Comparative Analysis of Data Mining Classification Methods in Predicting Credit Payments

Nafla Putri Kinanti[1)], Nabila Khansa Raefa[2)], Diah Ayu Kamila[3)], Wahyunengsih[4)]

[1-4)] Departement of Matematika, UIN Syarif Hidayatullah Jakarta

Correspondence Author: ***wahyu.nengsih@uinjkt.ac.id***

Article Info :

Article History :

Received :
05 July 2024
Revised :
20 July 2024
Accepted :
18 August 2024
Available Online :
28 August 2024

Keyword :

Credit Payments, Naive Bayes, Research Method.

ABSTRACT

Credit activities in the form of saving money from members, providing loans to members and managing existing funds are still intuitive and can cause errors in the credit process. So that this crediting process does not occur and can run smoothly, a payment prediction using data mining is needed. Therefore we conducted this research. The research method we use is the knowledge discovery in database process model, where this process is divided into several stages, namely selection, pre-processing or cleaning, transformation, data mining and evaluation.

1. INTRODUCTION

This research explores the efficacy of these algorithms in predicting credit repayment behaviour. This research utilizes a qualitative descriptive methodology, focusing on literature review and algorithmic analysis. The literature review highlighted the importance of data mining in automating the analysis of extensive data sets to uncover hidden patterns, especially in the context of credit repayment. The research utilized trusted sources, such as academic journals and institutional repositories, for data collection to ensure reliability and relevance.

Two main algorithms, Decision Tree C4.5 is a calculation utilizing the concepts of getting data picked up and decreasing entropy values to choose the optimal division. Here, it can discover the preferences of C4.5, and Naïve Bayes is a simple probabilistic classification algorithm that performs calculations by summing frequencies and combinations of values from a given data set.

Performs calculations by summing frequencies and combinations of values from a given data set. The C4.5 Decision Tree algorithm was evaluated for its effectiveness in classifying credit applicants based on attributes such as income level, residential status, spouse's occupation, and existing credit. The C4.5 Decision Tree algorithm was highlighted for its ability to achieve 100% accuracy in classification, which is due to its structured decision-making process that utilizes entropy and information gain calculations.

In contrast, the Naïve Bayes algorithm, although more straightforward and faster, achieved an accuracy rate of 86.67%. It relies on probabilistic calculations assuming independence among attributes, which may affect its prediction accuracy in specific scenarios. This research concludes by advocating the Decision Tree C4.5 algorithm as the preferred choice for credit prediction due to its superior accuracy. The study also recommends future research directions to refine the

ISSN Print : 1979-7141
ISSN Online : 2541-1942

algorithmic methods and explore hybrid approaches to improve classification results in credit-scoring scenarios. Based on the statement above, the problem formulation is:

1. What is the comparative effectiveness of the Naive Bayes and C4.5 classification methods in predicting credit payments?
2. How do we analyze each classification method's relative advantages and disadvantages in the context of credit payment prediction?

2. METHOD

2.1 Research Methodology

This research uses a literature search method. The author used a qualitative descriptive method because the research was carried out by collecting data from several journals and other library sources. Qualitative Descriptive Research is research carried out by describing and interpreting interrelated things. This research will collect data from customers who use credit payments. Next, the data will be analyzed using several algorithm models to determine which algorithm model has better effectiveness and accuracy. Based on previous research that has been examined as evidence that a prior research process has been carried out, this earlier research can be used as an illustration in conducting research. The test results show that the algorithm with the highest score is the decisionc4.5 algorithm, but several researchers show that Naïve Bayes is superior (Sri Widaningsih 2019).

2.2 DataCollection

Data was collected through a comprehensive literature study, including examination and analysis of articles sourced from trusted sources such as Garuda. Kemendikbud, media.elite, journal sites, and other trusted publications. This rigorous process ensures that the data collected is reliable, current, and reflects the latest trends and developments in the field. By utilizing various authoritative sources, researchers can access a wealth of information and perspectives, thereby increasing the depth and breadth of their research. Data collection techniques are used to compare literature studies to produce information based on the topics discussed.

2.3 Data Analysis

Data analysis in this research uses the algorithm models. consisting of Decision Tree C4.5,and Naïve Bayes. The stage of applying the classification model aims to predict class classification from the data in the testing data based on the classification model that has been made.

a. Decision Tree C4.5

C4.5 Choice tree calculation Utilizing the concepts of getting data pick up and decrease entropy values to choose the optimal division. Here, able to discover out the preferences of C4.5, as expressed by "Sumit Saha:" This calculation intrinsically employments a Single Pass Pruning Handle to Diminish overfitting. It can work with Discrete, and Ceaseless Information C4.5 can handle the issue of inadequate information exceptionally well. Its shortcomings, agreeing to the Serang Raya College "Data Frameworks" diary, C4.5 can overcome the powerlessness of non-informative qualities and exactness estimations that depend on just one likelihood by making a choice tree demonstrate. Be that as it may, choice trees got to progress when confronted with traits and huge sums of data.Works with both discrete and nonstop data, and C4.5 can handle the issue of insufficient data uncommonly well. The calculation of the entropy regard can be seen in Condition 1 underneath :

ISSN Print : 1979-7141
ISSN Online : 2541-1942

$$\textbf{\textit{Entropy}}\ \textbf{(S)} = \sum -\textbf{pi}\ .\ \textbf{\textit{log2pi}}$$

Information:

S: Set of cases

N: Number of partitions

S Pi: Proportion of Si to S

Meanwhile, the value of information gain can be calculated using the equation.

$$Gian(S, A) = Entropy(s) - \sum_{i=1}^{n} \frac{|Si|}{|S|} + Entropy(Si)$$

Information:

S: Set of cases

A: Attributes

n: Number of attribute A partitions

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

|S|: Number of cases in S

b. Naive Bayes

The naïve Bayes algorithm is a simple probabilistic classification algorithm that performs calculations by summing the frequency and combination of values from a dataset that is Given. Naive Bayes has a weakness: if the conditional probability is zero, the prediction probability will also be zero. The assumption that each variable is independent results in reduced accuracy because usually there is a correlation between one variable and another variable. Accuracy cannot be measured using one probability. Just. Need other evidence to prove it. As stated by "mochammad haldi widianto" the advantages are Can be used for quantitative and qualitative data, Does not require a large amount of data, No need to do a lot of training data. If there are missing values, they can be ignored in calculations, Calculations are fast and efficient, Easy to understand, Easy to create as stated by Mochammad Haldi Widianto."In general, Bayes' theory is written into an equation.

$$P(X|H) = \frac{P(X|H)P(H)}{P(X)}$$

Information:

X: Data with unknown class

H: Data hypothesis, which is a specific class

P(H|X): Probability of hypothesis H based on condition X (posterior probability)

P(H): Probability of hypothesis H (Prior probability)

ISSN Print : 1979-7141
ISSN Online : 2541-1942

P(X|H): Probability of X based on the conditions in the hypothesis

HP(X): Probability of X

2.4 The limitation of the study

The study's limitations are that journals regarding research on certain algorithm models are often limited and only accessible through institutional subscription access. Second, there was limited time and a restricted data collection method, where we could only gather information from a few journals and were unable to access the objects directly. Third, using C4.5 to measure accuracy relies on a decision tree design. Each variable in Naive Bayes is independent so that accuracy can be reduced

3. RESULTS AND ANALYSIS

3.1 Categories

In this research, the theme taken is data mining, especially in applying algorithm models in predictions that are non-fluent in making payments. Data mining is the automated analysis of large amounts of data or complex data to find important patterns or tendencies that are usually unaware of. (Pramudiono, 2006). Of course, data mining also experiences this phase, but what differentiates it is that in Data Mining, the input is the dataset, the process is the algorithm or method in data mining itself, and the output is Decision Tree C4.5 and Naïve Bayes.

3.2 Interpretation of the Finding

This study explains that the comparison aims to get more definite results on lending to customers. This can be seen from the accuracy obtained from the algorithm's application results. In addition to problems in research, it is also essential to know the data used.

Table 1. Customer Data

alternatif	penghasilan	Status rumah	Pasangan kerja	Kredit lainnya	Hasil keputusan
A1	Tinggi	sewa	Ya	Tidak	Diterima
A2	Sedang	sewa	Tidak	Ya	Ditolak
A3	Sedang	sewa	Tidak	Tidak	Diterima
A4	Tinggi	sewa	Ya	Ya	Diterima
A5	Tinggi	sewa	Ya	Tidak	Diterima
A6	Rendah	sewa	Ya	Tidak	Ditolak
A7	Sedang	Milik pribadi	Tidak	Ya	Ditolak
A8	Tinggi	Milik pribadi	Tidak	Ya	Diterima
A9	Rendah	Milik pribadi	Ya	Tidak	Ditolak
A10	Sedang	sewa	Ya	Ya	Diterima
A11	Tinggi	sewa	Ya	Tidak	Diterima
A12	Rendah	Milik pribadi	Tidak	Ya	Ditolak
A13	Sedang	Milik pribadi	Ya	Tidak	Diterima
A14	Tinggi	Milik pribadi	Tidak	Tidak	Ditolak
A15	Sedang	sewa	Ya	Tidak	Diterima

(Dison Librado , Asyahri Hadi Nasyuha, 2023

1.Naive bayes

Table 2. Submission data

No	Alternatif	income	home	work partner	other credits	Results
1	A1	medium	one'own	yes	no	???

The previous table shows data on new customers who will apply for credit. From this data, a decision-making process will be carried out. However, before the decision-making process takes place, the probability value must first be calculated as a basis for decision-making; as in previous researchers, the "accepted" probability value is the product of 0.44, 0.22, 0.78, and 0.67. and that "rejected" is 0.33, 0.67, 0.33, 0.5 from each criterion. (Disson Librado, Asyahri Hadi Nasyuha, 2023)

ISSN Print : 1979-7141
ISSN Online : 2541-1942

a. Define Classes

The last stage carried out in Naïve Bayes Algoirtma is the process of determining the final class. From the process done in the previous research, it is known that for the class value "Accepted," the probability value is 0.030352608. In contrast, for the class "Rejected," it has a probability value of 0.00145926. From the results of the process, a decision-making process can be made that the result is "Accepted ."This is because the value obtained in the "Accepted" class is greater than that obtained in the "Rejected" class. the value received in the "Rejected" class. The decision-making picture can be seen in the following table. (Dison Librado, Asyahri Hadi Nasyuha, 2023)

Tble 3. Results Table

No	Alternatif	income	House Status	work partner	other credits	result
1	**A1**	**medium**	**one'own**	**yes**	**no**	Accepted

It can be seen as the final result when making a decision. From the attribute values contained in tables such as Income is Medium, Home Status is Owned, Spouse is Yes, and there are Other Credits is No, the result of the decision-making carried out is "Accepted" for the credit application process.

3.3 Reliability and validation

The final stage carried out at Algoirtma Naïve Bayes is determining the final class. The process that has been done shows that for the class value "Accepted," the probability value owned is 0.030352608, while for the "Rejected" class, there is a probability value of 0.00145926. The results of the process can then be used to decide that the result is "Accepted."

In the table, the final results of the decision can be seen. From the attribute values contained in the table, such as Income is Medium, Home Status is Owned, Spouse is Yes, and there are Other Credits is No. Then, the result of the decision-making is "Accepted" for the credit application process. (Dison Librado , Asyahri Hadi Nasyuha, 2023)

2. Decision tree C4.5

After the process with the Naive Bayes algorithm and the decision-making process are complete, the next process continues with the C4.5 algorithm. The process carried out by the C4.5 algorithm begins with calculations: the entropy value of each attribute, the gain value of the attribute, and the total attribute value.

The results in previous studies show that the Income attribute has the highest Gain value of 0.34362, so the Income attribute was chosen as the root of the decision tree. There are still two attribute values from the completed process with more than 1 (one) decision result, so the process of calculating nodes to form a decision tree is still ongoing. (Dison Librado, Asyahri Hadi Nasyuha, 2023)

Table 4. Income table

Alternative	Income	House Status	Work Partner	Other Credits	Decision
A1	High	Rent	Yes	No	Accepted
A4	High	Rent	Yes	Yes	Accepted
A5	High	Rent	Yes	No	Accepted
A8	High	One's Own	No	Yes	Accepted
A11	High	Rent	Yes	No	Accepted
A14	High	One's Own	No	No	Rejected

ISSN Print : 1979-7141
ISSN Online : 2541-1942

Table 5. Approval table

No	Alternatif	income	House Status	Work Partner	Other Credits	Disicion
1	**A1**	**medium**	**one'own**	**yes**	**No**	**Accepted**

This table contains the C4.5 algorithm search results. The decision-making process can be seen based on the rules that emerge. In this case, it arises from the attribute income = average. Because income = average cannot yet be determined, it arises from the next attribute, namely working partner. In the colleague attribute = Yes, the decision taken is Accepted. In this case, the result of the decision is "Approved."

After the calculation and decision-making process using the Naive Bayes and C4.5 algorithms, the next step is the comparison process. The comparison process is carried out based on the accuracy value obtained by each algorithm.

	true accepted	True Rejected
In front of the Dieters	9	0
In front of the Dieters	0	5

$$\text{Accuracy} = \frac{9+5}{9+5+0+0} \times 100\% = 100\%$$

$$\text{Precision} = \frac{9}{9+0} \times 100\% = 100\%$$

$$\text{Recall} = \frac{9}{9+0} \times 100\% = 100\%$$

Within the picture, it can be seen that the precision rate of the C4.5 calculation is 100%. After knowing the level of exactness of the C4.5 calculation, at that point the level of exactness against the Naïve Bayes calculation. Where the level of accuracy can be seen within the taking after figure:

	true accepted	True Rejected
In front of the Dieters	8	1
Pred rejected	1	5

$$\text{Accuracy} = \frac{8+5}{8+5+1+1} \times 100\% = 86{,}6\ \%$$

$$\text{Precision} = \frac{8}{8+1} \times 100\% = 88{,}8\%$$

$$\text{Recall} = \frac{8}{8+1} \times 100\% = 88{,}8\%$$

ISSN Print : 1979-7141
ISSN Online : 2541-1942

3.4 Conclusion of The Finding

The ultimate result of the consider may be a prepare of concluding, where the steps taken within the problem-solving handle lead the research to conclude that information mining can be utilized to unravel the inquire about issue, particularly within the prepare. Exit for information handling. Information mining with Gullible Bayes and C4.5 calculations have been completed and can be classified for decision-making; both calculations have the same choice result, to be specific "Elude." but, the level of exactness accomplished is the same. The Gullible Bayes calculation has an precision rate of 86.67%, whereas the C4.5 calculation has an precision rate of 100%. (Dison Librado , Asyahri Hadi Nasyuha, 2023)

4. DISCUSSION

It can be concluded that the C4.5 / Choice tree calculation has the most excellent classification precision compared to the Gullible Bayes calculation for classifying credit beneficiaries. The inquire about expressed that the Gullible Bayes calculation had an precision rate of 86.67%, whereas the C4.5 algorithm had an precision rate of 100%. This can be gotten by comparing the rate of precision comes about from the two calculations. The solution uses the Naïve Bayes Algorithm by calculating all the attribute values in the data. Based on the new data, a calculation process will be carried out (A. Sentimen, 2022) [11]. Each attribute value in the latest data will be calculated based on Bayes' Theorem. Decision-making uses the Naïve Bayes algorithm to compare probability values (I. Verawati and B. S. Audit, 2022) [13].

Meanwhile, the C4.5 algorithm is based on calculating Entropy and Gain values (Ubaedi and Y. M. Djaksana, 2022) [1]. All attributes will be calculated for each Entropy value. After calculating the entropy value, the grain value for each attribute is calculated. The attribute with the highest Gain value will become the main root of the decision tree (R. Girsang, E. F. Ginting, and M. Hutasuhu, 2022) [16].

Theoretical Implications here aim to convince examiners regarding the contribution to science in the theories used to solve research problems. They are based on the application of theory in previous research. The research has the same objective as the research conducted by Dedi Darwis et al in 2021 with the research title "Application of the Naive Bayes Algorithm for Dedi's National BMKG Twitter Data Review Sentiment Analysis," where the research results showed that the classification process carried out made it easy for users to can see every positive, neutral or negative opinion and the accuracy rate obtained is 69.97% (D. Darwis,dkk, 2021) [18]. Apart from the Naive Bayes algorithm, there are also several references to the C4.5 algorithm, such as research conducted by Desi Marlina and Muhammad Bakri in 2021 with the research title "Application of Data Mining to Predict Customer Transactions Using the C4.5 Algorithm" that the results of the research obtained are Algorithms can be used to predict the size of transactions carried out by customers which are used to help determine the provision of loans (D. Marlina and M. Bakri, 2021)[21].

To guarantee simplification and transferability, it can be surveyed based on the handling of past credit client information to get a design of which clients are qualified to induce the credit or which clients may have trouble paying credit, which causes losses for the company. The proper way to unravel the issue in information mining is through the classification handle. The strategy comparison handle within the inquire about was carried out to get more clear comes about. This is

ISSN Print : 1979-7141
ISSN Online : 2541-1942

often based on giving credit to the correct individual so that everything runs easily when completing credit charge installments.

The down to earth suggestion is that it is carried out utilizing the Naïve Bayes and C4.5 calculations. These two calculations are presently exceptionally commonplace for tackling classification issues in mining information. Separated from the basic handle, the level of exactness gotten from these two calculations is additionally the premise for the truth that these algorithms are broadly utilized to assist fathom issues. Within the usage handle, giving credit to clients must be suitable. Giving suitable credit to clients will make it less demanding for the company to carry out collections until the credit charge is paid in full. In the event that charging is postponed or hampered, there will too be issues with the trade forms and charging carried out (I. Nurjanah,dkk, 2023) [5], [6].

The limitation of the alternative explanation here is that it often causes problems such as difficulties with billing, payments that are usually late or not on time, and even worse, customers cannot pay the bill each month. For customers unable to pay their monthly bills, this will undoubtedly impact the company. The problems must be resolved immediately; continuing and not providing credit to the right customers will be very detrimental to the company if the issues continue. Apart from bad credit, inability to pay and not paying off can also have a bad impact, causing the company to go bankrupt because the company no longer has operational costs to run the company.

5. CONCLUSION

The problem/objective of this study is to compare the performance of four classification algorithms: Decision Tree C4.5 and Naive Bayes, based on evaluation metrics such as accuracy, precision, recall, and execution time. This research aims to find out which algorithm is most effective in classifying a dataset and provide a better understanding of each algorithm's advantages and disadvantages.

Thus, this research is expected to provide practical guidance in choosing a classification algorithm for a particular problem based on the characteristics and needs of the dataset and the availability of computing resources. The algorithm with the best classification is the Decision tree/C4.5 algorithm.

This is shown by the Decision tree/C4.5 algorithm being the most effective in classifying a particular dataset with 100% accuracy. However, even with 100% accuracy, we still need to consider the advantages and disadvantages of both algorithms.

6. DECLARATION OF COMPETING INTEREST

We declare that we have no conflict of interest.

7. REFERENCES

Gupitha, R 2018, '*Penerapan Klasifikasi Status Pegawai Menggunakan Metode Naive Bayes di RSU H. Syaiful Anwar'*, Vol. 5 No. 1, hh. 28-37.

ISSN Print : 1979-7141
ISSN Online : 2541-1942

Anggraini, L, Yamasari, Y 2023, *'Klasifikasi Citra Wajah Untuk Rentang Usia Menggunakan Metode Artificial Neural Neutwork',* Vol. 5 No. 2, hh. 185-191

Widaningsih, Wida, S 2021, *'Analisis Perbandingan Metode Klasifikasi Data Mining Dalam Memprediksi Pembayaran Kredit',* Program Sistem informasi, Universitas Widyatama

Librado D. & Hadi A 2023, 'Penerapan Data Mining untuk Klasifikasi Penerima Kredit dengan Perbandingan Algoritma Naive Bayes dan Algoritma', *Jurnal Media Informatika Budidarma.,* Vol 7 No. 4, hh. 1952-1961

Sari, D 2016, 'Komparasi 5 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan', *Techno Nusa Mandiri,* Vol. XIII, No. 1, hh. 60-66

Pratiwi, F, Hartama, D, Damanik, I, Irawan, E, Saragih, I 2020, 'Implementasi Algoritma Naive Bayes Dalam Memprediksi Kenaikan Golongan Karyawan', *Prosiding Seminar Nasional Riset Information Science,* Vol. 2(2020), hh 206-215

Pangestu, B 2023, 'Data Mining Menggunakan Algortima Naive Bayes Classifier Untuk Evaluasi Kinerja Karyawan', *Jurnal Riset Matematika,* Vol. 3 No. 2, hh 177-184

Kusrini, Taufik, E 2009, '*Algoritma Data Mining',* STMIK Amikom Yogyakarta, Penerbit Andi

ISSN Print : 1979-7141
ISSN Online : 2541-1942