Comparison of Data Mining Classification Methods for Predicting Credit Appropriation through Naïve Bayes and Decision Tree Methods

Rendi Irawan[1], Agustinus Eko Setiawan[2], and Kurnia Muludi[3]
[1]*IIB Darmajaya*
[2]*Aisyah University*
[3]*The University of Lampung*
[3]ren.irawan33@gmail.com[2]tynuskicenk@gmail.com, [3]kmuludi@yahoo.com,

Abstract: The problem statement of this study was seen on inaccurate assessment of the debtors' ability in paying off the loan of their businesses so that it often caused credit problems. Data Mining was used in assessing or predicting creditworthiness for a prospective debtor. The author attempted to compare the data mining classification to analyze the credit feasibility prediction through Naïve Bayes and Decision Tree methods. The data of the prospective debtors had been processed through the stages of data mining – Naïve Bayes and Decision Tree. The data were tested through k-folds cross-validation (k = 10). The result of this study was that the accuracy of the method of Decision Tree (J-48) was higher than that of the method of Naïve Bayes. The result of the comparison of the two algorithms was that the Decision Tree (J-48) algorithm had an accuracy of 95.24% and the Naïve Bayes algorithm had an accuracy of 79, 59%.

Keywords: Credit, Naïve Bayes, Decision Tree, K-Folds Validation

1. INTRODUCTION

The Bank is a business entity whose activities are to collect funds from the public in the form of savings and channel them to the public in credit or other forms. (UU Perbankan, 1988) Based on the law, all from of credit must be based on a loan agreement. Inaccurate initial assessment before becoming a customer will be a problem of bad credit due to the lack of optimal decision making in terms of predicting the feasibility of providing credit to costomers. The research problems were formulated as follows:

a. Was the data mining classification methods able to be used to predict credit feasibility at Bank XYZ?
b. Which the data mining classification methods – Naïve Bayes and Decision Tree – are the most accurate for predicting the credit feasibility to customers?

2. LITERATURE REVIEW

a. Meaning of Credit

According to Act No.10 of 1998 concerning the amendment of the Act No.7 of 1992 on the banking law, the chapter 1 of the general provisions of article 1, the meaning of credit was the provision of equivalent money or claims that were in accordance with the agreement or

borrowing agreement between the bank and other parties requiring the borrower to repay the debt after a certain period with the grant of interest.

b. Data Mining

Data mining was a software used to discover hidden patterns, trends, and rules contained in large bases and generate rules used to predict future behavior [1]

c. Classification

Classification in data mining was the method of learning data to predict the value of a group of attributes. The classification algorithm produced a set of rules used as an indicator to be able to predict the class of data to be predicted [2]. Classification was used in many fields, and theoretically the same classification algorithm as the human brain.

d. K-Folds *Cross-Validation*

In this research, the method used to test the classification pattern was by k-fold cross-validation method. K-fold cross validation data was divided into k section, D1, D2 ... Dk, and each D had the same amount of data. Testing with k = 5 or k = 10 was used to estimate the error rate because the training data on each fold was quite different from the original training data. Calculating the value of its accuracy was done using the equation:

$$\text{Accuracy}=\frac{number\ of\ correct\ classification}{Total\ test\ data}\ X\ 100\ \%\text{Algorithm}$$

e. Naive Bayes Algorithm

Naïve Bayes was the machine learning using probability calculations through the Bayesian approach. Naïve's word, the impressed derogatory, was derived from the assumption of the independence of the influence of the value of an attribute of the probability of a given class on the value of another attribute [4]. The use of the Bayes theorem in the Naïve Bayes algorithm was to combine the prior probabilities and conditional probabilities in a formula that was able to be used to calculate the probability of each possible classification [4].

Naïve Bayes formula was:

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

f. Decision Tree

The decision tree itself was a "divide and conquer" approach in studying the problem of an independent set of data illustrated in the tree chart [5]. Here was the equation of data in tuple D.

$$Info(D) = \sum_{i=1}^{n} -p_i\ log_2(p_i)$$

Where p_i was the probability of tuples in D being class C_i assuming | C_ (i, D) | / | D |. Info (D), also called the entropy of D, was the average of information needed for the identification of tuples in D.

g. Variable Selection

The selection of variables, also called attribute selection, was used on the dataset to find important patterns in data mining [5]. The choice of variables was used for dimensional

dimensions in the dataset. The selection of variables made data mining algorithms faster and more effective.

h. Testing Accuracy and Validation Method of Data Mining Classification
To test the model, this research used the Confusion Matrix method.

Confusion matrix
This method used the matrix table as in Table 1, if the data set consisted of only two classes, one class was considered positive and the other negative [4].

Table 1. *Model Confusion Matrix*

Correct classification	*Classified as* +	-
+	*True Positives*	*False Negatives*
-	*False Positives*	*True Negatives*

To calculate the used equation below [6].

Sensitivity = TP/P
Specificity = TN/P
Precision = TP/(TP+FP)
Accuracy = sensitivity $\frac{P}{P+N}$ + specificity $\frac{N}{P+N}$

i. Study Overview
The research model of Hendra Marcos and Indriana Hidayah on Implementation of Data Mining for Classification of Credit Clients Bank "X" using Classification Rule [8] showed some classification algorithm tested to the training data with algorithm C4.5 and it generated the highest accuracy value. Furthermore, The research model of Shary Armonitha Lusiana by using Algorithm C4.5 in analyzing Credit Feasiblity (Case Study in Employee's Co-Operation) (KP-RI) Lengayang Pesisir Selatan, Painan, Sumatera Barat[9] showed that the application of the C4.5 algorithm assisted the Cooperative in determining not only the credit members who were approved in credit submission but also the amount of credit to be disbursed. In addition, the research model of Jayanti and Noeryanti's using K-Nearest Neighbor Method and discriminant analysis for analyzing credit risk at savings and loan cooperative at “KopinkraSumberRejeki " [10] showed an accurate prediction of credit risk with the K-NN method of 84,33% at k = 7. Moreover, the research model of Siti Masripah about Comparison of Data Mining Classification Algorithm for Evaluation of Credit [11] showed the better accuracy level through the C4.5 algorithm by 88.90% accuracy value and Naïve Bayes by 80.00%

3. METHOD

a. Research Methods
The research method was the way of procedure to search, obtain, collect, and record data used in preparing a research report. The dataset was in the form of the training data by 147. Furthermore, 77 data outside of the dataset was used for data testing. The 147 data consisted of 12 attributes used to produce the required data.

b. Sample Selection Method
The sampling was the process of selecting some elements of the population. The probability sampling was a sampling technique providing the equal opportunity for each element or member of the population selected as sample [12]. In the credit analysis, the data were obtained from Bank XYZ within 2018, consisting of 12 attributes from which 11 attributes were predictors and 1 attribute was a label.

c. Method of collecting data
The researchers sought, studied, and explored various literatures through journals, books, or other references related to this research topic.

d. The technique of Analysis, Design, and Testing
1. Analysis Technique
The obtained data were divided into two sets e.g., training data (tested data) and evaluation data. The results of each method with the train data compared the results of the test by using k-fold cross-validation with k = 10 to obtain the result, precision, recall, ROC curve.

2. Classification Processes Design on Prototype
The process designed in the prototype system included:
- Import excel data
Import data was done to enter the data to be predicted into the prototype to be designed. The data format was xls.
- Preprocessing
The checking process was the process of checking on missing values, a difference of data format, and others.
- The prediction process
The prediction result was the prediction accuracy of customer loan determination.

3. Testing Technique
The testing technique for the method was done through k-folds cross-validation with k = 10. This method divided the trainer data randomly into 10 sections with almost the same amount in each group. Test results were obtained by calculating the average test statistic values on the whole iteration.

e. Research Steps
The steps used in this study used several steps existed in the CRISP-DM model (Cross Standard Industries Process for Data Mining). There were 6 stages as follow:

1. Research Understanding Phase
The collected data was seen on several variables: names, addresses, areas, ages, genders, last educations, status, number of family dependent, residence status, occupation, desired loan amount, loan objective, incomes/years, and credit funds.

2. Data Understanding Phase
The retrieved data was data only in 2016. The data that were obtained during this phase were the data of the customer (name, citizen id number, place/date of birth, residence address, last education, marital status, residence status, house occupancy, home phone, cell phone, NPWP, email address, mother's name) NPK / NIP, working hours, office address, office phone, department/section, monthly income), the data of the credit card reference, and the data of desired loan facilities (desired loan amount, loan purpose)

3. Data Processing Phase
   The data used as datasets were the date of names, genders, last educations, ages, marital status, number of dependents in the family, homeownership, occupation type, annual income, desired loan amount, loan purpose, and incomes.

4. Modeling Phase
   Comparing, selecting, and applying data mining classification modeling techniques were used through Naïve Bayes and Decision Tree. The data obtained from the data processing phase were used in this process.

5. Evaluation Phase
   The data mining classification was used to determine whether the model was able to achieve the objectives applied to the business understanding phase The accuracy or precision was expressed as errors in forecasting. The method used in this phase was k-fold cross-validation.

4. RESULT AND DISCUSSION

a. Data Analysis

Table 2. List of Data Sample Attributes

No	Atribut	Information
1	Gender	Female Male
2	Last education	Senior High School - S3
3	Age	35 - d 55
4	Status	Single/Married/Widow/Widower
5	Number of family dependents	0 - 4
6	Status of residence	Rent Owned Persons Owned Family House of Country
7	Profession	Civil Servants/PNS Private Employes Military/Police Teacher/Lecturer Other
8	Income/Year	65 - 11.250
9	Loan Amount	50 - 5.500
10	Loan Purpose	Personal Loan Housing Loan Car Loan
11	result	Approval Reject

To simplify the selection process, the results of the comparison of the evaluation classification of the classification was presented in the form of a confusion matrix table.

1. Confusion Matrix with Decision Tree
   The test was done with a confusion matrix consisting of accuracy done on the dataset by 147 data processed through the Decision Tree.

Table 3.Confusion Matrix with Decision Tree

	Dataset		
	N	**Y**	**Precision**
N	66	5	92,96%
Y	2	74	97,37%
Recall	97,06%	93,67%	

The accuracy value of the confusion matrix was as follows:

$$\text{accuracy} = \frac{(\text{TN} + \text{TP})}{(\text{TN} + \text{FN} + \text{FP} + \text{TP})}$$

$$= \frac{(66 + 74)}{(66 + 5 + 2 + 74)}$$

$$\text{accuracy} = 95{,}24\%$$

2. Confusion Matrix with Naïve Bayesalgorithm
   The result of the confusion matrix testing for datasets processed used the Naïve Bayes algorithm was seen below:

Tabel 5. Confusion Matrix with Naïve Bayes

	dataset		
	N	**Y**	**Precision**
N	68	3	95,77%
Y	27	49	64,47%
Recall	71,58%	94,23%	

The accuracy value of the confusion matrix was as follows:

$$\text{accuracy} = \frac{(\text{TN} + \text{TP})}{(\text{TN} + \text{FN} + \text{FP} + \text{TP})}$$

$$= \frac{(68 + 49)}{(68 + 3 + 27 + 49)}$$

$$\text{accuracy} = 79{,}59\%$$

3. Evaluation and Validation
   - Confusion Matrix with Decision Tree

The test was done with a confusion matrix consisting of accuracy done on 77 data testing processed by using Decision Tree. Confusion matrix testing for data testing processed using Decision Tree was seen in the table below.

Tabel 6.Confusion Matrix with Decision Tree

	Dataset		
	N	**Y**	**Precision**
N	37	2	94,87%
Y	4	34	89,47%
Recall	90,24%	94,44%	

The accuracy value of the confusion matrix was as follows:

$$\text{accuracy} = \frac{(\text{TN} + \text{TP})}{(\text{TN} + \text{FN} + \text{FP} + \text{TP})}$$

$$= \frac{(37 + 34)}{(37 + 2 + 4 + 34)}$$

$$\mathbf{accuracy} = \mathbf{92,21}\%$$

- Confusion Matrix algorithm Naïve Bayes
  There were 77 testing data through the confusion matrix testing for datasets processed through the Naive Bayes algorithm as they were seen below.

Table 8.Confusion Matrix with Naïve Bayes

	Dataset		
	N	**Y**	**Precision**
N	36	3	92,31%
Y	11	27	71,05%
Recall	76,60%	90,00%	

The accuracy value of the confusion matrix was as follows:

$$\text{accuracy} = \frac{(\text{TN} + \text{TP})}{(\text{TN} + \text{FN} + \text{FP} + \text{TP})}$$

$$= \frac{(36 + 27)}{(36 + 3 + 11 + 27)}$$

$$\mathbf{accuracy} = \mathbf{81,82}\%$$

4. Comparative Results
   The Decision Tree and Naïve Bayes algorithms for predicting the credit feasibility to customers showed inaccurate result. The customer dataset was regarded as the tested data.

Decision Tree algorithm had the highest accuracy of 93.72%; while, the Naïve Bayes algorithm had an accuracy of 80.71%. It was seen on table 10.

Tabel 9. Comparison of Accession Levels of Decision Tree and Naïve Bayes

Data Mining Methods	Accuracy		Comparison
	Training	***Testing***	**Value accuracy**
Decision Tree	95,24%	92,21%	93,72%
Naïve Bayes	79,59%	81,82%	80,71%

5. CONCLUSION

The Decision Tree method had a good accuracy of 95.24% and the Naïve Bayes method had an accuracy of 79.59%. Although the Decision Tree algorithm had a high accuracy, the further research was needed such as 1) combining more methods in data analysis and problem-solving so that a system was more effective and efficient in processing or presenting information; 2) managing the time research maximally; 3) The roles of respondents was very important in supporting this research, especially the respondents were directly involved in research; .4) this research was able to be developed with the other classification algorithms contained in data mining e.g., Neural Network algorithm, K-NN, K-Means, or SVM (Support Vector Machine).

REFERENCES

[1] A. Kadir, Pengenalan Sistem Informasi Edisi Revisi, no. Penerbit Andi. Yogyakarta: Andi, 2014.
[2] C. Vercellis, Business Intelligence: Data Mining and Optimization for Decision Making (Google eBook), no. 2004. 2011.
[3] F. Gorunescu, Data Mining Concepts, Models, and Techniques. Springer-Verlag, 2011.
[4] M. Bramer, Principles of Data Mining. Springer-Verlag London, 2013.
[5] I. H. Witten, E. Frank, and M. a. Hall, Data Mining Practical Machine Learning Tools and Techniques Third Edition, vol. 277, no. Tentang Data Mining. 2011.
[6] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. 2012.
[7] H. Marcos, I. Hidayah, J. Teknik, E. Dan, T. Informatika, and U. G. Mada, "Implementasi Data Mining Untuk Klasifikasi Nasabah Kredit Bank ' X ' Menggunakan Classification Rule," pp. 1–7, 2014.
[8] S. A. Lusinia, S. Kom, M. Kom, and F. I. Komputer, "Algoritma C4.5 dalam menganalisa kelayakan kredit(studi kasus di koperasi pegawai Republik Indonesia(KP-RI))Lengayang Pesisir Selatan, Painan, Sumatera Barat," vol. 1, no. 2, pp. 6–10, 2014.
[9] J. R. Dwi and Noeryanti, "Aplikasi Metode K-Nearest Neighbor dan Analisis diskriminan untuk analisa resiko kredit pada koperasi simpan pinjam di Kopinkra Sumber Rejeki," pp. 275–284, 2014.
[10] S. Masripah, "Komparasi Algoritma Klasifikasi Data Mining untuk Evaluasi Pemberian Kredit," vol. 3, no. 1, pp. 187–193, 2016.
[11] S. Guritno, Sudaryono, and U. Rahardja, Theory and Application of IT Research-Metodologi dan Penelitian Teknologi Informasi. Yogyakarta: Andi, 2011.