Implementation of Data Mining to Predict the Eligibility Level for Prospective KPR (Home Ownership Credit) Subsidized Housing Customers Mitra Griya Indah Using the C4.5 Algorithm

Mia Anggraini [1], Fahmi Ruziq [2], Roy Nuary Singarimbun [3]
[123] Univesitas Battuta, Medan, Indonesia
[123] Fakultas Teknologi, Univesitas Battuta, Medan, Indonesia
[1] wahyukren787@gmail.com [2] fahmiruziq89@gmail.com [3] roy90singarimbun@gmail.com

Article Info

Article history:

Received April 24, 2024
Revised May 01, 2024
Accepted May 10, 2024

Keywords:

Data Mining
C4.5 Algorithm
Home Ownership Credit

ABSTRACT

In the housing industry, data mining plays an important role in assisting the home loan application process by extracting knowledge from historical data, this process allows lenders to identify potentially high-risk home loan applicants and decide whether to approve or reject the loan application. Data mining helps in effective marketing strategies. By optimizing this process, response time to home loan applications can be accelerated, operational efficiency increased, and credit risk can be better managed. In the practice of providing KPR (Home Ownership Credit) to prospective consumers, there are possible problems that will occur like most other people, namely late installment payments or defaulted payments so that it will make it difficult for the bank to maintain the level of credit risk on the credit provided, this is because Mitra Griya Indah Housing has not paid much attention to data regarding the history of credit granting decisions, in other words, it has not maximally utilized data on previous credit granting decisions in supporting credit granting decisions. To solve this problem, the researcher designed a calculation information system. In this case the author uses the waterfall method in the research process. For system design, the author uses the PHP programming language with a database format using MySql. Finally, with this information system, it can facilitate the decision-making process for prospective customers of Home Ownership Credit. [1]

Corresponding Author:

Mia Anggraini
Universitas Battuta
Email: wahyukren787@gmail.com

1. INTRODUCTION

In the housing industry, data mining plays an important role in assisting the home loan application process by extracting knowledge from historical data, this process allows lenders to identify potentially high-risk home loan applicants and decide whether to approve or reject the loan application. By optimizing this process, response time to home loan applications can be accelerated, operational efficiency increased, and credit risk can be better managed. Mitra Griya Indah Housing is a company engaged in residential property in Medan city located in Hamparan Perak District, Deli Serdang Regency. In practice, the provision of KPR (Home Ownership Credit) to prospective consumers there will be possible problems that will occur like most other people, namely late installment payments or defaulted payments so that it will make it difficult for the bank to maintain the level of credit risk on the credit provided, this is because Mitra Griya Indah Housing has not paid much attention to data regarding the history of credit granting decisions, in other words, it has not maximally utilized data on previous credit granting decisions in supporting credit granting decisions. [2] In addition, the bank will assess the performance in finding the right consumer to be a bad image for Mitra Griya Indah Housing. Therefore, the researcher has the intention to carry out an analysis process of credit granting decision data, namely by classifying prospective debtors based on the criteria adopted based on aspects of credit assessment in general, then further using a suitable method to carry out the classification process, namely the data mining classification technique by building a decision tree using the C4.5 algorithm which will later produce results in the form of hidden rules or rules in determining the feasibility of receiving credit loans which can later be used to improve the quality of credit analysis results and as evaluation material for determining future creditworthiness. [3]

According to Marvin Kristianto's research entitled "Classification of Potential Customer Credit Payments with the C4.5 Method at Pt. Autochem Industry" concluded that by using data analysis using the C45 algorithm, it can be concluded that the results of prediction accuracy to prevent bad customer credit, reduce bad debts, and as a result of making customer credit decisions cannot be used or used as a benchmark. Based on the root of the problem, research was conducted that could produce an information system that could help improve the accuracy and quality of credit analysis to help determine the feasibility of granting credit to prospective customers. [4]

2. METHOD

In this study the data used is secondary data or data derived from journals, papers, books and some other information related to the research cited. [5] Researchers also use several techniques to collect data, including the following; a. Literature Study, A writing method carried out to obtain data and information by reading various writing materials, and scientific essays on issues related to writing. b. Interview, Researchers conducted interviews with company directors to obtain the information needed as material for writing reports. [6] From the results of interviews in this study, the results of data that can be used as material for research needs are as follows; [7] [8]

Table I. Customer Data

Gender	Marriage status	Dependents	Income	Age	Results
Male	Marriage	Have income	4 Million	24	Current
Male	Marriage	Have income	5 Million	30	Current
Male	Marriage	No revenue	5,5 Million	41	Current
Male	single	No revenue	3,9 Million	26	Current
Male	Marriage	No revenue	4,6 Million	32	Current
Male	Marriage	No revenue	4,1 Million	35	Current
Male	Marriage	No revenue	3,7 Million	27	Current
Female	single	Have income	4 Million	34	Current
Female	single	Have income	5,5 Million	23	Current
Male	Marriage	Have income	6 Million	31	Current
Female	Marriage	No revenue	4,7 Million	42	Current
Male	Marriage	No revenue	5,8 Million	27	Current
Male	Marriage	No revenue	4,4 Million	24	Current
Male	Marriage	No revenue	5,3 Million	29	Current
Male	Marriage	No revenue	3,6 Million	24	Current
Female	Marriage	Have income	3,5 Million	45	Stuck
Female	Marriage	Have income	4,3 Million	41	Stuck
Male	Marriage	No revenue	4,4 Million	36	Current
Female	Marriage	No revenue	6 Million	38	Current
Male	Marriage	No revenue	3,8 Million	27	Stuck

Male	Marriage	No revenue	3,9 Million	29	Current
Female	single	Have income	3,8 Million	27	Stuck
Male	Marriage	No revenue	4,3 Million	26	Current
Male	Marriage	No revenue	4,7 Million	24	Current
Male	Marriage	No revenue	3,6 Million	24	Current
Female	Marriage	Have income	4 Million	42	Stuck
Male	single	Have income	3,9 Million	27	Stuck
Male	Marriage	Have income	3,6 Million	46	Stuck
Male	Marriage	Have income	4,2 Million	45	Stuck
Male	Marriage	No revenue	3,6 Million	27	Stuck

The term data mining has several equivalents, such as knowledge discovery or pattern recognition. The term knowledge discovery is appropriate to use because the main purpose of data mining is to get the knowledge that is still hidden in the data lumps. The term pattern recognition is also appropriate to use because the knowledge to be extracted is in the form of patterns that may also still need to be extracted from the chunks of data at hand. [9] Data mining is the process of obtaining useful information from large databases and needs to be extracted to become new information and can help in decision-making. One of the techniques created in data mining is how to explore existing data to build a model, then use the model to recognize other data patterns that are not in the stored database. The need for prediction can also utilize this technique. In data mining, data clustering can also be done. [10] [11]

There are three main steps in the data mining process systematically, including data exploration or preprocessing which consists of data cleaning, data normalization, data transformation, handling incorrect data, dimension reduction, feature subset selection and others. The second process is to build a model and validate it by analyzing various models and selecting the model with the best prediction performance. In this step, methods such as classification, regression, cluster analysis, anomaly detection and so on are used. The third or final step is application, which means applying a model to new data to produce predictions of the problem being investigated. The training methods used in data mining techniques can be divided into two approaches, namely: a. Unsupervised learning, this method is applied without training and without a teacher. The teacher here is the label of the data. b. Supervised learning, which is a learning method with training and trainers. [12]

Data mining analyzes data using tools to find patterns and rules in data sets. The software is tasked with finding patterns by identifying rules and features in the data. Data mining tools are expected to recognize these patterns in the data with minimal input from the user. Classification method, Classification is one of the most common learning in data mining. Classification is defined as a form of data analysis to extract a model that will be used to predict class labels. Classification is divided into two stages, namely the learning stage and the classification model formation stage. Decision three is used for decision making. Decision three will find a solution to the problem by making the criteria as interconnected nodes like the root of a tree. Decision three is a prediction model for a decision using a hierarchical or tree structure. Each tree has branches, branches that represent as attributes that must be met to go to the next branch so that it ends in the leaf. The concept of data in decision three is data expressed in tabular form consisting of attributes and records. Decision tree is one of the methods to classify data. The decision tree model is a tree consisting of root nodes. While root nodes and internal nodes are variables/features, terminal nodes are class labels. In performing classification, a query data will browse the root node and internal nodes until the terminal node. The class labeling of the query data is based on the labels in the internal nodes. In traditional decision trees, the data used is data with definite feature values. [13]

The selection of attributes as nodes, either root nodes or internal nodes is based on the highest Gain value of the existing attributes. The Gain value calculation uses the formula as in Equation 1.

$$\textbf{Gain (S,A) = Entropy(S)} - \sum_{i=1}^{n} * \boldsymbol{Entropy(Si)}$$

S : Case set

A : Attribute n: Number of partitions of attribute A

|Si| Number of cases in the i-th partition

|S| : Number of cases in S

To calculate the Entropy value can be seen in Equation 2

$$\textbf{Entropy(S) = } \sum_{i=1}^{n} - \textbf{pi x log}^{2}\textbf{ pi}$$

n : Number of partitions S
pi : The proportion of Si to S

The system development method that the author uses is the waterfall method, the waterfall method describes the flow of system development in order. The stages contained in the research method with waterfall use structured analysis and design. Waterfall is a system development technique that is interconnected between one process and another. These processes will be explained as follows; [14]

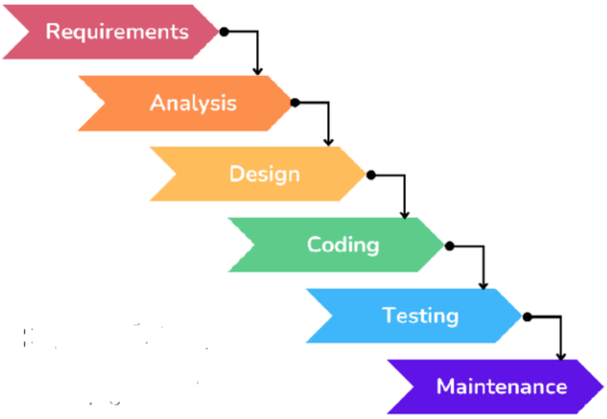


Figure I. Waterfall Method

3. RESULTS AND DISCUSSION

Data Mining Calculations, Steps to determine the decision tree using the C4.5 algorithm using 30 training data, namely:a. Prepare training data as much as 30 data used in this study. Primary data that had previously been grouped into their respective classes. b. Calculate the entropy and gain values after calculating entropy and gain, the following values are obtained: To calculate entropy, the formula used is:

Entropy(S) $= \sum_{i=1}^{n} -$ **pi x log² pi** Results: $Entropy\ S = \left(-\left(\frac{21}{30}\right) * log_2\left(\frac{21}{30}\right)\right) + \left(-\left(\frac{9}{30}\right) * log_2\left(\frac{9}{30}\right)\right) =$ **0.881290899**

The age attribute has a value of 23-30 years there are 17 cases consisting of 13 "Current" and 4 "Stalled", 31-40 years have 6 consisting of 6 "Current" and 0 "Stalled", 41-50 years have 7 consisting of 2 "Current" and 5 "Stalled" then calculate entropy.

(23-30 years old) $=\left(-\left(\frac{13}{17}\right) * log_2\left(\frac{13}{17}\right)\right) + \left(-\left(\frac{4}{17}\right) * log_2\left(\frac{4}{17}\right)\right) = \mathbf{0.787126586}$

(31-40 years old) $= \left(-\left(\frac{6}{6}\right) * log_2\left(\frac{6}{6}\right)\right) + \left(-\left(\frac{0}{6}\right) * log_2\left(\frac{0}{6}\right)\right) = \mathbf{0}$

(41-50 years old) $= \left(-\left(\frac{2}{7}\right) * log_2\left(\frac{2}{7}\right)\right) + \left(-\left(\frac{5}{7}\right) * log_2\left(\frac{5}{7}\right)\right) = \mathbf{0.863120569}$

Login Display, The login page functions to enter the next page according to the access rights that have been registered by the admin. On this page there is data that must be entered such as username, password in order to enter the next page. The login page display can be seen in the picture below:

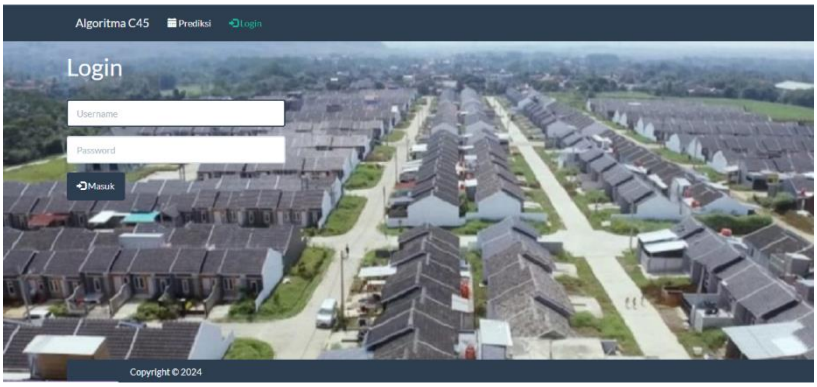
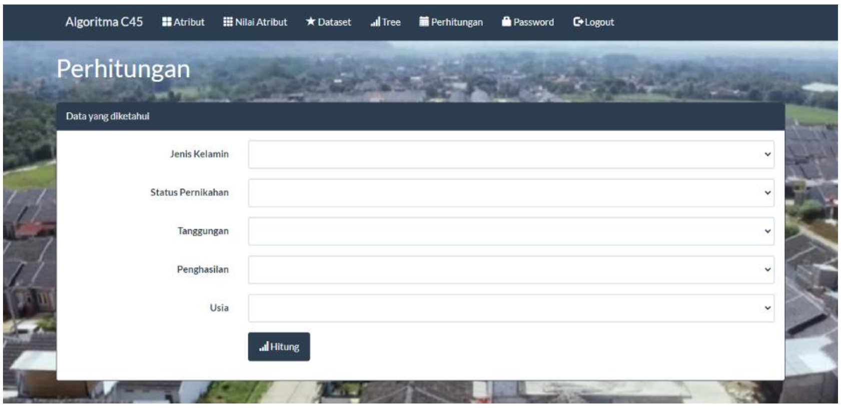


Figure II. Login View

Calculation Display, on this page serves to find out the results of calculations that will be used as predictions by the company. The display of the calculation page can be seen in the picture below:



Perhitungan

Data yang diketahui

Jenis Kelamin

Status Pernikahan

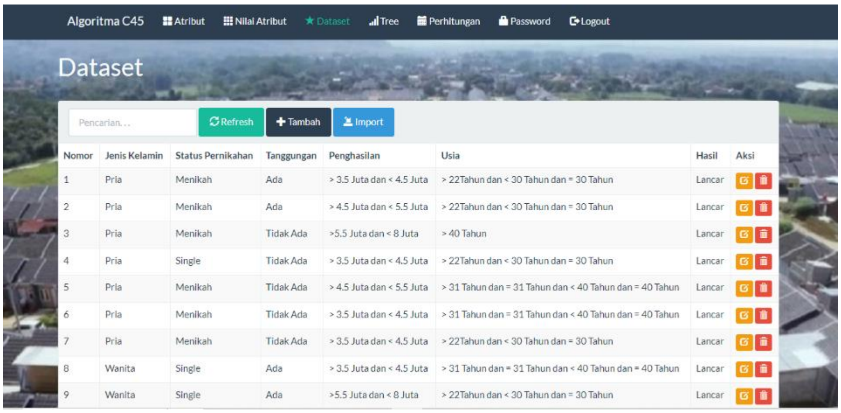
Tanggungan

Penghasilan

Usia

Figure III. Calculation View

Dataset Display, On this page there is information on consumer data that has been inputted to become a prediction reference. The dataset page display can be seen in the image below:



Dataset

Nomor	Jenis Kelamin	Status Pernikahan	Tanggungan	Penghasilan	Usia	Hasil	Aksi
1	Pria	Menikah	Ada	> 3.5 Juta dan < 4.5 Juta	> 22Tahun dan < 30 Tahun dan = 30 Tahun	Lancar	
2	Pria	Menikah	Ada	> 4.5 Juta dan < 5.5 Juta	> 22Tahun dan < 30 Tahun dan = 30 Tahun	Lancar	
3	Pria	Menikah	Tidak Ada	>5.5 Juta dan < 8 Juta	> 40 Tahun	Lancar	
4	Pria	Single	Tidak Ada	> 3.5 Juta dan < 4.5 Juta	> 22Tahun dan < 30 Tahun dan = 30 Tahun	Lancar	
5	Pria	Menikah	Tidak Ada	> 4.5 Juta dan < 5.5 Juta	> 31 Tahun dan = 31 Tahun dan < 40 Tahun dan = 40 Tahun	Lancar	
6	Pria	Menikah	Tidak Ada	> 3.5 Juta dan < 4.5 Juta	> 31 Tahun dan = 31 Tahun dan < 40 Tahun dan = 40 Tahun	Lancar	
7	Pria	Menikah	Tidak Ada	> 3.5 Juta dan < 4.5 Juta	> 22Tahun dan < 30 Tahun dan = 30 Tahun	Lancar	
8	Wanita	Single	Ada	> 3.5 Juta dan < 4.5 Juta	> 31 Tahun dan = 31 Tahun dan < 40 Tahun dan = 40 Tahun	Lancar	
9	Wanita	Single	Ada	>5.5 Juta dan < 8 Juta	> 22Tahun dan < 30 Tahun dan = 30 Tahun	Lancar	

Figure IV. Dataset View

4. CONCLUSION

The implementation of data mining using the C4.5 algorithm in predicting the eligibility level of prospective subsidized KPR consumers at Mitra Griya Indah Housing shows positive results. The C4.5 algorithm successfully processes prospective customer data that includes various attributes such as age, income, employment status, and credit history, to produce an accurate and reliable predictive model. With a high level of accuracy, this model can help speed up the selection process of potential customers, reduce the risk of bad debts, and ensure that subsidies are given to those who are truly deserving. This implementation not only improves the company's operational efficiency but also supports more informed decision-making.

REFERENCES

[1] G. K. Gupta, *Introduction to data mining with case studies*. PHI Learning Pvt. Ltd., 2014.

[2] I. S. Damanik, A. P. Windarto, A. Wanto, Poningsih, S. R. Andani, and W. Saputra, "Decision tree optimization in C4. 5 algorithm using genetic algorithm," in *Journal of Physics: Conference Series*, 2019, vol. 1255, no. 1, p. 12012.

[3] A. Cherfi, K. Nouira, and A. Ferchichi, "Very fast C4. 5 decision tree algorithm," *Appl. Artif. Intell.*, vol. 32, no. 2, pp. 119–137, 2018.

[4] J.-S. Lee, "AUC4. 5: AUC-based C4. 5 decision tree algorithm for imbalanced data classification," *IEEE Access*, vol. 7, pp. 106034–106042, 2019.

[5] P. Chen, "The application of an improved C4. 5 decision tree," in *2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC)*, 2021, pp. 392–396.

[6] A. Z. Abdullah, B. Winarno, and D. R. S. Saputro, "The decision tree classification with C4. 5 and C5. 0 algorithm based on R to detect case fatality rate of dengue hemorrhagic fever in Indonesia," in *Journal of Physics: Conference Series*, 2021, vol. 1776, no. 1, p. 12040.

[7] C. Deng and Z. Ma, “Research on C4. 5 Algorithm Optimization for User Churn,” in *2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE)*, 2021, pp. 75–79.

[8] M. R. Wayahdi, S. H. N. Ginting, and D. Syahputra, “Greedy, A-Star, and Dijkstra’s algorithms in finding shortest path,” *Int. J. Adv. Data Inf. Syst.*, vol. 2, no. 1, pp. 45–52, 2021.

[9] M. R. Wayahdi, D. Syahputra, and S. H. N. Ginting, “Evaluation of the K-Nearest Neighbor Model With K-Fold Cross Validation on Image Classification,” *INFOKUM*, vol. 9, no. 1, Desember, pp. 1–6, 2020.

[10] S. H. N. Ginting, “The Utilization Of The Simple Multi Attribute Rating Exploiting Ranks Can Enhance The Performance Of The Aco Algorithm,” *J. Minfo Polgan,* vol. 12, p. 1325, 2023, doi: doi.org/10.33395/jmp.v12i1.12743.

[11] M. R. Wayahdi and F. Ruziq, “KNN and XGBoost Algorithms for Lung Cancer Prediction,” *J. Sci. Technol.*, vol. 4, no. 1, pp. 179–186, 2022.

[12] F. Ruziq and M. R. Wayahdi, “Sistem Pendukung Keputusan Seleksi Karyawan Baru dengan Simple Additive Weighting pada PT. Technology Laboratories Indonesia,” *J. Minfo Polgan*, vol. 11, no. 2, pp. 153–159, 2022.

[13] B. Santoso, “Expert System Utilizing Bayesian Theorem Method for Hernia Disease,” *J. Technol. Comput.*, vol. 1, no. 1, pp. 18–22, 2024.

[14] A. Roy and others, “Applying the SMART methodology within decision support systems to evaluate the suitability of oil palm fruit for production,” *J. Technol. Comput.,* vol. 1, no. 1, pp. 1–5, 2024.