e-ISSN: 2829-6192, p-ISSN: 2829-6184
DOI: https://doi.org/10.38035/ijam.v2i2
Received: 30 June 2023, Revised: 13 July 2023, Publish: 21 July 2023

Prediction of Customer Data Classification by Company Category Using *Decision Tree* Algorithm (Case Study: PT. Teknik Kreasi Solusindo)

Ryo Nicholas Reinaldo[1], Saruni Dwiasnati[2]
[1)] Universitas Mercu Buana, Jakarta, Indonesia, Email : 41519110159@student.mercubuana.ac.id
[2)] Universitas Mercu Buana, Jakarta, Indonesia, Email : saruni.dwiasnati@mercubuana.co.id

Corresponding Author: 41519110159@student.mercubuana.ac.id

Abstract: Classification using a *decision tree* for grouping customers with a case study of PT Teknik Kreasi Solusindo is a problem that exists in the company. Where the classification of customer data grouping in the company PT Teknik Kreasi Solusindo previously had no basis which resulted in the classification results not being entirely good. To overcome this problem, this study uses a classification method that exists in the data mining process, namely the *decision tree* algorithm. This study uses a *Decision tree* because the data used has a discrete type and the classification process is simple and fast. The data in the research used are product offering attributes of PT Teknik Kreasi Solusindo from 2018-2023. The source of the data obtained is the results of interviews with representatives from PT Teknik Kreasi Solusindo and also several bidding files in the company. Based on this data, a classification will be carried out with the Google Colabolatory service. And with this method, the accuracy of the *decision tree* method will be seen as a reference for the desired classification.

Keywords: Classification, Company Data, Data Mining, Decision Tree.

INTRODUCTION

PT Teknik Kreasi Solusindo is a company engaged in the field of machine manufacturing services, where the majority of their customers are companies that need machines to run their business, some of which are label machines, sticker machines, and packaging machines. However, there are several stages for a company to become a customer at PT Teknik Kreasi Solusindo, one of which is to offer and demo the machine. Usually, to do that, the company requires costs incurred, for example, the cost of transportation and fuel services and of course the costs incurred also vary depending on several variables, such as the number of visits of course the company wants to continue to provide the best for customers who have good prospects for special offers, because this can increase revenue and save costs for the offer process but this is a problem in itself because the company has difficulty determining whether a prospective customer is eligible or not to be given a special offer. So that data mining can predict whether potential customers can be said to be eligible or not to receive special offers,

special offers or special promotions can generate profits for a company, according to one study, sales promotion can strengthen brand loyalty behavior(Mendez, 2012).

Data mining is an activity to mine knowledge from large data (Susanto, 2014), And can be used to make an important business decision (Arti, Idrawan, Dantes, 2019). So that it can be said that Data Mining is a process to explore knowledge that can be useful for the business interests of a company.

This study uses a classification technique. Classification in data mining is a technique based on machine learning algorithms that use math, statistics, probability distribution and artificial intelligence.To predict group membership for data items or to represent a descriptive analysis of data items for effective decision making (Satyanarayana, Ramalingaswamy, Ramadevi, 2014).

The data mining method used in this study is the *Decision tree*. *Decision tree* is a simple representation of classification techniques which is the process of learning an objective function that maps each set attribute to one of the previously defined classes (Nurzahputra, Safitri, Muslim, 2016), in another study, the *decision tree* was also used to classify loyalty to company services, namely 78.61%, with a total of 11 parameters. where the purpose of this research is to identify customers who are loyal to the company, customers for a company are very important and vital, because customers are the key to the success of a business run by the company(Pradana, Saputro, 2020). Whereas the study on determining the customer satisfaction index using the *decision tree* itself has an accuracy of 84.61% with 21 parameters, the purpose of this research is to determine the accuracy of the marketing strategy, by examining the characteristics of consumer behavior (Febriyani, Prayoga, Nurdiawan, 2021). in another study that we conducted at one of the Bekasi City Public Middle Schools implementing the COVID-19 Vaccination program for their students. This research aims to do interviews, surveys and investigations at one of the Bekasi City Public Middle Schools to determine the assessment criteria AEFIs that have been received by recipients of the COVID-19 vaccine. Then, the classification of AEFIs was carried out on Sinovac vaccine recipients use the Decision Tree algorithm(Yusuf, F. A., Alfaridzi, M., & Herdi, T, 2022).

The purpose of this research is to get the accuracy of the *decision tree* method, so that the information obtained can be useful for the company's future strategy in improving services.

METHODS

This study uses datasets originating from companies in the form of physical data and interview sessions with someone as part of the company as many as a thousand records. Where the parameters used are the number of visits, nominal price, number of workers, company scale, number of purchases, 5% turnover, month of offer, and type of customer.

This study also uses the Google Collaborative service, so that after the data is collected it will be inputted and then preprocessed to be able to carry out the classification process using a *decision tree*, a complete example of this research can be seen in Table 1.

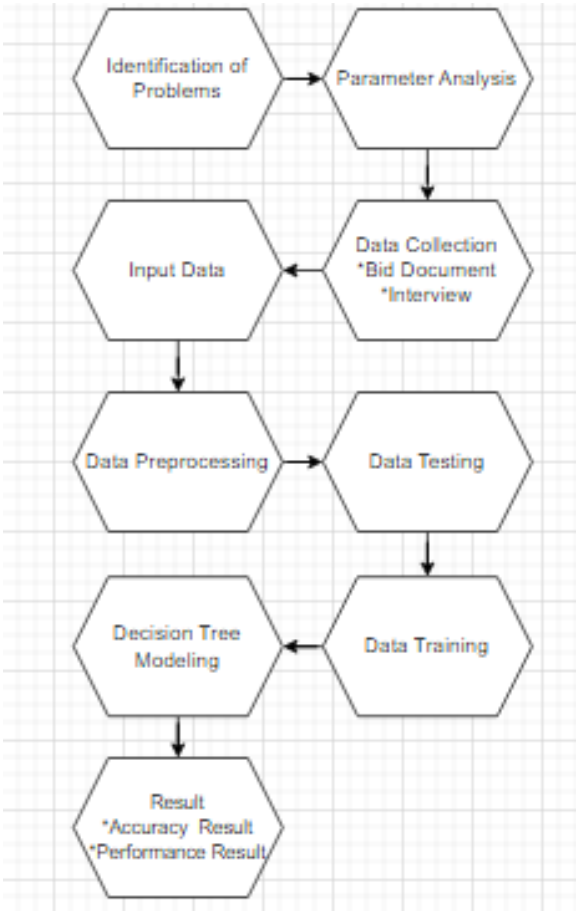

Figure 1. Research Method

A. Identification of Problems

The first part of the research stage is the identification of problems such as whether the *decision tree* algorithm can classify the types of prospective customers, Can the results of the classification accuracy help companies in making decisions.

B. Parameter Analysis

The second part is analyzing the parameters, which are used to determine what parameters will be included in the dataset for data mining. The parameters that have been determined include the number of visits, nominal price (in rupiah), number of workers, company scale, number of purchases, 5% turnover, the month of offer, and the type of customer.

C. Data Collection

Then proceed with collecting data from January 2018 to May 2023, obtained by interviewing a company representative and also data from bidding documents.

Table 2. Dataset

Number	Name
1	the number of visits
2	nominal price
3	number of workers
4	company scale
5	number of purchases
6	5% turnover
7	the month of offer
8	the type of customer

D. Input Data

After collecting the data, the next step is to input the data into Excel software with the .xlxs extension. Where this format will be used on Google Collaboration

Table 3. Input Data

the number of visits	nominal price	number of workers	company scale	number of purchases	5% turnover	the month of offer	the type of customer
2	145296000	more than 100	big business	0	yes	10	worthy
1	2948000	more than 100	big business	1	no	10	worthy
2	304565000	more than 100	big business	0	yes	6	not feasible
1	187000000	more than 100	big business	1	no	8	worthy
1	322275593	more than 100	big business	0	yes	10	not feasible

The picture above is an example of five rows of input dataset.

E. Data Preprocessing

There are 4 categorical variables, namely the number of workers, company scale, 5% turnover, and type of customer. The *Decision tree* in the Google Colab feature can only process numeric data, so categorical data needs to be preprocessed as follows

Table 4. Variable preprocessing results

Variable	Before	After
number of workers	more than 100	2
	above 50 and below 100	1
	under 50	0
company scale	big business	2
	medium business	1
	small business	0
5% turnover	yes	1
	no	0
Type of customer	worthy	1
	not feasible	0

F. Data Testing

The next data mining process that we need to do is separate training and testing data. Training data is data used to train data mining models or algorithms so that they can understand patterns in data and can predict accurately on data that has never been seen before.

G. Data Training

Data testing is data used to test the performance of models or data mining algorithms on data that has never been seen before. In this study we will divide the data by the proportion of Test 30% and Training 70%.

H. Decision *tree* Modelling

After we do preprocessing and distribute training and testing data, then we can create a *decision tree* model to predict the types of employees who are eligible or not to be given special offers, this can be done by classifying. *decision tree* builds a classification model with the shape of the tree structure where each the nodes of the tree are associated with the from attribute data, in this case, is the data variable (Latifah, Wulandari, Kreshna, 2019). The bottom of the existing node in the *decision tree* or commonly referred to as leaves, in the optimal case, have a value that is characteristic of the label class (Chiena, Rodriguez, Exposito, Batista, Moreno-Vega, 2018)Here is an example of how the *decision tree* algorithm works.

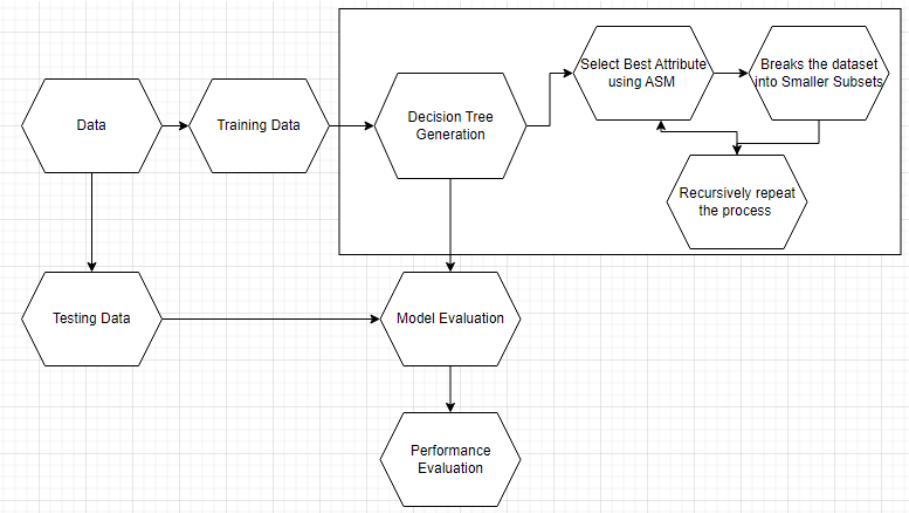

Figure 2. How decision tree work

For the select best attribute section using ASM which means Attribute Selection Measure is one way to choose criteria separator for classify data as best as possible (Latifah, Wulandari, Kreshna, 2019), for ASM used in this study is Entropy. Entropy is used in measuring the impurity of each node in the *decision tree*. The lower the entropy value, the "pure" or homogeneous the data at that node.

$$Entropy(s) = \sum_{i=1}^{n} - pi * log_2 pi$$

Where pi is the proportion of S_i to S.

I. Result

After the *decision tree* model is obtained, the next step is to evaluate the model. Evaluation in this research is the accuracy of the model and also the performance result, the confusion matrix is a table used to evaluate the performance of a classification model. The confusion matrix displays the number of correctly and incorrectly classified data based on the actual and predicted classes.

RESULT AND DISCUSSION

The results of this study only use the presentation of test data which is equal to 30% which is a rule of thumb in the *decision tree* method. In this study, the ASM setting was also carried out in the *decision tree* method by setting the criterion parameter with an entropy value which resulted in an accuracy above 0.7. Whereas if not, it will produce an accuracy of 0.68, thus showing ASM with better *entropy* than the default ASM *decision tree* method on Google Colab, namely *gini*. For the performance results section using the confusion matrix visualization.

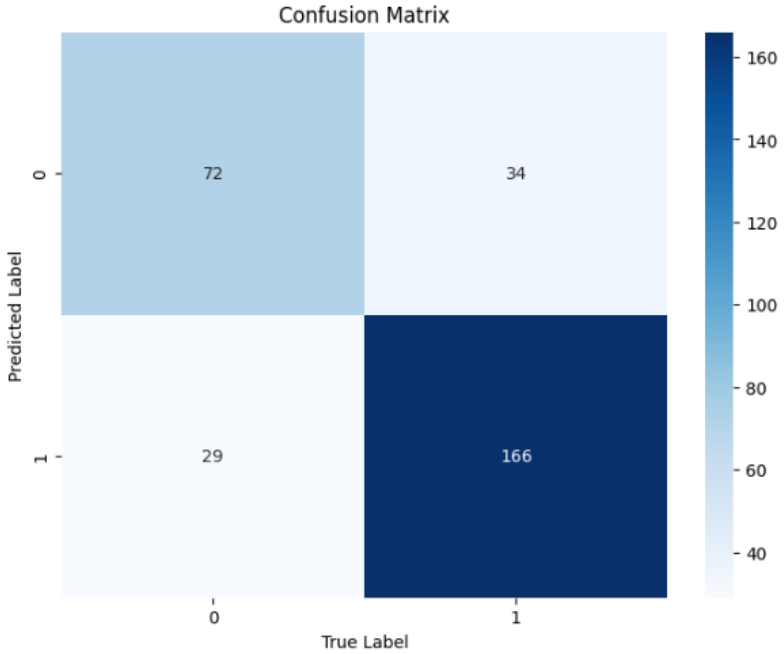

Figure 3. Confusion Matrix

The confusion matrix is a table used to evaluate the performance of a classification model. The confusion matrix displays the number of correctly and incorrectly classified data based on the actual and predicted classes. This table usually contains four cells, namely True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN), True Positive (TP): The number of data correctly classified as positive (positive actual class, positive predicted class). False Negative (FN): The number of data that is incorrectly classified as negative (positive actual class, negative predicted class). False Positive (FP): Number of data incorrectly classified as positive (negative actual class, positive predicted class). True Negative (TN): The amount of data that is correctly classified as negative (negative actual class, negative predictive class), can be seen from the picture above for the True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) values are 72, 34, 29, and 166.

To visualize the *decision tree* results with ASM(Attribute Selection Measure) *entropy* can be seen in the following figure 4.

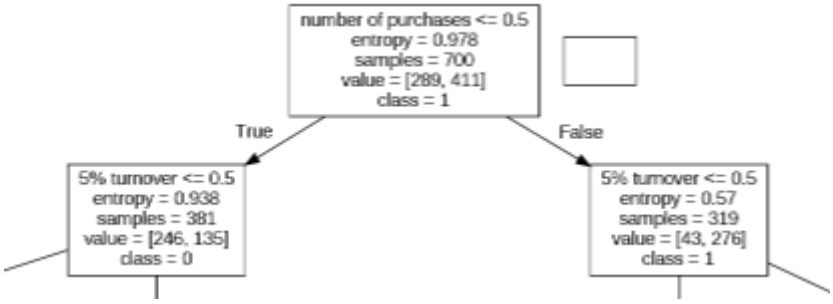

Figure 4. Visualization root node decision tree results

The image above is the result of a root node or parent node visualization in a *decision tree*, which is used to divide data based on the best features that will provide the most significant separation between different classes or targets.

According to the results of the visualization, it can be seen that the variable number of purchases with a value of 1 or worthy of being given a special offer is chosen as the main separator because it has an entroypy value of 0.978, the greater the value, the means that the variable has a high uncertainty value in the selection of the *decision tree*, so *decision tree* will do the separation to reduce the entropy value.

This an explanation of why the entropy value of the number of purchases variable is 0.978, known from the visualization of the *decision tree*, The total number of samples = 700 the number of samples in class 1 = 41, the number of samples in class 0 = 289, calculate the proportion of each class proportion of the sample in class 1 = 411 / 700 ≈ 0.587 Proportion of the sample in class 0 = 289 / 700 ≈ 0.413, so that based on the known entropy formula, we will calculate the final result

$$\begin{aligned} Entropy(s) &= (-0.587 * log_2(0.587)) + (-0.413 * log_2(0.413)) \\ &= (-0.587 * -0.7686) + (-0.413 * -1.2758) \\ Entropy(s) &= 0.452 + 0.526 \\ Entropy(s) &= 0.978 \end{aligned}$$

The following entropy calculations will be carried out continuously for each selected variable to produce a small entropy value.

Below is an example of a *decision tree* result to predict which customers qualify for a special offer.

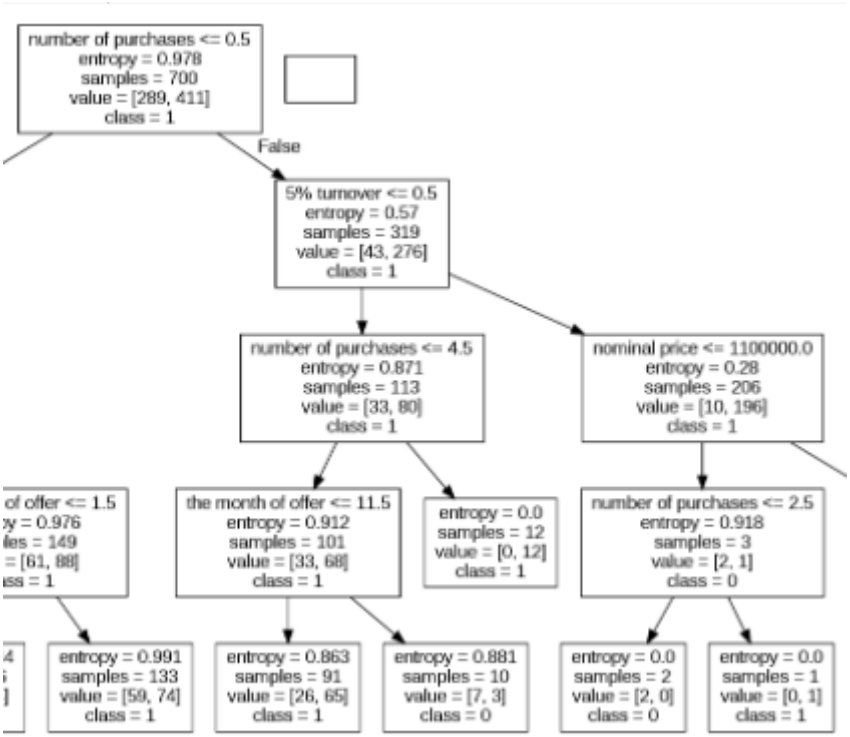

Figure 5. Example visualization decision tree prediction

The results of the following *decision tree* show that the number of purchases variable with a value below 0.5 will be classified as class 1, followed by if the number of purchases variable has a value above 0.5 it will proceed to the calculation of the 5% turnover variable if the value is below 0.5 then classified as class 1, if the variable is correct below equal to 0.5 then it will proceed to the calculation of the number of purchases variable below equal to 4.5 then it is classified as class 1, if yes then the prospective customer will be categorized as class 1 because the final entropy calculation result is 0, which means no uncertainty or variation.
Below is the complete result of the *decision tree* model:

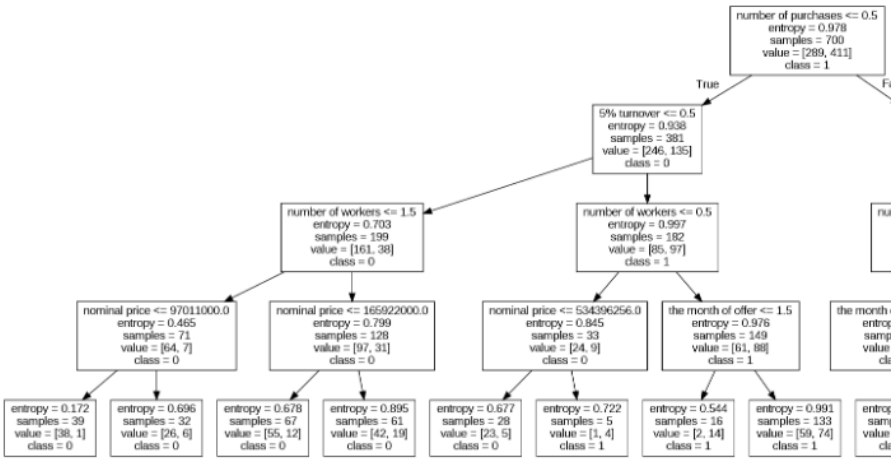

Figure 6. Visualization decision tree part 1

For the *decision tree* results from figure 6 is a true statement from the prediction results, in the figure given it will be classified whether the number of purchases variable is less or equal to 0.5, if not it will enter the *decision tree* classification on the right which means true, likewise for nodes after where if the node is on the left it means true, otherwise it means false.

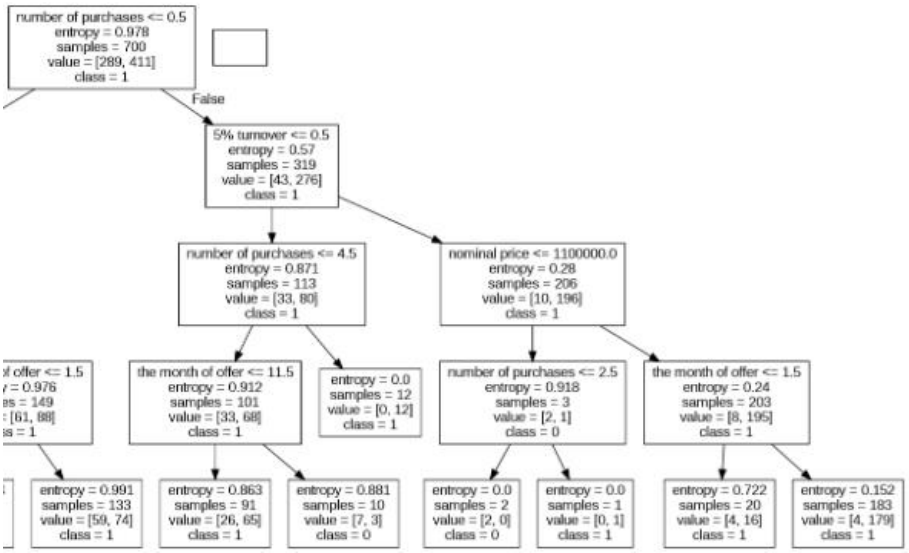

Figure 7. Visualization decision tree part 2

For part figure 7 it means that the *decision tree* classification has a false value from the separator node. It can be seen that if the variable number of purchases is not less than or equal to 0.5 then it will enter the separator node on the right.

For the calculation of each entropy value for each variable according to the formula and examples that have been given.

CONCLUSION

Based on the use of the *Decision tree* method, results were obtained with an accuracy of 0.7. In this case, the confusion matrix shows the following results: there are 72 True Positive (TP) which is a true positive classification, 34 False Negative (FN) which is a wrong negative classification, 29 False Positive (FP) which is a wrong positive classification , and 166 True Negative (TN) which is a true negative classification.

From these results, it can be concluded that the *Decision tree* model has a fairly good level of accuracy with a fairly good ability to correctly classify positive data and negative

data. However, it should be noted that there were cases that were misclassified as either false negatives or false positives.

To improve model performance, it is necessary to carry out further evaluation and parameter adjustments as well as more careful data handling. Although an accuracy of 0.7 indicates decent performance, there is still room for improvement in reducing misclassification.

BIBLIOGRAPHY

Adhinata, K. B. (2019). *Implementasi Algoritma Decision tree Classifier Untuk Klasifikasi Pelanggan Provider X Pada E-commerce Sepulsa*. Surabaya, Indonesia: Universitas Ciputra.

Aguilar-Chinea, R. M., Castilla Rodriguez, I., Exposito, C., Melian-Batista, B., & Moreno-Vega, J. M. (2018). *Using a decision tree algorithm to predict the robustness of a transshipment schedule*. La Laguna, Spain: Universidad de La Laguna.

Alleyda, I. S., Ballya, V. H., Endah, P., Yohanne, M. S., Andhika, W. W., & Desta, S. P. (2021). *Klasifikasi Data Penjualan pada Supermarket dengan Metode Decision tree*. Jakarta, Indonesia: Universitas Pembangunan Nasional "Veteran" Jakarta.

Arfandi, A. P. W., Ilham, S. S., & Saragih, I. S. (2021). *Penerapan Data Mining Klasifikasi Pada Calon Pelanggan Baru Indihome dengan C.45*. Pematang Siantar, Indonesia: STIKOM Tunas Bangsa.

Basri, W. G., Risnandar, & Nusa Mandiri, U. (2020). *Analisis Loyalitas Pelanggan Berbasis Model Recency, Frequency, Dan Monetary (RFM) Dan Decision tree Pada PT. SOLO*. Jakarta, Indonesia: Universitas Nusa Mandiri Kampus Kramat Raya.

Febriyani, A., Prayoga, G. K., & Nurdiawan, O. (2021). *Index Kepuasan Pelanggan Informa dengan Menggunakan Algoritma C.45*. Cirebon, Indonesia: STMIK IKMI.

Hikmatulloh, R., Mahaerani P., H., & Aini, Q. (2020). *Penerapan Decision tree untuk Prediksi Kepuasan Pengguna Bus Transjakarta*. Banten, Indonesia: Universitas Islam Negeri Syarif Hidayatullah.

Latifah, R., Wulandari, E. S., & Kreshna, P. E. (2019). *Model Decision tree untuk Prediksi Jadwal Kerja menggunakan Scikit-Learn*. Jakarta, Indonesia: Universitas Muhammadiyah Jakarta.

Mendez, M. (2012). *Sales promotions effects on brand loyalty*. Florida, United States of America: Nova Southeastern University.

Musthofa, G. P., & Saputro, P. H. (2020). *Komparasi Metode Naïve Bayes dan C4.5 Dalam Klasifikasi Loyalitas Pelanggan Terhadap Layanan Perusahaan*. Yogyakarta, Indonesia: Universitas Alma Ata Yogyakarta.

Nasrullah, A. H. (2021). *Implementasi Algoritma Decision tree untuk Klasifikasi Produk Laris*. Gorontalo, Indonesia: Universitas Ichsan Gorontalo.

Oktaviani, A., Cornelia, R., & Wibisono, D. (2022). *Peningkatan Loyalitas Pelanggan pada PT Home Center Indonesia Menggunakan Metode Algoritma C4.5 dan Metode CSI (Customer Satisfaction Index)*. Jakarta, Indonesia: Universitas Indraprasta PGRI.

Satyanarayana, N., Ramalingaswamy, CH., & Ramadevi, Y. (2014). *Survey of Classification Techniques in Data Mining*. Hyderabad, India: CVR College of Engineering.

Solehuddin, M., Syafei, W. A., & Gernowo, R. (2022). *Metode Decision tree untuk Meningkatkan Kualitas Rencana Pembelajaran dengan Algoritma C4.5*. Semarang, Indonesia: Universitas Diponegoro.

Susanto, H., & Sudiyatno. (2014). *Data Mining untuk Memprediksi Prestasi Siswa Berdasarkan Sosial Ekonomi, Motivasi, Kedisiplinan, dan Prestasi Masa Lalu*. Surakarta, Indonesia: SMK Negeri 4 Surakarta.

Wardani, N. W., & Ariasih, N. K. (2019). *Analisa Komparasi Algoritma Decision tree C4.5 dan Naïve Bayes untuk Prediksi Churn Berdasarkan Kelas Pelanggan Retail*. Singaraja, Indonesia: Universitas Pendidikan Ganesha.

Wardani, N. W., Dantes, G. R., & Indrawan, G. (2018). *Prediksi Customer Churn Dengan Algoritma Decision tree C4.5 Berdasarkan Segmentasi Pelanggan Pada Perusahaan Retail*. Jurnal Resistor, 1(1), 1-10.

Yusuf, F. A., Alfaridzi, M., & Herdi, T. (2022). Penerapan Algoritma Decision Tree Untuk Klasifikasi KIPI Vaksin Covid-19. *Jurnal Ilmiah FIFO*, *14*(2), 155. https://doi.org/10.22441/fifo.2022.v14i2.005