COMPILER
Vol. 12, No. 2, November 2023, pp. 103~108
ISSN: 2252-2403 (Print), ISSN: 2549-2403 (Online)
Accredited Fourth Grade by Kemenristekdikti, Decree No:200/M/KPT/2020
DOI: 10.28989/compiler.v12i2.1911

Implementation of the Decission Tree Algorithm to Determine Credit Worthiness

Abdussomad[1], Ilham Kurniawan[2], Agung Wibowo[3*]
[1,2,3]Information System, Universitas Bina Sarana Informatika, Indonesia

Article Info

Article history:

Received October 23, 2023
Accepted November 20, 2023
Published November 30, 2023

Keywords:

Credit analysis
Decision tree models
Decision tree algorithms

ABSTRACT

Credit is a loan from a bank that needs to be repaid with interest. In practice, problematic credit or bad credit often occurs due to less thorough credit analysis in the credit granting process, or from bad customers. This research aims to predict creditworthiness using the Decision Tree Classification Algorithm and find a solution for determining it. This research uses the CRISP-DM (Cross-Industry Standard Process for Data Mining) method. This research method tests the effects of using the decision tree, Support Vector Machine, and Naïve Bayes model with the Decision Tree Classification Algorithm. The decision tree classification algorithm accurately analyzed problem loans and non-problem debtors at 93.49%. The decision tree algorithm test results are better than the support vector machine by 3.45%, and naïve bayes by 13.03%. The results of our study were also 4.16% better than the previous study. This research has also implemented the selected model in the form of website application deployment.

Corresponding Author:

Agung Wibowo,
Information System
Universitas Bina Sarana Informatika,
Jl. Kramat Raya No.98, Senen Jakarta, Indonesia.
Email: *agung.awo@bsi.ac.id

1. INTRODUCTION

Personal credit has become an important credit indicator for individuals in modern economic society [1]. Individuals in modern economic society consider personal credit as an important credit indicator. Credit scoring assesses the risks of credit products using historical data and statistical or machine learning techniques [2][3][4]. The main focus of credit assessment is to determine whether a credit applicant has a place in the creditworthy or non-creditworthy group. The credit assessment process is not a one-step process; periodically, financial institutions do it in various steps, such as application assessment, behavioral assessment, collection assessment, etc.[4]. When a credit applicant applies for a loan, financial institutions collect information from the applicant. This information is called application data and consists of demographic information, for example, the number of dependents, current address, current employment, etc. Bureau information is also collected from local bureaus and includes the number of inquiries, assessments, amounts outstanding, etc. Once the accepted population, i.e., credit applicants who have been granted a loan, is identified, their loan repayment history is tracked for a certain period, e.g., 24 months [5]. To overcome this problem, it is necessary to carry out a credit analysis. A credit analysis is a study carried out to determine the feasibility of a credit problem. Through the results of credit analysis, it can be seen whether the customer's business is feasible, marketable (business results can be marketed), profitable, and can be repaid on time [6]. The company can use historical data on approved debtors as a benchmark for approving or rejecting a debtor. However, we should also note that being approved does not guarantee that all debtors are good credit-payers. The company has approved some debtors, but their payments are in arrears several months later. So we need proper analysis to determine the creditworthiness of customers who will apply for credit. One way to reduce bad credit cases is by using machine learning, specifically the Decision Tree Algorithm, to predict future bad credit. We can implement this algorithm on an application or website.

2. RESEARCH METHOD

The research method used in this study is divided into six stages as follows:

1. Business Understanding

Initially, we assess business goals and requirements to determine machine learning problem areas.

2. Data Understanding

During the Data Understanding stage, we collect and analyze data to identify information and assess its quality. This helps you figure out the percentage of data we'll analyze, including the amount to retrieve and the data that's stuck or running smoothly [7].

3. Data Preparation

At this stage, all the activities involve creating the final dataset, which we will include in the model we are creating. Data preparation involves all the activities of building a data set for a model [8].

4. Modelling

During the modeling process stage, the author initiates the selection of the model to be used in the experiment. This stage will look at or review past literature and identify prediction models that were commonly used previously [9].

5. Evaluation

We evaluate the model at this stage, looking back at previous steps to ensure it meets our business goals. In the evaluation stage, the confusion matrix is used to determine the accuracy value of the resulting model or algorithm test results. The following is the level of accuracy in the confusion matrix [10]: Accuracy 0.90–1.00 = Excellent classification, Accuracy 0.80–0.90 = Good classification, Accuracy 0.70–0.80 = Fair classification, Accuracy 0.60–0.70 = Poor classification and Accuracy 0.50–0.60 = Failure

6. Deployment

Model deployment (Deployment) is the final and most challenging stage of the machine learning life cycle. This is because we will implement the model from the test results in real applications, using real data in the field and depicted in figure 1. Research method as follow.

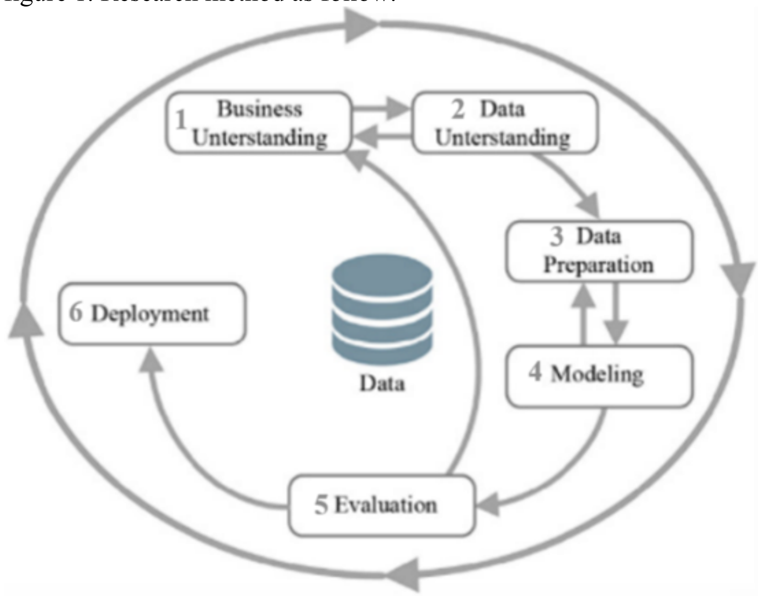

Figure 1. Research method [11]

3. RESULTS AND ANALYSIS

The implementation of the decision tree algorithm in this research was carried out through stages by the previously selected methodology. First, the dataset was preprocessed to handle missing values and outliers. Then, feature selection techniques were applied to identify the most relevant variables for building the decision tree. Next, the decision tree was constructed using an appropriate algorithm, such as ID3 or C4.5. Finally, the performance of the decision tree model was evaluated through various metrics, such as accuracy, precision, recall, and F1-score. Overall, this systematic approach ensured a robust and reliable implementation of the decision tree algorithm in the research study methodology. First, the dataset was preprocessed to handle missing values and outliers. Then, feature selection techniques were applied to identify the most relevant variables for building the decision tree. Next, the decision tree was constructed using an appropriate algorithm, such as ID3 or C4.5. Finally, the performance of the decision tree model was evaluated through various metrics, such as accuracy, precision, recall, and F1-score. Overall, this systematic approach ensured a robust and reliable implementation of the decision tree algorithm in the research study. the CRISP-DM model, with the following:

3.1. Business Understanding

Based on the condition of credit customers at one of the leasing companies in Karawang Regency in 2015, the percentage of customers with bad credit status was higher than that of customers with current status. The higher data on problematic customers shows a lack of accurate analysis in determining the feasibility of providing credit to consumers.

3.2. Data Understanding

The data used in this research came from a leasing company in Karawang City in 2015. The amount of data used was 1044 records, with 11 predictor attributes and 1 as a label. See table 1. Data credits

Table 1. Data credits

No_kontrak	Kecamatan	Kabupaten	Status	Pekerjaan	Object	Dp_net	Otr	Tenor	Area	Angs_bln	Kondisi
1	Kadungwaringin	Kab. Bekasi	Pemohon tunggal	Wiraswasta non formal	Motor bekas	3122000	9500000	12	Karawang utara	648000	Lancar
2	Kadungwaringin	Kab. Bekasi	Penjamin	Peg.swasta formal	Motor bekas	4108600	11100000	12	Karawang utara	705000	Lancar
3	Pebayuran	Kab. Bekasi	Penjamin	Peg.swasta non formal	Motor bekas	3525000	11100000	35	Rengasdengklok barat	415000	Lancar
4	Talagasari	Kab. Karawang	Pemohon tunggal	Wiraswasta non formal	Motor baru	3300000	16050000	33	Karawang timur	710000	Lancar
5	Kutawaluya	Kab. Karawang	Pemohon tunggal	Peg.swasta formal	Motor baru	4700000	21150000	32	Rengasdengklok timur	872000	Lancar

Based on the data used, the author visualized current and bad conditions with the results that the level of credit congestion is higher than the current credit condition. The data shows that there are 620 problem customers (59.4%) and 424 data of non-problem customers (59.4%) Current) (40.6 %) as seen at Figure 1 Visualization of credit conditions.

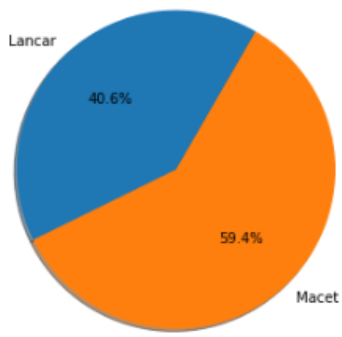

Figure 1. Visualization of credit conditions

3.3. Data Preparation

At this stage the author prepares the data before entering it into the model, several stages carried out in the data preparation are:

a. Data validation

Data validation was performed to identify and remove outliers, inconsistent data, and attributes that were not descriptive of the credit conditions. As a result, the dataset was reduced to 8 predictor attributes: Status, Job, Object, Dp_Net, Otr, Tenor, Area, and Installment, describes in Table 2. Data validation.

Table 2. Data validation

Status	Pekerjaan	Object	Dp_net	Otr	Tenor	Area	Angs_bln	Kondisi
Pemohon tunggal	Wiraswasta non formal	Motor bekas	3122000	9500000	12	Karawang utara	648000	Lancar
Penjamin	Peg.swasta formal	Motor bekas	4108600	11100000	12	Karawang utara	705000	Lancar
Penjamin	Peg.swasta non formal	Motor bekas	3525000	11100000	35	Rengasdengklok barat	415000	Lancar
Pemohon tunggal	Wiraswasta non formal	Motor baru	3300000	16050000	33	Karawang timur	710000	Lancar
Pemohon tunggal	Peg.swasta formal	Motor baru	4700000	21150000	32	Rengasdengklok timur	872000	Lancar
Penjamin	Wiraswasta non formal	Motor bekas	6497600	11100000	18	Rengasdengklok timur	337000	Lancar
Pemohon tunggal	Wiraswasta non formal	Motor baru	5000000	23275000	26	Rengasdengklok timur	1096000	Lancar
Pemohon tunggal	Peg.swasta formal	Motor bekas	2993000	10500000	18	Karawang utara	554000	Lancar

b. Missing Value Checking

Missing value checking was conducted, and it was found that there were no missing values in the dataset.

c. Data size reduction and categorization

Data size reduction and categorization were performed to transform continuous attributes into nominal values. This process, the author creates several categorical attributes with the following at Table 3. Data type conversion.

Table 3. Data type convertion

STATUS	PEKERJAAN	OBJECT	DP_NET	OTR	TENOR	AREA	ANGS_BLN	KONDISI
1	5	1	5	1	1	1	1	1
2	2	1	7	1	1	1	1	1
2	3	1	5	1	2	5	1	1
1	5	2	5	3	2	2	1	1
1	2	2	7	4	2	3	1	1
2	5	1	7	1	1	3	1	1
1	5	2	7	4	2	3	1	1
1	2	1	5	1	1	1	1	1

d. Data integration and transformation

to improve the accuracy and efficiency of the algorithm. The author carries out feature selection to increase accuracy in the model to be built, namely using feature selection (Feature Importance) with the following results:

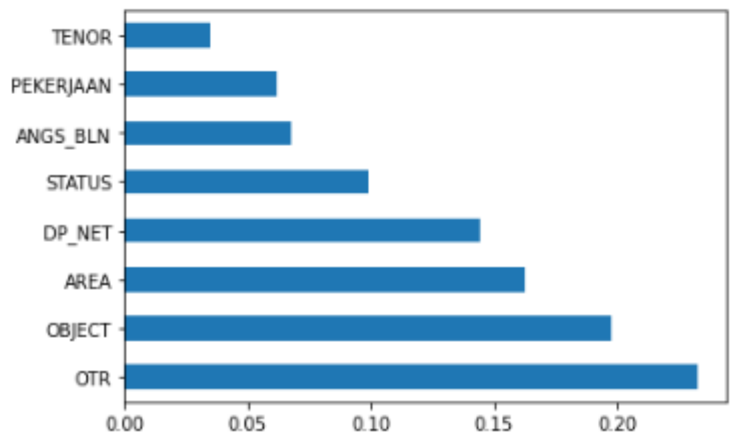

Figure 2. Feature Importance

In Figure 2 above, the process resulted in the creation of several categorical attributes to represent the problem of bad credit. For example, the tenor attribute, which had the lowest feature relevance for class attributes, was removed from the dataset to improve the accuracy of the model.

3.4. Modelling

The model used by the author in this research was the decision tree (DT) classification algorithm, Support Vector Machine and Naïve Bayes. Researchers used the decision tree (DT) algorithm to solve prediction problems in determining creditworthiness because it was proven to work well in previous research conducted by [12] by comparing several classification algorithms including k-nearest neighbor, naive Bayes and decision trees. It is proven that the decision tree classification algorithm outperforms other classification algorithms with an accuracy value of 98.00%. Based on the total existing data, namely 1044, the author divides the data into a). Training Data (75%) = 783 and b). Testing Data (25%) = 261. Testing was carried out using the 10-fold cross-validation technique against the decision tree (DT) classification algorithm, resulting in a confusion matrix. Confusion matrix is a method that can be used to measure the performance of a classification method [10]. Basically, the confusion matrix contains information that compares the classification results carried out by the system with the classification results.

3.5. Evaluation

The evaluation was carried out to determine the accuracy value of the model used. The author used K-fold 10-Cross Validation, a model evaluation technique that is quite popular and widely used. The author chose 10 for the K value. This means that for each iteration, the K-fold algorithm evaluates the model using 10% test data and 90% training data. Based on the TN, FP, FN, and TP values in the evaluation matrix table for measuring the classification algorithm (accuracy, precision, recall, and F1-score) from the model used, namely the decision tree, the results are as follows:

$$Accuracy = (TP+TN)/(TP+TN+FP+FN) \times 100\% \quad (1)$$
$$Precision = TP/(FP+TP) \times 100\% \quad (2)$$
$$Recall = TP/(FN+TP) \times 100\% \quad (3)$$
$$F1\text{-}Score = (2 \times Recall \times Precision)/(Recall+Precision) \quad (4)$$

Table 4. Confusion matrix

Algorithms	True Positive	True Negative	False Positive	False Negative
Decission Tree	142	102	7	10
Support Vector Machine	144	91	18	8
Naive Bayes	124	86	23	28

Table 5. Evaluation result

Algorithms	Accuracy	Precision	Recall	F1-Score
Decission Tree	0,935	0,953	0,934	0,943
Support Vector Machine	0,900	0,888	0,947	0,917
Naive Bayes	0,805	0,843	0,815	0,829

Table 5 above shows that the evaluation results of the decision tree model used have an accuracy value of 0.930 (93%) and are included in the Excellent Classification category. The F1-Score describes the comparison between the average precision and recall values. If the dataset has closeness between false negatives and false positives (symmetric), we choose to use accuracy as a reference for algorithm performance. However, if the numbers are not close, then you have to use the F1-Score as a reference. Based on table 4. Confusion matrix above, it shows that the Decision Tree algorithm is considered good at classifying and predicting credit worthiness with an ROC (accuracy) value close to 1, namely 0.930. Next, the author deploys the implementation of the results of the decision tree algorithm model using Flask Python.

3.6. Deployments

Next, the author deploys the implementation of the results of the decision tree algorithm model using Flask Python, with results like Figure 3 below:

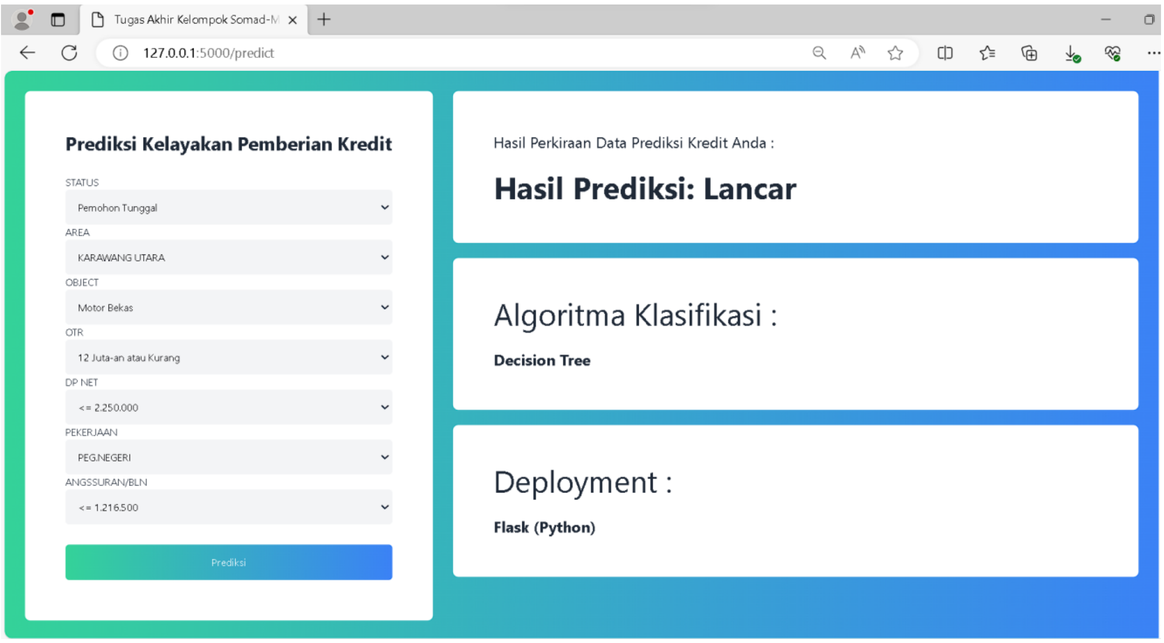

Figure 3. Deployment model.

4. CONCLUSION

The decision tree algorithm is more accurate than the analysis carried out by credit analysts, with research evaluation results showing that the decision tree classification algorithm can analyze problem loans and non-problem debtors, which produces an accuracy value of 93.00%. Compared to support vector machine and naïve bayes algorithms, this algorithm model is 3.45% to 13.3% better. Our study also had an increase in accuracy of 4.16% compared to the results of the previous study [SIAPA].

This research only uses the decision tree classification algorithm for the implemented model website. It is hoped that for further research, other methods can be used so that comparisons can be made. For better accuracy values, you should use optimization methods such as PSO (particle swarm optimization), GA (genetic algorithm), and others.

REFERENCES

[1] J. Zhang, R. Tan, dan Y. dong Li, “Design of Personal Credit Information Sharing Platform Based on Consortium Blockchain,” *Commun. Comput. Inf. Sci.*, vol. 1286, no. 2019, hal. 166–177, 2020.

[2] R. A. Mancisidor, M. Kampffmeyer, K. Aas, dan R. Jenssen, “Knowledge-Based Systems Deep generative models for reject inference in credit scoring ☆,” *Knowledge-Based Syst.*, vol. 196, hal. 105758, 2020.

[3] D. Tripathi, R. Cheruku, dan A. Bablani, “Relative Performance Evaluation of Ensemble Classification with Feature Reduction in Credit Scoring Datasets,” *Springer*, vol. 705, no. January 2018, hal. 201–211, 2018.

[4] D. Tripathi, D. R. Edla, V. Kuppili, dan A. Bablani, “Evolutionary Extreme Learning Machine with novel activation function for credit scoring,” *Eng. Appl. Artif. Intell.*, vol. 96, no. February, hal. 103980, 2020.

[5] X. Dastile, T. Celik, dan M. Potsane, “Statistical and machine learning models in credit scoring: A systematic literature survey,” *Appl. Soft Comput. J.*, vol. 91, hal. 106263, 2020.

[6] D. O. Cardoso *et al.*, “Author ’ s Accepted Manuscript Financial credit analysis via a clustering weightless neural classifier,” *Neurocomputing*, 2015.

[7] Z. A. Andriawan *et al.*, “Prediction of Hotel Booking Cancellation using CRISP-DM,” *ICICoS 2020 - Proceeding 4th Int. Conf. Informatics Comput. Sci.*, hal. 0–5, 2020.

[8] C. Gonçalves, D. Ferreira, C. Neto, A. Abelha, dan J. Machado, “Prediction of mental illness associated with unemployment using data mining,” *Procedia Comput. Sci.*, vol. 177, hal. 556–561, 2020.

[9] F. Martinez-Plumed *et al.*, “CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories,” *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, hal. 3048–3061, 2021.

[10] P. Darussalam Adhis, “Jurnal Ilmiah Wahana Pendidikan,” *J. Ilm. Wahana Pendidik. https//jurnal.unibrah.ac.id/index.php/JIWP*, vol. 7, no. 1, hal. 391–402, 2021.

[11] S. Huber, H. Wiemer, D. Schneider, dan S. Ihlenfeldt, “DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model,” *Procedia CIRP*, vol. 79, hal. 403–408, 2019.

[12] T. T. Muryono, A. Taufik, dan I. Irwansyah, “Perbandingan Algoritma K-Nearest Neighbor, Decision Tree, Dan Naive Bayes Untuk Menentukan Kelayakan Pemberian Kredit,” *Infotech J. Technol. Inf.*, vol. 7, no. 1, hal. 35–40, 2021.