International Journal of Advanced Science and Computer Applications (2023); 2(2):
DOI: 10.47679/ijasca.v2i2.35
https://ijasca.org/index.php/ijasca

RESEARCH ARTICLE

Comparison of K-Nearest Neighbor, Naive Bayes Classifier, Decision Tree, and Logistic Regression in Classification of Non-Performing Financing

Rizky Ferdiansyah Putra[1,*)], Iis Dewi Ratih[2]

Published online: 14 September 2023

Abstract

The Non-Performing Financing (NPF) indicator of one of the Islamic Banks in Indonesia in the 1st to 3rd quarter of 2021 in a row of 9.69%; 9.97%; 9.46%. The NPF movement tends to improve slightly from time to time but still exceeds the maximum limit stipulated in Bank Indonesia's Regulation Number 23/2/PBI/2021, which is no more than 5%. This shows that the Islamic bank has a financing performance that can be said to be less good. Preventive steps that can be taken to reduce the NPF ratio in order to improve the health of the bank is to classify prospective financing customers. This research was conducted using the K-Nearest Neighbor (KNN), Naive Bayes Classifier (NBC), Decision Tree, and Logistics Regression classification methods to predict potential financing customers. The dataset is divided into 80% training and 20% testing. It was found that the best classification result was the Naive Bayes Classifier in the proportion of distribution of 80% training data and 20% testing data with an accuracy value of 84.69%, sensitivity of 58.25%, and specificity of 90.16%.

Keywords: Classification, Decision Tree, K-Nearest Neighbor, Logistic Regression, Naive Bayes Classifier, Non-Performing Financing.

Introduction

Financing is an activity of providing funding support for the needs of procurement of certain goods / assets / services which generally involves three parties, namely the funder, the provider of goods / assets / services, and the recipient of funds. In carrying out this financing, of course, there are risks, namely the failure of the party who is given the loan in fulfilling their obligations, or what is referred to as non-performing financing. Non-performing financing occurs when in the financing there are arrears of principal installments and or profit sharing / margin [1]. The arrears will have an impact on the Non-Performing Financing (NPF) ratio. NPF is the ratio between non-performing financing and total financing. This ratio describes the risks that must be borne by BPRS in financing. NPF in BPRS from the second quarter to the third quarter of 2021 decreased by 0.51% from 9.97% to 9.46%, but the value still exceeded the limit set by Bank Indonesia Regulation Number 23/2/ PBI/ 2021, which is 5%.

Unhealthy financing performance makes BPRS obliged to conduct a financing analysis before realizing financing. However, the problem is that financing analysis takes a long time because the bank has to go to the field directly to see the condition of prospective customers whether it is feasible to be given financing or not, so a method is needed to classify customers.

Predictions of non-performing financing customers are carried out using the KNN, NBC, Decision Tree, and Logistic Regression methods. KNN is a classifying method based on the similarity of a new object to an object whose label is already known [2]. NBC was chosen because of its very simple algorithm yet produces good accuracy [3]. Decision Tree is a classification method based on a decision tree. This method is robust against the emergence of noise and accuracy remains good even though the number of predictors used is large [4].

[1*),2]Department of Business Statistics, Faculty of Vocational, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

*) *corresponding author*

Rizky Ferdiansyah Putra

Email: rizkyfp15@gmail.com

Received 1 May 2023 | Received in revised form 15 Juli 2023 | Accepted 2 August 2023
Available online 14 September 2023

Logistic regression is a statistical method to determine the causal relationship between predictor variables (numerical and categorical) and response variables (categorical) [5].

Based on the background, in this study, the classification of problem financing customers will be carried out using the K-Nearest Neighbor, Naive Bayes Classifier, Decision Tree, and Logistic Regression methods. All of these methods will be compared and selected one of the best models to classify potential new financing customers.

This research paper will be divided into several parts. Part 2 explains about related work. Part 3 explains about the brief theory of the methodology to be applied. Part 4 explains about the characteristics of financing customers and comparisons between methods. Part 5 explains the conclusions in the form of the best classification models selected in this study.

Related Work

Annur& Efendi Lasulika in a study entitled Classification of Cooperative Credit Customers Using the K-Nearest Neighbor Algorithm stated that using K=1, a model accuracy of 77.78% was obtained [2]. Ciptohartono with the research title Naive Bayes Classification Algorithm for Assessing Credit Worthiness stated that the Naive Bayes Classifier method was able to produce an accuracy of 92.53% [6]. Sarimuddin et al in their research entitled Classification of Data on Aging Arrears of Customers Using the Decision Tree Method at ULaMM Unit Kolaka stated that decision trees are able to produce model accuracy of 95% [4]. Islahulhaq in his research entitled Classification of Non-Performing Financing Customers in PT. BPRS Lantabur Tebuireng Using Logistic Regression – Synthetic Minority Over-Sampling Technique states that the binary logistic regression model combined with the Synthetic Minority Over-Sampling Technique has an accuracy of 81% [7].

Method

Dataset

The data used in this study is secondary data, namely financing customer data recorded in one of BPRS's Information and Technology (IT) department. The initial data of 3258 was preprocessed so that the final data became 3004 data. The data is divided into 80% (2403 data) training and 20% (601 data) testing. The variables used in this study can be seen in Table 1.

Tabel 1: Research Variables

Variable	Description	Category
Y	Financing Quality	1: Good Financing 2: Non-Performing Financing
X_1	The Amount of Financing (Rupiah)	-
X_2	Type of Use	1: Working Capital 2: Investment 3: Consumption
X_3	Financing Period (Months)	-
X_4	Gender	1: Male 2: Female
X_5	Marriage Status	1: Married 2: Single 3: Separated
X_6	Latest Level of Education	1: ≤ Elementary School/ Equivalent 2: Junior High School/ Equivalent 3: Senior High School/ Equivalent 4: 1st Diploma 5: 2nd Diploma

Table 1: Research Variables (cont)

Variable	Description	Category
		6: 3rd Diploma 7: Bachelor Degree 8: Master Degree 9: Doctoral Degree
X_7	Type of Financing	1: Murabahah 2: Musyarakah 3: Multijasa 4: Mudharabah 5: Qordhiyu

X_8	Type of Collateral	1: Deposits/ Savings 2: 2-Wheeled Vehicles (BT) 3: 2-Wheeled Vehicles (FE0) 4: 2-Wheeled Vehicles (Power of Attorney) 5: 2-Wheeled Vehicles (Notary) 6: 4-Wheeled Vehicles (BT) 7: 4-Wheeled Vehicles (FE0) 8: 4-Wheeled Vehicles (Notary) 9: Institutional Decision Letter 10: Land Assets (APHT) 11: Land Assets (BT) 12: Land Assets (Power of Attorney) 13: Land Assets (Notary) 14: Tanah (SKMHT) 15: Others
X_9	Occupation	1: Labor 2: Medical Personnel 3: Law 4: Housewife 5: Consultant/ Analyst 6: Enterpreneur 7: Marketing 8: Military 9: Fisherman 10: Government Employee 11: Informal Employee 12: Art Worker 13: Student/ College Student 14: Teacher/ Lecturer 15: Retired 16: Hotel & Restaurant 17: Carpenter & Craftsman 18: Farmer 19: Breeder 20: Broker 21: Others
X_{10}	Age (years)	-

K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a classification algorithm based on similarity values. The similarity value is obtained by comparing the distance between training and testing data [8]. The most commonly used distance is the Euclidean distance [3]. The Euclidean distance equation can be seen in Equation 1.

$$d(x,y)=\sqrt{\sum_{i=1}^{n}\left(x_i-y_i\right)^2} \quad (1)$$

Description:
d(x,y) = The distance between x data to y data
x_i = i^{th} testing data
y_i = i^{th} training data

Naive Bayes Classifier

Naive Bayes Classifier (NBC) is a simple probability-based classification method that in calculating the probability of its occurrence comes from the number of frequencies and combinations of a given dataset. In other words, NBC can predict the probability of an event in the future based on past data [9]. The general equation of NBC can be shown in Equation 2.

$$P(Y\mid X)=\frac{P(X\mid Y).P(Y)}{P(X)} \quad (2)$$

Description:
P(Y|X) = The probability of *Y* condition when it is in class *X*
P(X|Y) = The probability of *X* condition when it is in class *Y*
P(X) = The prior probability of *X*
P(Y) = The prior probability of *Y*
If the type of data is numerical, then gaussian distribution calculations are used according to Equation 3.

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

Decision Tree

Decision Tree is an algorithm that converts training data into a decision tree, so that patterns / information can be clearly known in the form of decision trees [10]. This algorithm consists of a collection of nodes connected by branches and then moving from the root (the top node) to the bottom to the leaf node. Leaf nodes that can no longer be broken down are the result of predictions from testing data that the class is to know [3].

Logistic Regression

Logistic regression is a method for determining the model of the relationship between a predictor variable (numerical/ categorical) and a response variable (categorical) [11]. A logistic regression model whose response variables consist of 2 categories ("success" and "failure") is called binary logistic regression [5]. The general form of the binary logistic regression opportunity model can be seen in Equation 4.

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Model Evaluation

The evaluation of the model is based on the confusion matrix. Confusion matrix is a matrix-shaped table used to measure the performance of the classification model [12]. The general form of the confusion matrix table can be seen in Table 2.

Table 2: Table of Confusion Matrix

Actual	Predicted		Total
	Positive	Negative	
Positive	True Positive (TP)	False Negative (FN)	TP+FN
Negative	False Positive (FP)	True Negative (TN)	FP+TN
Total	TP+FP	FN+TN	TP+FP+FN+TN= n

The four values contained in Table 2 produce several indicators for measuring the performance of the classification model which can be seen in Table 3.

Table 3: Equations of *Confusion Matrix*

Indicator	Formula
Accuracy: Percentage of observations that can be correctly classified by the model	$\frac{(TP+TN)}{n}$ (5)
Sensitivity: Percentage of positive observations correctly classified as positive	$\frac{TP}{(TP+FN)}$ (6)
Specificity: Percentage of negative observations correctly classified as negative	$\frac{TN}{(TN+FP)}$ (7)

The Proposed Method

The series of analysis processes from preprocessing data to determining the best model can be described as follows.

1. Preprocessing financing transaction data which includes:

Data *Cleaning*

Discard duplicated financing customer data and delete observations containing missing values. Dispose of customer data that has more than one type of guarantee and collateral that is not proportional in value to the nominal financing and term.

Data *Selection*

Delete data on financing customers who have more than one type of guarantee.

Data *Transformation*

The values on the variables need to be re-encoded to fit into predetermined categories.

2. Describe the characteristics of financing quality based on financing transaction data.
3. Divide the data into training data and testing data with a ratio of 80:20.
4. Form a classification model using the K-Nearest Neighbor, Naive Bayes Classifier, Decision Tree, and LogisticRegression methods.
5. Measure model performance using confusion matrix tables based on accuracy, sensitivity, and specificity.
6. Choose the best model based on accuracy, sensitivity and specificity.
7. Draw conclusions and suggestions from the results of the analysis.

Methods

This section discusses the results of a descriptive analysis of customer characteristics and performance results for models using the KNN, NBC, Decision Tree, and Logistic Regression methods. The data used for data characteristics and the formation of classification models are 3004 preprocessed data.

The Characteristics of Financing Customers

The general characteristics of financing customers in the form of a proportion of customers according to the quality of financing visually can be seen in the pie chart Figure 1.

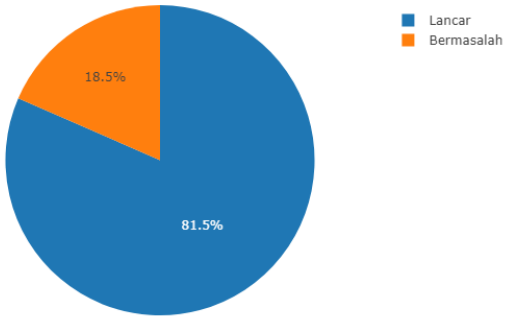

Fig 1: Proportion of Customers According to Financing Quality

Figure 1 can be seen that 81.5% (2,449 people) are classified as the good quality of financing, while the other 18.5% (555 people) are classified as the quality of non-performing financing, so that most of the financing customers are dominated by the good financing quality.

Classification of Financing Customers

The classification model of problem financing customers was formed using the KNN, NBC, Decision Tree, and Logistic Regression methods by including all predictor variables. The selection of the best model prioritizes high sensitivity indicators due to data imbalances, namely the number of data classes of problem customers is much lower than that of current customers, but still considers accuracy and specificity. Model performance results for each method can be seen in Table 4.

Table 4: Model Performance Results of Each Method

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
K-Nearest Neighbor	86.36	39.81	95.98%
Naive Bayes Classifier	**84.69**	**58.25**	**90.16%**
Decision Tree	87.02	41.75	96.39%
Regresi Logistik	85.86	42.72	88.88%

Table 4 shows that when viewed based on accuracy and specificity, the best model obtained is the Decision Tree with an accuracy of 87.02% and a specificity of 90.16%, while if based on sensitivity, the best model obtained is the Naive Bayes Classifier with a sensitivity of 58.25%. The best overall model performance is produced by the Naive Bayes Classifier model. This model was chosen because it has the highest sensitivity but with accuracy and specificity that is not much different compared to other models. The Naive Bayes Classifier model is able to produce an accuracy of 84.69% which means that 84.69% of customers as a whole can be classified correctly, sensitivity of 58.25% which means that 58.25% of non-performing financing customers are correctly classified as non-performing customers, and a specificity of 90.16% which means that 90.16% of good financing customers are correctly classified as good financing customers. Indicators of accuracy, sensitivity, and specificity were measured based on the Confusion Matrix results of entering 601 testing data into the model. The Confusion Matrix is presented in Table 5.

Table 5: Confusion Matrix of NBC Model

		Actual		Total
		Non-Performing Financing	Good Financing	
Prediksi	Non-Performing Financing	60	49	109
	Good Financing	43	449	492
	Total	103	498	601

Table 5 shows that of the 498 good financing customers, 449 of them were able to be predicted precisely as good financing customers and of the 103 non-performing financing customers. 60 of them are capable of precisely being non-performing financing customers.

Conclusions and Recommendations

Financing customers are dominated by the current category with the male gender. The Naive Bayes Classifier model was chosen as the best classification model because it has better sensitivity than other methods, but accuracy and specificity are not much different from other methods.

The suggestion for the next research is to use methods to overcome data imbalances and try more advanced methods and add other predictor variables, so as to obtain better model performance.

Declarations

All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

Ethical Approval

Not applicable as the research is using secondary datasets provided by DOE and DOS, Malaysia.

Consent to Participate

Not applicable as the research is using secondary datasets provided by DOE and MOH, Malaysia.

Consent for publication

The consent to publish is granted from the DG of health, Malaysia.

Competing Interests

The authors declare no competing interests.

References

[1] W. b. M. Cokrohadisumarto., A. G. Ismail., & K. A. Wibowo. (2016). BMT: Praktik dan Kasus. Rajawali Pers. Jakarta.

[2] Annur, Haditsah & Efendi Lasulika, Moh. (2019). Klasifikasi Nasabah Kredit Koperasi Menggunakan Algoritma K-Nearest Neighbor. *Jurnal Informatika Upgris*. 5(2): 126-129.

[3] J, Suntoro. (2019). Data Mining: Algoritme dan Implementasi Menggunakan Bahasa Pemrograman PHP. PT. Elex Komputindo. Jakarta

[4] Sarimuddin et al. (2020). Klasifikasi Data Aging Tunggakan Nasabah Menggunakan Metode Decision Tree pada ULaMM Unit Kolaka. *Informatics Journal*. 5(1): 26-32.

[5] Alwi, Wahidah., Ermawati, & Husain, Saddam. (2018). Analisis Regresi Logistik Biner untuk Memprediksi Kepuasan Pengunjung pada Rumah Sakit Umum Daerah Majene. *Jurnal MSA*. 6(1): 20-26.

[6] Ciptohartono, Claudia Clarentia. (2014). Algoritma Klasifikasi Naive Bayes untuk Menilai Kelayakan Kredit. Universitas Dian Nuswantoro. Semarang.

[7] Islahulhaq. Wibowo. W.. & Ratih. I. D. (2021). Classification of Non-Performing Financing Using Logistic Regression and Synthetic Minority Over-sampling Technique-Nominal Continuous (SMOTE-NC). *Int. J. Advance Soft Compu. Appl.* 13(3): 115-128.

[8] Prasetyowati, Erwin. (2017). Data Mining: Pengelompokan Data untuk Informasi dan Evaluasi. Duta Media Publishing. Pamekasan.

[9] Sidiq. Y. N.. Fathonah. R. N.. & Riza. N. (2020). Metode Klasifikasi Menentukan Kenaikan Level UKM Bandung Timur Dengan Algoritma Naive Bayes pada Sistem JURAGAN Berbasis Komunitas. Kreatif Indonesia Nusantara. Bandung.

[10] Bahri, Syaiful & Lubis, Akhyar. (2020). Metode Klasifikasi Decision Tree untuk Memprediksi Juara English Premiere League. *Jurnal Sintaksis*. 2(1): 63-70.

[11] Nugraha, Jaka. (2014). Pengantar Analisis Data Kategorik. Deepublish. Yogyakarta.

[12] Faisal. M. R. & Nugrahadi. D. T. (2019). Belajar Data Science: Klasifikasi dengan Bahasa Pemrograman R. Scripta Cendekia. Banjarbaru.

Notes on Contributors

Rizky Ferdiansyah Putra, He is a student at the Department of Business Statistics, Faculty of Vocation, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. He has experience as a lecturer assistant for 2 years. His recent project is about people's salt price survey this year.

Iis Dewi Ratih. She is an Associate Lecturer at the Department of Business Statistics, Faculty of Vocation, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. Her main research is on statistical modeling, multivariate, and extreme value theory.