67%, and the precision is 19.62%. This model is usable, because the accuracy is more than 50%, making it a better solution than pure wild guesses.

```
> (party_accuracy <- party$overall[1])
Accuracy
0.577728
> (rpart_accuracy <- rpart$overall[1])
 Accuracy
0.6823617
```

Figure 4. 13 Model comparison

Both of the models, party and rpart are usable. From both of the confusion matrixes, some difference could be seen. These differences appear because both of them use slightly different algorithms. The party package uses significance test procedure to select variables, whereas the rpart package selects variables that maximizes information measurement. From both of the confusion matrixes, generally the rpart package has higher statistics than the party model, such as the kappa, specificity, and precision. The accuracy of the rpart package is also higher than the party package. Therefore it could be concluded that the rpart model is the better model for this dataset.

There is only one version of Naïve Bayes algorithm model made for this research.

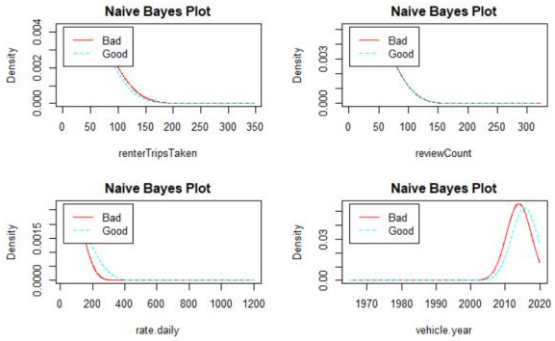

Figure 4. 14 Naive Bayes plot

Figure 4. 15 Naive Bayes confusion matrix visualization

```
          Bad Good
Bad       118  725
Good       47  448

               Accuracy : 0.423
                 95% CI : (0.3964, 0.45)
    No Information Rate : 0.8767
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0351

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.71515
            Specificity : 0.38193
         Pos Pred Value : 0.13998
         Neg Pred Value : 0.90505
             Prevalence : 0.12332
         Detection Rate : 0.08819
   Detection Prevalence : 0.63004
      Balanced Accuracy : 0.54854

       'Positive' Class : Bad
```

Figure 4. 16 Naive Bayes confusion matrix

The confusion matrix of Naïve Bayes algorithm shows that the accuracy is 42.3%, the sensitivity is 71.51%, the specificity is 38.19%, and the precision is 13.99%. This model is unusable, because the accuracy is below 50%, making it worse than pure wild guess. The reason for this unfortunate result could be seen from the Naïve Bayes plots. From the plots, it could be seen that there are not much differences between the two classes, making it difficult for Naïve Bayes algorithm to classify between them.

C. Application

In the final stage, the application stage, the two algorithm models, Decision Tree and Naïve Bayes, are compared. From Decision Tree, the rpart package model will be chosen to represent the algorithm as it has higher accuracy. The two models are created using the same testing and training data, however they are also created using different methods and approach, therefore producing different results.

The results show that Decision Tree algorithm has a much better accuracy, with 68.24% accuracy, than the Naïve Bayes algorithm, with 42.3% accuracy. Therefore, the Decision Tree algorithm model is a better choice for predicting the rating of car rental data.

V. CONCLUSION

The two classification algorithm models produce different results based on their approach in each methods. Decision Tree is the algorithm that uses nodes to describe different conditions that might be formed, whereas Naïve Bayes is the algorithm that uses probability of each class to produce prediction model. To successfully predict the target variable of the dataset, an accurate prediction model is needed. In this research, it is found that in determining the rating of cars in car rental business, the Decision Tree is the more suitable choice than the Naïve Bayes. Therefore, it would be better to use the Decision Tree algorithm when dealing with similar dataset and target variable in the future.

REFERENCES

[1] Oracle. (2020, November). *Classification*. Retrieved from Oracle: https://docs.oracle.com/en/database/oracle/oracle-database/19/dmcon/classification.html#GUID-3D51EC47-E686-4468-8F49-A27B5F8E8FE4

[2] Supriyanto, E., Bakti, I., & Furqon, M. (2021). The Role of Big Data in The Implementation of Distance Learning. *Paedagoria : Jurnal Kajian, Penelitian dan Pengembangan Kependidikan, 12*(1), 61-69. Retrieved from http://journal.ummat.ac.id/index.php/paedagoria/article/view/3902/pdf

[3] Maryanto, B. (2017). Big Data dan Pemanfaatannya Dalam Berbagai Sektor. *Media Informatika, 16*(2), 14-19. Retrieved from https://jurnal.likmi.ac.id/Jurnal/7_2017/0717_02_BudiMaryanto.pdf

[4] BusinessPlanTemplate. (2021). *Car Rental Business Plan Template [2021 Updated]*. Retrieved from BusinessPlanTemplate.com: https://www.businessplantemplate.com/car-rental-business-plan-template/2/

[5] Wiley, J. (2015). *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data.* Indianapolis: John Wiley & Sons Inc.

[6] Sharma, H., & Kumar, S. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. *International Journal of Science and Research (IJSR), 5*(4), 2094-2097. Retrieved from https://www.researchgate.net/profile/Sunil-Kumar-310/publication/324941161_A_Survey_on_Decision_Tree_Algorithms_of_Classification_in_Data_Mining/links/5aebdfe6a6fdcc8508b6e8bb/A-Survey-on-Decision-Tree-Algorithms-of-Classification-in-Data-Mining.pdf.

[7] Bruce, P., & Bruce, A. (2017). *Practical Statistics for Data Scientists 50 Essential Concepts.* Sebastopol: O'Reilly Media.

[8] Aggarwal, C. C. (2015). *Data Mining The Textbook.* New York: Springer.

[9] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R Second Edition.* New York: Springer.

[10] Rosandy, T. (2016). Perbandingan Metode Naive Bayes Classifier dengan Metode Decision Tree (C4.5) untuk Menganalisa Kelancaran Pembiayaan. *Jurnal TIM Darmajaya, 2*(1), 52-62. Retrieved from https://jurnal.darmajaya.ac.id/index.php/jtim/article/view/648/429

[11] Zhu, W., Zeng, N., & Wang, N. (2020). Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations. *Health Care and Life Sciences, 1*(1), 1-9. Retrieved from https://www.lexjansen.com/nesug/nesug10/hl/hl07.pdf