Journal of Actuarial, Finance and Risk Management (JAFRM)
Vol. 1(2), November, 2022.

Prediction of Loan Status Using Logistics Regression Model and Naïve Bayes Classifier

Christabell[1], Edwin Setiawan Nugraha[2*], Karunia Eka Lestari[3]

[1,2] *President University, Jl. Ki Hajar Dewantara, Mekarmukti, Cikarang Utara, Bekasi Regency, 17550, Indonesia*
[3]*Mathematics Education Departement, Universitas Singaperbangsa Karawang*
[*]*Corresponding author: edwin.nugraha@president.ac.id*

Abstract—Conducting an evaluation process of prospective debtors is important for creditors to reduce the risk of default. For this reason, the research aims to construct a model that can determine whether a prospective applicant's credit application is recommended to be accepted or rejected by using the method of logistic regression and naïve Bayes classifier. We used a dataset of gender, married, dependent, education, self-employed, applicant income, co-applicant income, loan amount, loan amount term, credit history, and property area as predictor variables and loan status as a response variable. The results show that the performance measures, including accuracy, precision, recall, and F1 score of the logistics regression method, are 85.9%, 83.82%, 100%, and 91.2%, while the naïve Bayes classifier is 84.62%, 83.58%, 98.2%, and 90.32%. Since the performance measures of logistic regression are bigger than naïve Bayes classifier, it suggests that logistic regression is better than naïve Bayes classifier.

Keywords—Categorical Data; Credit Risk; Logistic Regression; Naïve Bayes

I. INTRODUCTION

Credit is a financial facility that allows a person or business entity to borrow money to buy a product and pay it back within a specified period of time and interest. Credit has a critical role in finances, both for creditors and debtors. Credit aims to assist the availability of funds to finance national production activities, storing materials, financing credit sales, transportation of goods, and trading activities. The role of credit is quite dominant in a developing country to develop economic potential [1]. The provision of smooth credit will develop and improve the financial activities of a nation. The position of creditors is very vulnerable because of the provision of credit, which contains a "Degree of Risk" which does not rule out the possibility of bad credit [2].

Bad credit is uncollectible receivables or credit that have substandard criteria, doubtful because they have difficulty paying off due to certain factors. Bad credit occurs when the creditor finds it challenging to ask for installments from the debtor for some reason. The possibility of bad credit is included in credit risk. Credit risk is a risk of loss caused by the inability of the debtor to fulfill their obligations to pay debts, both principal and interest payable. Therefore, providing credit to prospective debtors is an important decision that must be carefully considered. The creditor needs a model that can classify whether prospective debtors have a possible default or not. The prediction results of the model are not always correct, but it can minimize the possibility of lenders providing credit to applicants who will default.

There were various differences and similarities between this research and previous research. There are some related studies from Ernest [3], Ginting et al. [4], Sunitha et al. [5], and Tabagari [6] are to examine how the factors that influence and predict the results of a classification in credit criteria. Several independent variables used also intersect, such as gender, marital status, education level, number of dependents, type of jobs, income, and credit history. In contrast, the differences are that this research analyzes logistic regression and naïve Bayes classifier and compares classification results. There is no research comparing the classification results of logistics regression and naïve Bayes classifier in the credit sector.

This paper aims to determine which factors affect the approval or rejection of credit applications by debtors using the logistics regression model and naïve Bayes and how accurate the two models are in predicting the possibility that the applicant will fail to pay or not. In this paper, the credit criteria that will be analyzed are short term credit, where the loan term is less than or equal to one year (360 days). The structure of this paper is as follows. Section II explains the methodology used in this research, namely logistics regression and naïve Bayes classifier. Section III demonstrates the process of generating a model for this research and shows the results obtained. Section IV is the conclusion of this research.

II. METHOD

A. Research Method

Qualitative and quantitative methodologies are the two most frequently used in scientific research. This research utilized the quantitative method, which involved collecting measurable data and applying statistical

analysis. The researcher had to get and calculate the data, perform statistical analysis, and interpret the result. A quantitative methodology likewise requires the researcher to construct hypotheses based on the theory. The statistical analysis result would determine whether the hypotheses were accepted or rejected. The influence or impact of independent variables on the dependent variable can be studied using quantitative techniques [7]. In this research, the methods used are the logistics regression method and naïve Bayes with RStudio.

B. Research Flowchart

To start the research analysis, researchers clean the data to remove incomplete or not available data that can affect the analysis process, input the clean data and divide the data into 2 types; training and testing data. For the flow of the analysis using logistic regression, the researcher inputs the training data before performing the likelihood ratio test. Likelihood ratio test aims to determine whether the independent variables contained in the model have a significant effect on the entire. The next step is to do a Z test to see whether any of the model's independent variables has a significant or partial effect on the dependent variable. After obtaining which independent variables affect the dependent variable, goodness of fit is carried out to test how effectively the dependent variable can be classified by the independent variable. The final logistic regression model has been obtained if the result of goodness of fit show that the model fits the data. Next, input the testing data into the final model, and classify how accurate the results of the data testing's prediction with actual testing data.

For the analysis flow using naïve Bayes, the researcher input training data, and calculates the previous probability [P(C)] and attribute probability for each class [$P(F_1 \ldots F_n)|C$]. Next, calculate the multiplication of the probability with the probability attribute in each class [P(C) P(F1…Fn)|C] and find the maximum value of the multiplication result for the naïve bayes' model. After getting the result, the researcher input the testing data into the model and classify how accurate the prediction results of the testing data with the actual testing data.

The steps above are illustrated in the flowchart in Figure 1.

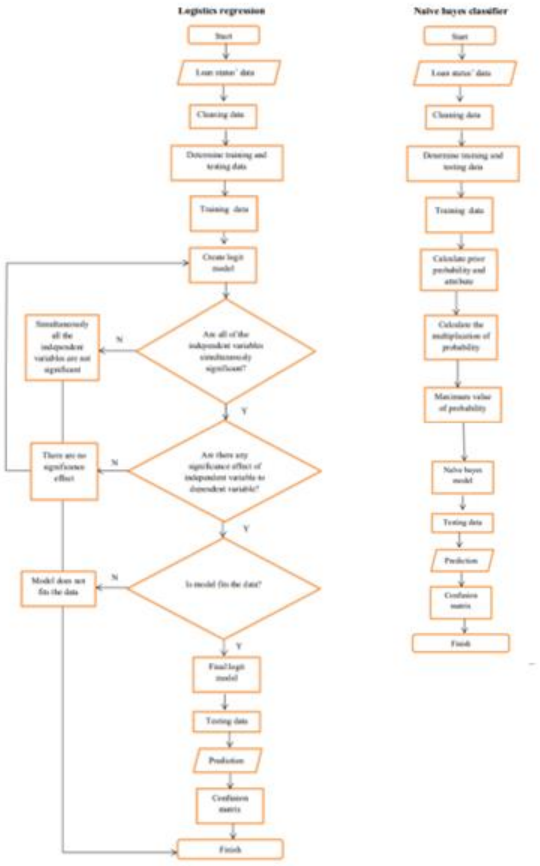

Figure 1. Research Flowchart

C. Sampling Design

1) Training Data

Training data is previously dataset where we know all the attributes, including the objective class property. Training data is utilized to shape a predictive mode. In this paper, 75% of the total data is training data.

2) Testing Data

Testing data is another data set that used to check the validity of model. In this paper, 25% of the total data is testing data.

D. Data Analysis Method

1) Logistics regression

Binary logistic regression is a data analysis method that is used to discover the relationship between binary response variables and predictor variables [8]. The response variable y_i has a binary data (value 1 or 0)

$y_i = 1$ if its success,

$y_i = 0$ if its failure,

where $i = 1,2,3,\cdots,n$

In such cases, for each observation, the variable y follows the Bernoulli distribution. The probability function for each observation is as follows [9]:

$$f(y) = \pi^y (1-\pi)^{1-y} \quad y=0,1 \tag{1}$$

with y_i is a realization of a random variable from Y_i. The probability of Y is

$$P(Y_i = 1) = \pi i \text{ and } P(Y_i = 0) = (1-\pi_i) \tag{2}$$

The logistic regression function can be written as follows.

$$f(y) = \frac{e^y}{1+e^y} \tag{3}$$

When $y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$ the value of y is between $-\infty$ and $+\infty$. Therefore the value of $f(y)$ lies between 0 and 1 for any given value of y. This shows that the logistic model actually describes the probability or risk of an object. As indicated by Hosmer and Lemeshow (2000), a logistic regression model with p predictor variables is formed with a value of $(x) = (Y = 1 | x)$.

The logistic regression model is as follows.

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 X_1+...+\beta_p X_p)}}{1+e^{(\beta_0+\beta_1 X_1+...+\beta_p X_p)}} \tag{4}$$

a) Maximum likelihood estimator

Maximum likelihood estimation is a parameter value that maximizes the likelihood function based on the sample data (y). The L(π) function has a maximum value of π =0 when y=0.

Suppose a sample consists of n independent experiments, where i is the response variable from the i-th observation (i = 1, 2, …, n) with binomial distribution with probability of success [$\pi(x)$] and probability of failure [$1-\pi(x)$], and xi is the dependent variable on the i-th observation (i=1,2,…,n) then the probability function for each pair is as follows.

$$P(Y_i = y_i) = \pi_i^{yi} (1-\pi_i)^{1-yi} \tag{5}$$

where $(y_i) = 0,1$, and

$$\pi(x_i) = \frac{e^{(\Sigma_{j=0}^{p} \beta_j x_j)}}{1+e^{(\Sigma_{j=0}^{p} \beta_j x_j)}} \tag{6}$$

where if $j = 0$, then the value of $x_{ij} = x_{10} = 1$

Each pair of observations is assumed to be independent, that the likelihood function is a combination of the distribution functions of each pair as follows:

$$l(\beta) = \Pi_{i=1}^{n} f(x_i) = \Pi_{i=1}^{n} \pi(x_i)^{y_i} (1-\pi(x_i))^{1-y_i} \tag{7}$$

The likelihood function is easier to maximize in log β and it is expressed by (β).

$$L(\beta) = \ln l(\beta)$$

$$= \sum_{i=1}^{n} \{ y_i \ln[\pi(x_i)] + (1-y_i) \ln[1-\pi(x_i)] \}$$
$$= \Sigma_{i=1}^{n} (\Sigma_{i=1}^{n} y_i X_{ij}) \beta_j - \Sigma_{i=1}^{n} \ln(1+e^{(\Sigma_{j=0}^{p} \beta_j x_j)}) \quad (8)$$

The maximum value of is obtained through the derivative (β) and the result is equal to zero.

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij} \left(\frac{e^{(\Sigma_{j=0}^{p} \beta_j x_j)}}{1+e^{(\Sigma_{j=0}^{p} \beta_j x_j)}} \right)$$
$$\Sigma_{i=1}^{n} y_i x_{ij} - \Sigma_{i=1}^{n} x_{ij} \hat{\pi}(x_i) = 0 \quad (9)$$

where $j = 0,1, \ldots, p$.

b) Likelihood ratio test

Likelihood ratio test is utilized to test the feasibility of the model acquired from parameter estimates which expects to decide if the independent variables contained in the model significantly influence on the entire [8]. This test compares the complete model (model with predictor parameters) against the model with only constants (model without predictor parameters) to see whether the model with only constants is significantly better than the complete model with the following formula:

$$G = -2 \ln \left[\frac{\left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n} \hat{\pi}_i^{\,y_i} (1-\hat{\pi}_i)^{(1-y_i)}} \right] \quad (10)$$

n_1 = the quantity of observations class 1

n_0 = the quantity of observations class 0

n = quantity of observations ($n_0 + n_1$)

Hypothesis:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ (predictor variables have no significance on the response variable)

H_1: $\beta_j \neq 0, j = 1,2,\ldots, p$ (at least one predictor variable has significance with response variable)

Reject H_0 if the G value is greater than $\chi^2_{(\alpha,v)}$ while $\chi^2_{(\alpha,v)}$ is the value that can be obtained from the chi-square table which has a significance level (α) and degrees of freedom (v).

c) Assumption of logistics regression

When the dependent variable is dichotomous, logistics regression is the best regression analysis to use. There are five key assumptions in logistics regression [10].

(1) Assumption of appropriate outcome structure. Binary logistic regression requires a binary dependent variable, whereas ordinal logistic regression requires an ordinal dependent variable

(2) Assumption of observation independence. Logistic regression necessarily requires that the observations be independent of each other

(3) Assumption of the absence of multicollinearity. Logistic regression requires that the independent variables have little or no multicollinearity. This indicates that the independent variables should not be overly correlated.

(4) Assumption of linearity of independent variables and Log odds. Although the dependent and independent variables are not required to be linearly related, the independent variables must be linearly related to the log odds.

(5) Assumption of a large sample size. Logistic regression frequently necessitates a large sample size.

2) Naïve bayes classifier

Naïve Bayes is a classification based on probability and statistical methods introduced by British scientist Thomas Bayes, predicts future opportunities based on past experiences, known as Bayes' Theorem. Naïve Bayes Classification Algorithm uses data with a target (class/label) in the form of categorical/nominal values. The naïve Bayes method is often called the HMAP (Hypothesis Maximum Apriori Probability) algorithm which is a simplification of the Bayes method. This method states the hypothesis of the calculation using probabilities depend on prior conditions [11].

$$P(H \mid X) = \frac{P(X \mid H).P(H)}{P(X)} \tag{11}$$

Note :

X	:Data with unknown class
H	:Data hypothesis X is a specific class
P (H|X)	:Probability hypothesis H based on condition X
P (H)	:Probability hypothesis H
P (X|H)	:Probability X based on the conditions in the hypothesis H
P (X)	:Probability X

To explain the naïve Bayes hypothesis, variable X in the above equation is changed to $F_1...F_n$ represents the qualities of the directions expected to perform the classification and variable H is changed to C represents the class. Therefore, Bayes hypothesis is changed as follows:

$$P(C \mid F_1...F_n) = \frac{P(C)P(F_1...F_n \mid C)}{P(F_1...F_n)} \tag{12}$$

The formula states that the probability of entering a sample with certain characteristics in class C (posterior) is the probability of the appearance of class C (before the sample's entry, often referred to as prior), multiplied by the probability of the sample's characteristics occurrence in class C (also known as likelihood), divided by the probability of the sample's characteristics occurrence globally (also called evidence).

$$posteorior = \frac{prior\ x\ likelihood}{evidence} \tag{13}$$

In a single sample, the value of evidence is always fixed for each class. Further elaboration of the Bayes formula is completed by explaining $C \mid F_1,...,F_n$ utilizing the multiplication rules as follows.

$$\begin{aligned} P(C \mid F_1...F_n) &\approx P(C)P(F_1...F_n \mid C) \\ &\approx P(C)P(F_1 \mid C)P(F_2,...,F_n \mid C,F_1) \\ &\approx P(C)P(F_1 \mid C)P(F_2 \mid C,F_1)P(F_3,...,F_n \mid C,F_1,F_2) \\ &\approx P(C)P(F_1 \mid C)P(F_2 \mid C,F_1)...P(F_n \mid C,F_1,...,F_{n-1}) \end{aligned} \tag{14}$$

It can be shows that the outcome of this elaboration is the more complicated conditional factors that affect the probability value, the more impossible it is to analyze one by one. This is where the assumption of very high (naïve) independence is utilized, that each piece of information is independent of each other.

$$P(P_i \mid F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = P(F_i) \tag{15}$$

For $i \neq j$,

$$P(F_i \mid C, F_j) = P(F_i \mid C)$$

From the above equation, it very well may be presumed that the assumption of naïve independence simplifies the probability conditions, that the calculation becomes possible. Furthermore, the interpretation of $P(C \mid F_1,...,F_n)$ can be rearranged to:

$$P(C \mid F_1,...,F_n) = P(C)P(F_1 \mid C)P(F_2 \mid C)... \ = P(C)\prod_{i=1}^{n} P(F_i \mid C) \tag{16}$$

The equation is a model of the naïve Bayes theorem, which will be used to classify the data.

3) Classification

The process of determining a model or function that describes or distinguishes a concept or data class with the objective of predicting the unknown class of an object is known as classification. In classifying data, there are two processes carried out, to be specific [10].

a) Training data

A training data set with known labels is used to create a model or function during the training process.

b) Testing data

To determine the accuracy of the model or function to be create on training process, the labels were predicted using data called a testing data set.

4) Confusion Matrix

Confusion matrix is a table that shows the amount of test data that is correctly and incorrectly classified. An illustration of a confusion matrix for classification is shown in Table I.

TABLE 1
CONFUSION MATRIX

Predicted	Actual	
	0	**1**
0	TN	FN
1	FP	TF

Notes:
a) True Positive (TP) : the number of data from class 1 that are correct and classified as class 1.
b) True Negative (TN) : the number of data from class 0 that are correctly classified as class 0.
c) False Positive (FP): the number of data from class 0 that are incorrectly classified as class 1.
d) False Negative (FN) : the number of data from class 1 that are incorrectly classified as class 0.

5) Accuracy, Precision, Recall and F1 Score

Accuracy is the most basic and widely used metric for evaluating a classifier's performance. However, accuracy isn't always a good metric, especially when the data is biased. The fundamental problem is that when the negative class is dominant, we can only achieve high accuracy if we predict negative the majority of the time. As a result, researchers must add another calculation to determine the validity of the prediction results, namely precision, recall, and F1 score. Precision is the degree of agreement between the information requested by the user and the response provided by the system. The level of success of the system in retrieving information is referred to as recall or sensitivity. Recall and precision values in a situation can have different weights. The measure that displays the reciprocity between Recall and Precision is the F1 score. Accuracy, precision, recall and F1 score are included evaluation metrics. Calculation of accuracy, precision, recall and F1 score are expressed in the equations:

$$\text{Accuracy} = \frac{TP+TN}{Total\ data} \times 100\% \quad (17)$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\% \quad (18)$$

$$\text{Recall} = \frac{TP}{TP+FN} \times 100\% \quad (19)$$

$$\text{F1 score} = \frac{2TP}{2TP+FP+FN} \times 100\% \quad (20)$$

E. Hyphotheses Testing

1) Partial test (Z test)

Z test is a partial test to see the effect of each independent variable having a significant or partial effect on the dependent variable.

$H_0 : \beta_j = 0$, j= 1,2,…,13; There is no significant influence of independent factor on the loan status.

$H_1 : \beta_j \neq 0$, j= 1,2,…,13; There is significant influence of independent factor on the loan status.

2) Overall test (G test)

$H_0 : \beta_1 = \beta_2 = ... = \beta_{13} = 0$, means there is no simultaneous significant effect of gender, married, dependent, education, self-employed, applicant income, co-applicant income, loan amount, loan amount term, credit history, property area: urban, property area: rural and property area: semiurban toward the loan status. H_1: at least one $\beta_j \neq 0$, j= 1,2,…,13 means there is a simultaneous significant effect of gender, married, dependent, education, self-employed, applicant income, co-applicant income, loan amount, loan amount term, credit history, property area: urban, property area: rural and property area: semiurban toward the loan status.

III. RESULTS AND DISCUSSION

A. Logistics Regression

1) Training and Testing Data

In the analysis utilizing the logistic regression method, the data is divided into two types, training data and testing data. Training data is previously data set where we know all the attributes, including the objective class property. While the testing data is the data used to test the classification rules acquired from the training data. For this study, the data is taken from GitHub "Credit Risk Train Data" where the data consists of 310 applicant data with eleven independent variables, which are; gender, married, dependent, education, self-

employed, applicant income, co-applicant income, loan amount, loan amount term, credit history, property area and dependent variable, which is loan status [12]. The proportion of training and testing data used is 75%: 25% where the training data is 232 data and the testing data is 78 data.

2) Overall and Partial Test

To determine the significant variables that have an effect on loan status, perform an overall test and partial test of all the independent variables. The overall testing is completed utilizing the likelihood ratio (G) test statistic, while the partial testing is done utilizing the Z test statistic.

First, researcher conducted an overall test with all independent variables. Based on calculation using Rstudio, the G value is 294.61. With a significance level of 5% obtained a value of $\chi^2_{(0.05,13)}$ is 22.362032. Because the value of G is greater than the value of $\chi^2_{(0.05,13)}$, then H_0 is rejected. It can be said that simultaneously all the independent variables are significant to the loan status. By utilizing Rstudio, it was found that the variables credit history, property area: urban and property area: rural are significant with the dependent variable because they have a p-value <0.05.

Next, researcher build model 1 with credit history, property area: urban and property area: rural variables, and then conducted an overall test with three variables. Based on calculation using Rstudio, the G value is 243.6. With a significance level of 5% obtained a value of $\chi^2_{(0.05,3)}$ is 7.814727903. Because the value of G is greater than the value of $\chi^2_{(0.05,3)}$, then H_0 is rejected. It can be said that simultaneously all the independent variables are significant to the loan status. However, when a partial test was conducted on the model 1, the researcher found that the variable property area: urban was no longer significant with loan status. The variable property area: urban has a *p*-value greater than 0.05. The partial test results show that there is no significant effect of property area: urban factor to loan status.

Then, researcher build model 2 with credit history, and property area: rural variables, and then do an overall test the two variables. Based on calculation using Rstudio, the G value is. 247.39 With a significance level of 5% obtained a value of $\chi^2_{(0.05,2)}$ is 5.99146. Because the value of G is greater than the value of $\chi^2_{(0.05,2)}$, then H_0 is rejected. It can be said that simultaneously all the independent variables are significant to the loan status. Furthermore, a partial test is carried out on the two variables and the results obtained, all variables are significant.

a) Credit history (β_{10})

H_0: There are no significance effect of credit history factor to loan status.

H_1: There are significance effect of credit history factor to loan status.

Significance test for credit history

1. Hypothesis
   H_0: $\beta_{10} = 0$
   H_1: $\beta_{10} \neq 0$
2. Level of significance = α 0.05
3. Z test
4. Reject to H_0 if *p*-value <0.05
5. *p*-value = $4.9*10^{-8}$
6. Decision = reject to H_0 because *p*-value <0.05
7. Conclusion = There are significance effect of credit history factor to loan status.

b) Property area : rural (β_{12})

H_0: There are no significance effect of property area: rural factor to loan status.

H_1: There are significance effect of property area: rural factor to loan status.

Significance test for β_{12} (Property area: rural)

1. Hypothesis
   H_0: $\beta_{12} = 0$

H_1: $\beta_{12} \neq 0$

2. Level of significance = 0.05
3. Z test
4. Reject to H0 if p-value <0.05
5. *p*-value = 0.01127
6. Decision = reject to H0 because *p*-value <0.05
7. Conclusion =There are significance effect of property area: rural factor to loan status.

3) Goodness of Fit

The model 2 is tested for goodness of fit with the Hosmer Lesmeshow test. Hosmer Lemeshow test values are utilized to see whether the data fit or not with the model.

TABLE 2
GOODNESS OF FIT

X-Squared	DF	Pf
0.017702	8	1

From Table II, hypothesis testing is carried out as follows:

1. Hypothesis

$H_0 = \hat{\pi}_i = y_i$ or the model fits (there is no significant difference between the observed results and the possible predictions of the model)

$H_1 = \hat{\pi}_i \neq y_i$ or the model does not fit (there is a significant difference between the observed results and the possible predictions of the model)

2. Level of significance = α 0.05
3. Reject to H_0 if *p*-value <0.05 or Reject $\chi^2_{\text{calculate}} > \chi^2_{(db;\ \alpha)}$
4. *p*-value = 1

   $\chi^2_{\text{calculate}} = 0.017702$, $\chi^2_{(8;\ 0.05)} = 15.51$
5. Decision = fail to reject to H_0 because *p*-value > 0.05,

   $\chi^2_{\text{calculate}} < \chi^2_{(8;0.05)}$
6. Conclusion = Model fits the data.

4) Logit Model

The logistics regression model for loan status is

$$In\left[\frac{\pi(x)}{1-\pi(x)}\right] = -1.2942 + 2.6898X_{10} - 0.8644X_{12}$$

where :

X_{10} = credit history

X_{12} = property area : rural

The interpretation of the logistics regression model can be expressed as follows:

1. The constant of -1.2942 indicates the size of the loan status variable when all independent variables were zero.
2. $\beta_j = 0,\ j = 1,2,...,9,11,13$, it is not significant, as *p*-value is more than 0.05. It shows that gender, married, dependent, self-employed, applicant income, co-applicant income, loan amount, loan amount term, education, property atea: urban, property area: semiurban have no significant influence on loan status.

3. β_{10} = 2.6898, it is the coefficient of the credit history variable. It shows that credit history has a positive influence on loan status.

4. β_{12} =-0.8644, it is the coefficient of property area: rural variable. It shows that property area: rural has a negative influence on loan status.

5) Prediction and Classification

Based on the predicted value (Table III), it can be shows that 10 records of rejected loan status are predicted to be correct as rejected loan status and 11 records are predicted to be incorrect as rejected loan status because these records are predicted to be approved loan statuses. Then 57 records in the approved loan status were predicted to be correct and no records were predicted to be incorrect as the approved loan status.

TABLE 3
CONFUSION MATRIX OF LOGISTICS REGRESSION

Predicted	Actual	
	0	**1**
0	10	0
1	11	57

B. Naïve Bayes

1) Training and Testing Data

In the analysis utilizing the naïve bayes method, the data is divided into two types, training data and testing data. Training data is previously data set where we know all the attributes, including the objective class property. While the testing data is the data used to test the classification rules acquired from the training data. For this study, the proportion of training and testing data used is 75%: 25% where the training data is 232 data and the testing data is 78 data.

2) Naïve Bayes Model

Basically, naïve Bayes method is a classification with probability and statistical methods, predicting future opportunities based on previous experiences.

3) Prediction and Classification

Based on the predicted value above, it can be shows that 10 records of rejected loan status are predicted to be correct as rejected loan status and 11 records are predicted to be incorrect as rejected loan status because these records are predicted to be approved loan statuses. Then 56 records in the approved loan status were predicted to be correct and 1 record were predicted to be incorrect as the approved loan status.

TABLE 4
CONFUSION MATRIX OF NAÏVE BAYES

Predicted	Actual	
	0	**1**
0	10	1
1	11	56

C. Comparison of Classification Results

Based on the classification result, it can be seen that the value of accuracy, precision, recall and F1 score of the logistic regression method has a higher value than naïve Bayes. According to the results of this research, logistic regression is the best method for classifying credit applications as approved or rejected.

IV. CONCLUSION

This work has discussed the prediction of loan status using logistics regression and a naïve Bayes classifier. It suggests that the percentage of approved loan status is 68% while for rejected loan status is 32%. According to logistic regression analysis, credit history, and property area: rural have a significant influence on loan status. It implies that the Logistics Regression model classification for the applicant's loan status on the test data has an accuracy of 85.9%, a precision of 83.82%, a recall of 100%, and an F1 score of 91.2%. The classification results

utilizing the naïve Bayes Classifier on the test data gave an accuracy of 84.62%, precision of 83.58%, recall of 98.25%, and F1 score of 90.32%.

Based on research results of loan status prediction, it concluded that the logistics regression method is better in classifying the loan status of applicants because the value of accuracy, precision, recall, and F1 score of the logistic regression method is greater than naïve Bayes classifier. For further research, other methods are needed to get good performance measures, such as neural networks and random forests.

REFERENCES

[1] Hermanto, "Faktor - faktor Kredit Macet pada PD. BPR BKK Ungaran Kabupaten Semarang," Universitas Semarang, 2006.
[2] A. Astusti, "Analisis Kredit Macet pada PT,BPR Restu Klaten Makmur," Universitas Sebelas Maret Surakarta, 2009.
[3] Y. B. Ernest, "Predicting Microfinance Credit Default," Kwame Nkrumah university, 2012.
[4] S. L. B. Ginting, J. Adler, Y. R. Ginting, and A. H. Kurniadi, "The Development of Bank Application for Debtors Selection by Using Naïve Bayes Classifier Technique," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 407, no. 1, 2018, doi: 10.1088/1757-899X/407/1/012177.
[5] T. Sunitha, M. Chandravallika, M. Ranganayak, G. Suma, and T. V. S. Jagadeesh, "Predicting the Loan Status using Logistic Regression and Binary Tree," no. Icicnis, pp. 708–715, 2020.
[6] S. Tabagari, "Credit Scoring by Logistics Rgression," Tartu University, 2015.
[7] U. Sekaran and R. Bougie, *Research Methods for Business: A Skill- Building Approach*, 6th editio. New York: Wiley, 2013.
[8] D. W. Hosmer and S. Lemeshow, *Applied Logistik Regression*, 2nd editio. Jakarta: John Wiley & Son, 2000.
[9] A. Agresti, *An Introduction to Categorical Data Analysis Second Edition*, Second. Wiley, 2007.
[10] E. Fajrila, "Perbandingan klasifikasi ketepatan waktu kelulusan mahasiswa menggunakan regresi logistik biner dan naïve bayes classifier," UNIVERSITAS ISLAM INDONESIA YOGYAKARTA, 2018.
[11] E. Prasetyo, *Data Mining Konsep dan Aplikasi Menggunakan MATLAB*, 1st ed. Yogyakarta: CV Andi Offset, 2012.
[12] A. Jana, "Credit Risk Train Data," *Github*, 2018. https://github.com/anup-jana/R-Machine-Learning/blob/master/R Scripts/Datasets/Credit_Risk_Train_data.csv.