Predicting Loan Repayment Reliability in Cooperative Societies using Naive Bayes Classifier: A Data Mining Approach for Risk Mitigation and Decision Support

Saiful Kabir
Department of Computer Science
and Engineering,
BGC Trust University Bangladesh,
Chattogram, Bangladesh

Sihabul Islam Safin
Department of Computer Science
and Engineering,
BGC Trust University Bangladesh,
Chattogram, Bangladesh

Marjahan Tanjin
Department of Computer Science
and Engineering,
BGC Trust University Bangladesh,
Chattogram, Bangladesh

Himu Akter
Department of Computer Science
and Engineering,
BGC Trust University Bangladesh,
Chattogram, Bangladesh

Rajib Ghose
Department of Computer Science
and Engineering,
Military Institute of Science and
Technology (MIST)

Abhijit Pathak
Department of Computer Science
and Engineering,
BGC Trust University Bangladesh,
Chattogram, Bangladesh

ABSTRACT

In this research, the primary analysis method is the Naive Bayes Classifier for predicting the reliability of loan repayment in cooperative environments to facilitate credit analysis for a cooperative's staff. Floor management which means how employees conduct assessments of loans can lead to the formation of non-performing loans as borrowers fail to pay as agreed. To avoid these risks the following evaluation procedures in disbursement of loans are highly relevant and necessary. Thus, drawing on historical member data, this research uses data mining, namely the Naive Bayes Classifier, to predict the chances of smooth loan repayment. The Naive Bayes method is based on the records of various attributes of the members, which include occupation, income, home ownership, amount of loan, and type of loan. These attributes are useful in the prediction of some of the qualities that are useful in decision-making in as much as the loan is concerned. Based on the assessment of the Naive Bayes classifier on the current model, an accuracy rate of 90% was obtained. 00%, an accuracy of 0. 880%, and a recall (Sensitivity) of 83%. , a recall of 33%, and the precision of 100%. These measures indicate the high performance of the proposed model in correctly classifying positive as well as negative loan repayment status. In the subsequent studies, it will be interesting to work on expanding a more appropriate dataset for improving the model's predictive capability by increasing the variation of individuals' examples. Hence, the expansion of more intricate computing approaches that consider attributes' interdependencies may shed light on enhanced methods of risk identification and loan approvals. Through progressive enhancement and creativity in the lending area of operations, the risks involved in lending as well as the overall loaning practices can be controlled and enhanced respectively.

General Terms

Loan Repayment Reliability, Cooperative Societies, Naive Bayes Classifier, Conditional Independence, Decision Support.

Keywords

Prediction, Smoothness of Installment Payments, Data Mining, Naive Bayes Classifier, Machine Learning.

1. INTRODUCTION

A cooperative is a company or other legal body whose members work together to raise their standard of living and well-being as a group. One of its main responsibilities is to provide loans with credit-based repayment plans. The distribution of money or comparable claims based on an agreement or loan arrangement that compels the borrower to return their debt within a specific term is defined as credit under Law No. 10 of 1998 (Banking Law). Although lending can be advantageous for cooperatives, there are a lot of dangers involved that could result in losses. These risks occur when borrowers postpone or neglect to make loan payments on schedule, which leads to non-performing or poor loans. Bad loans are caused by several things, such as members' bad repayment habits and officials' carelessness or mistakes when determining creditworthiness. To reduce potential losses at the Grameen Multipurpose Co-Operative Society Ltd., there is a need for a system that can forecast the seamless repayment of loans by prospective members. This can be accomplished by utilizing members' personal data and data mining techniques with the Naive Bayes classifier algorithm [21]. To find patterns, trends, and insights from huge datasets, data mining entails gathering, utilizing, and analyzing historical data. Subsequently, this data can be employed to facilitate decision-making and enhance subsequent choices. A simple probabilistic classification technique that determines the likelihood of different outcomes based on the provided dataset is the Naive Bayes classifier. This algorithm is also referred to as Bayes' Theorem since it can forecast future probabilities based on historical data. The Naive Bayes classifier was used in earlier research by Rizal and Raisa, et al. to forecast the smooth payment of MSME terrace rentals. With an accuracy of 81.81%, their study offered details on several outstanding payment statuses depending on particular parameters and regions. The findings of this study provided support for choices made to lower the incidence of delinquent payments. Similar predictions have been made using the Naive Bayes approach in other investigations [11].

In a different study, Sri and Rolly, et al. used six criteria—loan number, gender, marital status, kind of business, loan amount, and loan type—to predict the smooth repayment of loans by potential borrowers in a savings and loan cooperative. The accuracy of this study, which employed the K-NN approach, was 73.37%. Using the Naive Bayes approach, the current study seeks to outperform earlier research in terms of accuracy [12].

Because members typically postpone or forget to make loan repayments on time, cooperatives frequently face major risks while engaging in lending activities. Bad or non-performing loans cause financial losses for the cooperative. The present methods for assessing a potential borrower's creditworthiness are unreliable, which makes for poor decision-making and an increased risk of bad loans. To reduce these risks, an efficient system that can precisely forecast the repayment patterns of potential borrowers is therefore important. The goal of this study is to determine how well the Naive Bayes classifier algorithm predicts future members of the Grameen Multipurpose Co-Operative Society Ltd. will repay their loans, and how accurate it is in comparison to other prediction techniques like K-NN [13].

This research has several different goals as shown in figure 1. First, it seeks to create a prediction system based on personal data that forecasts potential borrowers' repayment behavior using the Naive Bayes classifier method. Second, by using this predictive approach, the research hopes to lessen the number of non-performing or bad loans inside the Grameen Multipurpose Co-Operative Society Ltd. Third, to assess the Naive Bayes classifier method's efficacy, its accuracy will be compared to that of other predicting techniques, particularly K-NN. Fourth, to help cooperative personnel make well-informed decisions about the creditworthiness of potential borrowers, the research aims to deliver dependable prediction results. Finally, by reducing loan defaults via improved forecasting and decision-making procedures, the cooperative's overall financial stability and profitability will be improved [14].

2. BACKGROUND STUDY

Cooperative Financial is a community-based entity that offers member-only financial solutions while aiming not at generating revenues but at maximizing members' earnings. Through the ownership of these institutions, such cooperatives empower individuals empowered within the economy. Members gain from training, voluntary support, and membership, a structure of membership that is in a manner that is inclusive and autonomous together with democratic control. The identification of the cooperative model lies in the cooperation, to increase efficiency with the objectives of improving the life of members and participating in the growth of the cooperative movement and maintenance of sustainable economic development of the community [19]. The profits realized are distributed back to the member-owners which creates a virtuous cycle where customers, members, and the cooperative all benefit when there is increased business.

As for the borrowers, they should pay the borrowed amount in due course and the cooperative also works towards the same. To the borrowers, its benefits include no penalties in case they delay making their payments, good ratings for repayment, reduced penalties and interest, and also a good working relationship with the cooperative [20]. Thus, for the cooperative, the loan repayments constitute one of the necessary forms of financing that guarantees the safe functioning and development of capital. On-time repayments are essential in maintaining the cooperative's financial health, are key to developing the credit necessary to encourage fund availability, help avoid a high risk of default, are indispensable in the cooperation's growth and development, keep the costs of interest low, helps coherence with the cooperative principle about member's creditworthiness, is important in developing member confidence in the cooperative, is also socially responsible as legal and ethical requirements must be met, improves cooperative operational effectiveness [15].

In this case, the authors experiment to compare the Naive Bayesian Probabilistic Method to the Support Vector technique when identifying the most lucrative lending plan. Applying the Naive Bayes algorithm was proved to be accurate and efficient in the classification and prediction investigation with an average precision of 0.975.00%, recall of 100.00%, Accuxtacy of 100%, and F1 score of 99.00 %, PICS threshold set to 95.00%. Overall, the investigations presented here suggest that the Naive Bayes algorithm remains highly reliable in determining the ability of borrowers to pay back the loans while also calculating the profitability of extending credit [1].

In this article, the authors discuss the use of structured decision trees Naïve Bayes algorithm to predict the reliability of loan repayment in cooperatives. They compared the performance of different Kernel selection methods for SVM including Linear Kernel, Polynomial Kernel, newly developed and more accurate SVM-P Kernel, Radial Base Function Kernel, and Sigmoid Kernel with Naïve Bayes. In understanding the results, the Naive Bayes algorithm was rated the highest in the four streams of scores, thus making it gain the best rank in this usage. In particular, the indicated indexes revealed that the performance of SVM-P ranked the lowest in terms of accuracy, F1, precision, and recall, as well as the ratio of predicted values to actual ones. While using the F1 score, precision as well as recall; SVM-R gave the best results, accurate Naive Bayes model yields the overall better results and is thus exceptional in the task of predicting loan repayments of cooperatives' reliability [2].

For this paper, six supervised machine learning algorithms were used with the primary goal of developing default prediction models for loan lenders; the attribute of early loan repayment was applied. In this study, it emerged that most models that have disregarded early loan repayment delivery have inferior performance indicators as compared to the ones that account for this delivery. The models considering early loan repayment depicted more accuracy, T recall, precision, RMSE, and ROC features. Of these all, the Random forest has given satisfactory results in predictive analysis of the data set by having an accuracy of 93 percent, RMSE of 11 percent, the precision of 90 percent, recall of 89 percent as well and a ROC value of 81 percent. Whereas Naïve Bayes was applied to set debt expectations, Random Forest stood as the most accurate model for determining loan repaying credibility [3].

In this article, the authors discuss a new way of analyzing the ability of borrowers to pay back their debts by working with artificial intelligence. As a result, it aims to improve the viability of loan default predictions by integrating further credit information and other personal details. Current statistical techniques are lacking when following the loan default risk, for example, there is a FICO score. As the study shows it is possible to achieve considerable gains in prediction accuracy through the use of an artificial intelligence approach. Of the many different algorithms explored, the Gradient Boosting model was found to be superior to the other models in achieving adequate predictive performance that yielded the best PR-AUC score equal to 0.957. In essence, this suggests that it is far superior at predicting loan defaults than the traditional methods used [4].

In this paper, the machine learning classification algorithms used in the prediction of loan default are as follows: Logistic Regression, Decision Tree, and Artificial Neuron Network. The research problem of the study was centered on the inability of the current models to identify a borrower's ability to repay a loan, coupled with the harshness or lenient credit-granting criteria to minimize the rate of defaults and yet to achieve maximum profitability. In the analysis done, the model with the best accuracy was the Logistic Regression model with Decision Tree as the second model with only displayed accuracy of 84.68%. This goes a long way in establishing that, by applying ensemble learning models in the process, the overall accuracy when predicting loan default is improved [5].

Loan default prediction is the area of interest of this research in which the Naive Bayes algorithm is used to determine if borrowers are reliable or not in repaying their loans from cooperatives. In this paper, the authors explain one of the machine learning methods called predictive modeling and use it for determining how appropriate a particular customer might be for availing a loan. They used the following eight models in an attempt to identify whether the borrower would ever be capable of repaying the loan. Out of these models, the Adaboost realizing model performed best with a high recall rate of 0.384 and, an accuracy of 59, classifier. 2% while a true-negative rate stood at 76.74%. Moreover, in the same study, the authors pinpointed that credit history is the most influential predictor in the identification of loan defaults [6]

3 METHODOLOGY

3.1 Research Stages

The research stages represent the workflow in addressing the discussed problem to obtain results and conclusions. The stages of the research can be seen in the figure 1 below:

Several important steps are included in Figure 1. First, some information about an applicant for a loan and his ability to pay back is compiled from such sources as the records of the cooperative society and the financial papers. Subsequently, cleaning steps are performed, feature selection, and transformation before proceeding with Exploratory Data Analysis to determine patterns as well as relations between them [7].

The phase of model building involves the usage of the Naive Bayes classifier which is trained on some portion of data under the hypothesis of mutual stochasticity of the features given the explained variable. Specifically, to apply the model, one has to run it on RapidMiner- a software for data science, where one sets up data processing in the form of workflows, manages parameters, and tests the accuracy of the model [16].

After getting the final model, the model's performance is tested using a different dataset, as well as, using accuracy, precision, and recall, the model's efficiency is validated using k-fold cross-validation. In results interpretation, emphasis is given to findings that concern the nature of the factors that determine the reliability of loan repayment and recommendations that may be used by the cooperative societies in avoiding risks by lending money.

Finally, the research discusses the results and model's evaluation and identifies directions for future research and possible model enhancements when it comes to loan repayment predictions in cooperative societies – data mining, risk management, and decision support are highlighted.

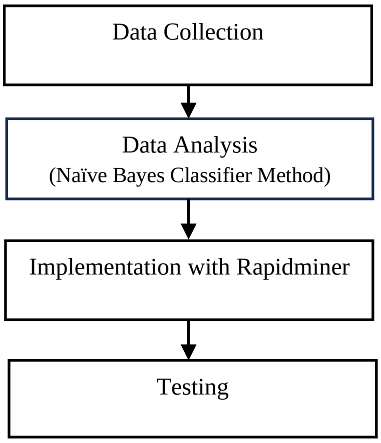
Data Collection
Data Analysis
(Naïve Bayes Classifier Method)
Implementation with Rapidminer
Testing

Fig.1: Research Stages

3.2 Data Attributes

Grameen Multipurpose Co-Operative Society Ltd. is therefore a multipurpose co-operative society specializing in the advancement of financial solutions to the economically disadvantaged and their societies through such products as; savings, various forms of loans, and financial services whereby people are educated on how to manage their finances. Huge care is considered with their mission to promote financial services access for all and enhance the socio-economic statuses. The above societies are nonprofitable member based where the profit arising from the activities are reinvested or shared amongst the members. In turn, Grameen collects and analyzes membership data as well as information relating to the sort of loans issued and financial transactions in an international bid to improve its services towards the development of the targeted communities and the global economy. Studies were made on the spot at Grameen Multipurpose Co-Operative Society Ltd., Bangladesh together with an administering structured interview for its staff, collectors, and administrators of Grameen Multipurpose Co-Operative Society Ltd [8]. Data that will be inputted for training will be the December 2021 members' data of the cooperative. Then the testing data will be the January 2022 members' data. This consists of 5 predictor attributes and 1 considered to be an attribute that will serve as the label. The training and testing data were grouped by attribute value, as was relevant to obtain the expected results based on the wants of the researcher [17].

Table 1. Data Attributes

Attribute	Type	Description	Variable
Occupation	Binomial	Permanent	X1
Income	Polynomial	High, Medium, Low	X2
Home Status	Polynomial	Own, Parental, Rent	X3
Loan Amount	Numeric	200,000 to 2,000,000	X4
Loan Type	Binomial	Daily, Weekly	X5
Classification	Label	On-time, Default	Y

3.3 Data Analysis

Data analysis is the act of gathering, selecting, processing, and rather changing data into useful information that can be used in guiding decision-making. In this stage, data mining techniques are applied in processing data into information using the Naive Bayes classifier algorithm [18].

3.4 Naive Bayes Classifier

The Naive Bayes algorithm is a non-rule-based technique that applies a part of mathematics known as probability theory. It does not depend on rules, but the most likely classification can be determined by looking at the frequency or the number of occurrences of each classification in the training data [10]. Investigations proved that Naive Bayes contributes to high accuracy and speed when applied to large databases. It's a probabilistic method where each feature in the data is independent of others. Although Naive Bayes is as powerful as decision trees and neural networks, the model relies on realistic assumptions. We express Bayes theorem with the following formula:

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

Explanation:

- **X:** Unknown data.
- **H:** Data hypothesis X is a specific class.
- **P(H|X):** Probability of hypothesis H based on condition X (Posteriori Probability).
- **P(H):** Probability of hypothesis H (Prior Probability).
- **P(X|H):** Probability of X based on hypothesis H.
- **P(X):**Probability

3.5 Testing

Testing is the last stage in data mining and is also referred to as model evaluation. Testing will be done to establish Naive Bayes' performance through measures such as accuracy, precision, and recall. Accuracy is a percentage value derived from the number of correctly identified training data. For the Naive Bayes approach, a performance confusion matrix will be used in testing [9]. The confusion matrix simply is a table with only predictions and actual values, turning over in positive and negative cases. An example of the resulting confusion matrix model is shown in the table below:

Tabel 2. Confusion Matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Here are the formulas to calculate accuracy, precision, and recall based on the table above.

Accuracy = $\frac{(TP+TN)}{(TP+FP+FN+TN)} \times 100\%$

Precision = $\frac{TP}{(TP+FP)} \times 100\%$

Recall = $\frac{TP}{(TP+FN)} \times 100\%$

4 Results and Discussion

Stages of Calculating the Naive Bayes Classifier Method:

4.1 Reading the Training Data

Training data, also known as training data, is the data collected to train the Naive Bayes classifier algorithm. The training data used consists of 85 data points, which include 54 data points in the smooth class and 31 data points in the congested class.

Table 3. Sample Training Data

Occupation	Income	House Status	Loan Amount	Loan Type	Classification
Permanent	High	Own House	500,000	Daily	Smooth
Freelance	Low	Rented	300,000	Weekly	Congested
Freelance	High	Own House	500,000	Daily	Smooth
Freelance	Medium	Own House	300,000	Daily	Smooth
Freelance	Medium	Own House	300,000	Daily	Smooth
Freelance	Medium	Parent's House	500,000	Weekly	Congested
Freelance	Low	Own House	300,000	Daily	Congested
Freelance	Low	Rented	500,000	Daily	Congested
Permanent	High	Rented	500,000	Daily	Smooth
Freelance	Medium	Rented	400,000	Daily	Congested
...	...	...	...	...	...
Permanent	High	Own House	1,000,000	Daily	Smooth

To calculate the standard deviation, use the following formula:

$$\sigma = \sqrt{\frac{(nl - mean)^2 + (n2 - mean)^2 + (n3 - mean)^2 + \cdots}{Data\ Quantity - 1}}$$

$$\sigma|Smooth = \sqrt{\frac{(500000-540740{,}741)^2+(500000-540740{,}741)^2+(300000-540740741)^2\ldots\ldots}{54-1}}$$

$$= 231906{,}39$$

$$\sigma|Congested = \sqrt{\frac{(300000 - 483870{,}968)^2 + (500000 - 483870{,}968)^2 + (300000 - 483870{,}968)^2 + \cdots}{31 - 1}}$$

$$= 175303{,}12$$

Based on the above calculations, the mean and standard deviation values are shown in Table 4 below:

Table 4. Mean and Standard Deviation Values

Classification	Mean	Standard Deviation
Smooth	540,740.741	231,906.39
Congested	483,870.968	175,303.12

4.3 Calculating Prior Probability Values

The next stage is to calculate the prior probability values, or the probability values for each category within the attributes of each class, once the mean and standard deviation values have been determined. There are 54 data points in the 'Smooth' class and 31 data points in the 'Congested' class. The data for each attribute in the same class and category should be added together, and the total number of data points in each attribute and class should be divided by this number to determine the prior probability values. Table 5 shows the details below:

Table 5. Prior Probability Values

Attributes	Total Data	Number of Occurrences	Probability
		Smooth	Congested
---------------	----------- -	--------	-------------- -
Total	85	54	31
Job			
Permanent	23	20	3
Freelance	62	34	28
Income			
High	15	14	1
Medium	50	37	13
Low	20	3	17
House Status			
Own House	57	41	16
Rented	17	6	11
Parent's House	11	7	4
Loan Type			
Daily	71	44	27
Weekly	14	10	4

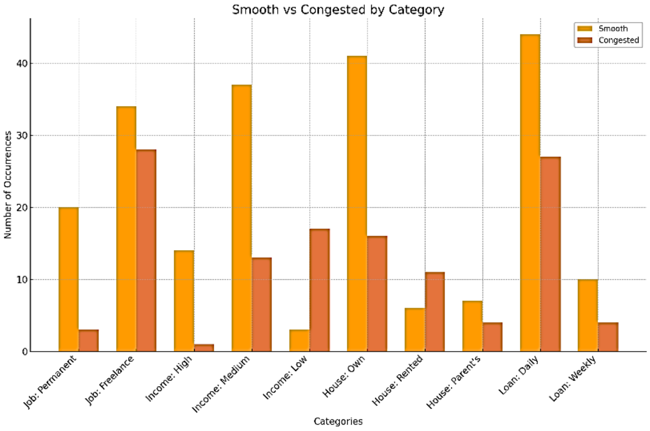

Fig.2: Analysis of Smooth and Congested Occurrences

4.4 Calculating Gaussian Values

The next step is to calculate the Gaussian values for the loan amount attribute in the testing data. Gaussian values are computed specifically for numerical or numeric data types. The testing data can be seen in Table 6 below:

Table 6. Testing Data

Occupation	Income	House Status	Loan Amount	Loan Type	Prediction
Permanent	Medium	Parent's House	500,000	Weekly	?
Freelance	Low	Own House	400,000	Daily	?

Gaussian values are calculated for each class. To find the Gaussian value, we can use the following formula:

$$\vartheta(x, \sigma, \mu) = \frac{1}{\sqrt{2\pi}\,.\,\sigma} exp \frac{-(x - \mu)^2}{2\sigma^2}$$

$$\vartheta(500000|Smooth) = \frac{1}{\sqrt{2\pi}.231906{,}39} exp \frac{-(500000 - 540740{,}741)^2}{2(231906{,}39)^2}$$

$$= 0{,}000001747$$

$$\vartheta(500000|Congested) = \frac{1}{\sqrt{2\pi}\,.\,175303{,}12} exp \frac{-(500000 - 483870{,}968)^2}{2(175303{,}13)^2}$$

$$= 0{,}000002286$$

Above, the Gaussian values for the loan amount attribute in the first set of testing data have been obtained. Perform the same procedure as above to calculate Gaussian values for the subsequent testing data.

4.5 Calculating Probability Values for Each Class

The next step is to calculate the probability values for each class in the testing data. Probability values can be calculated by multiplying all prior probability values according to the categories in the testing data.

P(X|Smooth)=P(Job|Permanent|Smooth)×P(Income|Medium|Smooth)×P(Home Status|Parents|Smooth)×P(Loan Amount=500000|Smooth)×P(Loan Type|Weekly|Smooth)×P(Classification|Smooth) …………… (1)

P(X|Smooth)=0.3704×0.6852×0.1296×0.000001747×0.1852×0.6353=0.00000000676

P(X|Congested)=P(Job|Permanent|Congested)×P(Income|Medium|Congested)×P(Home Status|Parents|Congested)×P(Loan Amount= 500000|Congested)×P(Loan Type|Weekly|Congested)×P(Classification|Congested) …………… (2)

P(X|Congested)=0.0968×0.4194×0.1290×0.000002286×0.1290×0.3647=0.00000000056

The last step is to normalize the probabilities to obtain a total of 1.

P(X|Smooth)=0.00000000676 / (0.000000000560 + 00000000676)=0.9235

P(X|Congested)=0.00000000056 / (0.000000006760 + 00000000056)=0.0765

To determine which class the testing data belongs to based on the highest probability, and according to the calculations above, it appears that the higher probability is associated with the class (Classification | Smooth). Therefore, it can be concluded that the prediction for the first testing data point falls into the "Smooth" class. To obtain predictions for the remaining testing data, repeat the steps outlined earlier. The overall prediction results for implementing this process in RapidMiner with 10 testing data points can be seen in Table 7 below:

Table 7. Prediction Results

Classification	Prediction	Confidence (Smooth)	Confidence (Congested)	Occupation	Income	Housing Status	Loan Amount	Loan Type
Smooth	Smooth	0.703	0.297	Permanent	Moderate	Parents-owned	500000	Weekly
Congested	Congested	0.418	0.582	Miscellaneous Low	Self-owned	400000	Daily	
Congested	Congested	0.089	0.911	Miscellaneous Low	Self-owned	300000	Daily	
Smooth	Smooth	0.668	0.332	Smooth	Smooth	1.000	0.000	Permanent
Smooth	Smooth	1.000	0.000	Permanent	Miscellaneous Medium High	Self-owned	500000	Daily
Congested	Congested	0.091	0.909	Miscellaneous Low	Self-owned	500000	Weekly	
Smooth	Smooth	0.673	0.327	Miscellaneous Medium	Self-owned	300000	Daily	
Congested	Smooth	0.917	0.083	Permanent	Moderate	Self-owned	500000	Daily
Congested	Congested	0.091	0.909	Miscellaneous Low	Self-owned	500000	Weekly	
Smooth	Smooth	1.000	0.000	Permanent	High	Self-owned	1000000	Daily

4.6 Testing Naive Bayes Classifier Method

The testing of the Naive Bayes Classifier method In the Naive Bayes algorithm, testing is conducted using a performance confusion matrix to measure the performance of the method by calculating accuracy, recall, and precision values. The higher the accuracy value, the better the performance. There are 2 classification classes: congested and smooth classes. In this study, testing was carried out using the RapidMiner application. The confusion matrix results of the Naive Bayes method testing for predicting the smoothness of installment payments can be seen in detail in Table 8 below.

Table 8. Confusion Matrix

	True Smooth	**True Congested**	**Class Precision**
Pred.Smooth	5	1	83.33%
Pred.Congested	0	4	100.00%
Class Recall	100.00%	80.00%	

Based on table 8 above, shows that testing the Naive Bayes method using the RapidMiner application achieved an accuracy of 90.00%. Out of the 10 testing data points, 5 people were correctly predicted as "Smooth" (True Positive), 4 people were correctly predicted as "Congested" (True Negative), 1 person was incorrectly predicted as "Smooth" when they were "Congested" (False Negative), and there were no False Positives or people predicted as "Smooth" when they were "Congested".

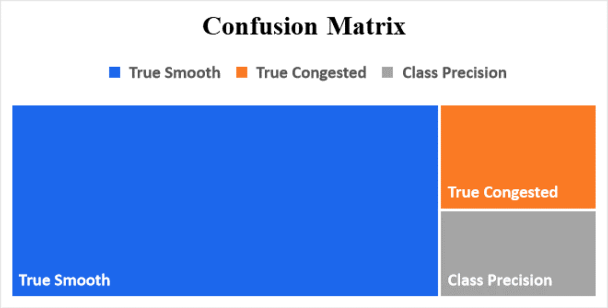

Fig. 3: Confusion Matrix

The accuracy value can be manually calculated using the formula:

Accuracy $= \frac{(5+4)}{(5+0+4+1)} \times 100\% = 90\%$

Precision $= \frac{5}{(5+1)} \times 100\% = 83.33\%$

Recall $= \frac{5}{(5+0)} \times 100\% = 100\%$

5. CONCLUSION

The Naive Bayes classifier has demonstrated its effectiveness in predicting the repayment behaviour of loan applicants using a dataset comprising 85 training data points and 10 testing data points. Leveraging 5 attributes as predictors and 1 attribute as the label, both manual calculations and the RapidMiner application consistently achieved a high accuracy of 90.00%. Additionally, the model attained impressive metrics with an Area Under the Curve (AUC) of 0.880%, Recall (Sensitivity) of 83.33%, and Precision of 100%, underscoring its robust performance in accurately identifying both positive and negative instances. Despite these achievements, several limitations warrant attention for future research to enhance the model's efficacy and applicability. Firstly, the quality and scope of data significantly influence predictive accuracy, necessitating efforts to expand and refine datasets for better representation and diversity. Secondly, Naive Bayes' assumption of independence among predictor variables may not always hold in complex real-world scenarios, prompting exploration into advanced models capable of handling attribute interdependencies for improved predictive performance. Addressing class imbalances within datasets remains crucial, with techniques like data resampling and alternative evaluation metrics being pivotal in mitigating biases and promoting model fairness. Continual refinement of feature engineering techniques is essential to capture nuanced relationships and enhance model interpretability. Furthermore, validating the model across various datasets and periods is essential to gauge its generalization capabilities and reliability in diverse contexts. By addressing these limitations and exploring innovative avenues, Naive Bayes classifiers can evolve to offer more accurate insights, empowering decision-makers in the lending industry with enhanced risk management strategies and informed lending practices.

6. REFERENCES

[1] Kusrini, Kusrini., M., Rudyanto, Arief. (2023). Komparasi Algoritma Naïve Bayes Dan Support Vector Machine (SVM) Untuk Klasifikasi Kelayakan Pemberian Pinjaman. Infotek, doi: 10.29408/jit.v6i2.20059

[2] Sugeng, Riyadi., Muhammad, Mizan, Siregar., Khairul, fadhli, Fadhli, Margolang., Karina, Andriani. (2022). Analysis of svm and naive bayes algorithm in classification of nad loans in save and loan cooperatives. JURTEKSI (Jurnal Teknologi dan Sistem Informasi), doi: 10.33330/jurteksi.v8i3.1483

[3] Awuza, Abdulrashid, Egwa. (2022). Default Prediction for Loan Lenders Using Machine Learning Algorithms. SLU Journal of Science and Technology, doi: 10.56471/slujst.v5i.222

[4] Ngo, Tien, Luu., Phan, Duy, Hung. (2021). Loan Default Prediction Using Artificial Intelligence for the Borrow - Lend Collaboration.. doi: 10.1007/978-3-030-88207-5_26

[5] Pham, Thanh, Binh., Nguyen, Dinh, Thuan. (2022). Predicting Loan Repayment Using a Hybrid of Genetic Algorithms, Logistic Regression, and Artificial Neural Networks. doi: 10.1007/978-981-19-8069-5_11

[6] Riktesh, Srivastava. (2022). Extrapolation of Loan Default using Predictive Analytics: A Case of Business Analysis. Samvad, doi: 10.53739/samvad/2021/v23/166261

[7] A. Alwi, P. Studi, T. Informatika, F. Teknik, and U. M. Ponorogo, “the Concept of Naive Bayes and Its Simple Use for Prediction Final Konsep Naive Bayes Dan Penggunaannya Secara Sederhana,” J. Tek. Inform., vol. 3, no. 1, pp. 133– 140, 2022.

[8] D. A. Kurniawan and Y. I. Kurniawan, “Aplikasi Prediksi Kelayakan Calon Anggota Kredit Menggunakan Algoritma Naïve Bayes,” J. Teknol. dan Manaj. Inform., vol. 4, no. 1, 2018, doi: 10.26905/jtmi.v4i1.1831.

[9] M. Guntur, J. Santony, and Y. Yuhandri, “Prediksi Harga Emas dengan Menggunakan Metode Naïve Bayes dalam Investasi untuk Meminimalisasi Resiko,” J. RESTI

(Rekayasa Sist. dan Teknol. Informasi), vol. 2, no. 1, pp. 354– 360, 2018, doi: 10.29207/resti.v2i1.276.

[10] Pathak, A., Chakraborty, A., Rahaman, M., Rafa, T. S., & Nayema, U. (2024). Enhanced Counterfeit Detection of Bangladesh Currency through Convolutional Neural Networks: A Deep Learning Approach. International Journal of Innovative Research in Computer Science & Technology, 12(2), 10–20. https://doi.org/10.55524/ijircst.2024.12.2.2

[11] S. A. Rizky, R. Yesputra, and S. Santoso, "Prediksi Kelancaran Pembayaran Cicilan Calon Debitur Dengan Metode K-Nearest Neighbor," JURTEKSI (Jurnal Teknol. dan Sist. Informasi), vol. 7, no. 2, pp. 195–202, 2021, doi: 10.33330/jurteksi.v7i2.1078.

[12] S. S. Khautsar, D. Puspitasari, and P. wida Mustika, "Algoritma Naïve Bayes Untuk Memprediksi Kredit Macet Pada Koperasi Simpan Pinjam," J. Inform., vol. 4, no. 2, 2018.

[13] Arna Chakraborty, Arnab Chakraborty, Abdus Sobhan, Abhijit Pathak. Deep Learning for Precision Agriculture: Detecting Tomato Leaf Diseases with VGG-16 Model. International Journal of Computer Applications. 186, 19 (May 2024), 30-37. DOI=10.5120/ijca2024923599

[14] A. Muzaki and A. Witanti, "Sentiment Analysis of the Community in the Twitter To the 2020 Election in Pandemic Covid-19 By Method Naive Bayes Classifier," J. Tek. Inform., vol. 2, no. 2, pp. 101–107, 2021, doi: 10.20884/1.jutif.2021.2.2.51.

[15] Corporation, N. C. D. (2012). Annual Report - National Cooperative Development Corporation. http://books.google.ie/books?id=LLTXXZjl_rEC&q=Grameen+Multipurpose+Co-Operative+Society+Ltd&dq=Grameen+Multipurpose+Co-Operative+Society+Ltd&hl=&cd=3&source=gbs_api

[16] M. E. Lasulika, "Komparasi Naïve Bayes, Support Vector Machine Dan KNearest Neighbor Untuk Mengetahui Akurasi Tertinggi Pada Prediksi Kelancaran Pembayaran Tv Kabel," Ilk. J. Ilm., vol. 11, no. 1, pp. 11–16, 2019, doi: 10.33096/ilkom.v11i1.408.11-16.

[17] F. Ariadi, "Analisa Perbandingan Algoritma DT C.45 dan Naïve Bayes Dalam Prediksi Penerimaan Kredit Motor," KERNEL J. Ris. Inov. Bid. Inform. dan Pendidik. Inform., vol. 1, no. 1, pp. 1–8, 2020, doi: 10.31284/j.kernel.2020.v1i1.1183.

[18] M. Sadikin, R. Rosnelly, R. Roslina, and ..., "Penerapan Data Mining Pada Penerimaan Dosen Tetap Menggunakan Metode Naive Bayes Classifier dan C4. 5," J. Media …, vol. 4, no. 4, pp. 1100–1109, 2020, doi: 10.30865/mib.v4i4.2434.

[19] Y. I. Kurniawan, A. Fatikasari, M. L. Hidayat, and M. Waluyo, "Prediction for Cooperative Credit Eligibility Using Data Mining Classification With C4.5 Algorithm," J. Tek. Inform., vol. 2, no. 2, pp. 67–74, 2021, doi: 10.20884/1.jutif.2021.2.2.49.

[20] Hossen, N. H., Shuvon, N. M. S. S., Barsha, N. J. B., Chy, N. a. A., & Pathak, N. A. (2023). Ultimate cricket experience: Dynamic web app for a real-time scoring system in university cricket. World Journal of Advanced Research and Reviews, 19(2), 1269–1280. https://doi.org/10.30574/wjarr.2023.19.2.1721.

[21] E. Sutoyo and A. Almaarif, "Educational Data Mining untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritme Naïve Bayes Classifier," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 4, no. 1, pp. 95–101, 2020, doi: 10.29207/RESTI.V4I1.1502