RESEARCH ARTICLE OPEN ACCESS

Data Mining Implementation Using Naïve Bayes Method for Prediction of Customers in Customer's Credit Payment in CS Finance

Herfandi[1], Royan Habibie Sukarna[2]
1(Computing Engineering, Budi Luhur University, and Jakarta
Email:herfandi3@gmail.com)
2(Computing Engineering, Budi Luhur University, and Jakarta
Email:rhsukarna@gmail.com)

Abstract:

Credit collectibility is a way to find out the quality of credit so that Central Santosa Finance can anticipate credit risk early because credit risk can affect the continuity of business enterprises. The quality is based on 3 main standards, one of which is the ability of customers to pay the principal and interest at the time agreed upon together. Whereas, to accurately predict the ability of prospective customers to repay loans based on the characteristics of manually owned data sets is very difficult. Data Mining with the Naïve Bayes method was chosen to find patterns in analyzing and predicting the ability to repay prospective borrowers. The test will be carried out by comparing the initial dataset and dataset characteristics using the attribute selector Gain Ratio Attribute algorithm with the help of WEKA tools. The results showed that there was a difference in the results of accuracy, and that the ROC or AUC (Area Under Cover) curve was greater in the dataset characteristics using the selector attribute by using the Gain Ratio attribute, although not too significant. And the results of this study produced an accuracy rate of 78.56% with a precision of 93.46% and a recall of 72.78%. The method used is included in Good Classification and will later become a reference for the management of Central Santosa Finance, to address the problems that may arise in decreasing the quality of credit (eg decreasing the company's ratio with customers).

***Keywords** — Data Mining, Credit Analysis, Customer Prediction, Naïve Bayes.*

I. INTRODUCTION

In the current era, Data Mining has become very popular in the banking sector. This is due to the development of the information age that has entered the business world, except that its utilization is still not optimal [1].

There are many classification algorithms that can be used such as Decision Tree Algorithm [2], Neural Network [3], Support Vector Machine [4] Naive Bayes [5], and K-Nearest Neighbo [6].

At present the transfortation tool is a necessity of the community in carrying out daily activities. This can be seen from the data issued by the Central Statistics Agency (BPS), where the development of the number of vehicles continues to increase every year. The biggest increase was in the type of passenger car vehicles (8.15%) and motorcycles (6.38%). The data was taken from the Traffic Corps of the Republic of Indonesia Police (Korlantas Polri). This is because the aspects of security and security are the reason for people to switch to private transportation, namely motorbikes and cars. To meet the need for Transfortasi Pribadi tools, multi finance companies provide a credit payment system to make it easier for people to have personal transportation equipment.

One of the challenges faced by the Central Santosa Finance company is to improve poor credit analysis, which can have an impact on the high credit risk experienced at Central Santosa Finance. The thing that becomes crucial is in determining the strategy and planning so that the quality of Central Santosa Finance's credit can be improved. Customer failure or inability to return the loan amount and interest in accordance with a predetermined period of time (non-performing loans) can be done by evaluation to continue to improve credit quality at Central Santosa Finance both in terms of management and marketing risk at Central Santosa Finance. Credit collectibility is a way to find out the quality of credit in companies so that Central Santosa Finance can anticipate credit risk early because credit risk can affect the continuity of business enterprises. The quality is based on 3 main standards, one of which is the ability of customers to pay the principal and interest at the time agreed upon together [5].

Low credit collectibility or collectibility status that is more popular as bad credit can affect monetary economic conditions at the deteriorating Central Santosa Finance and have a trickle down effect on the economy as a whole, which impacts on the company's growth and income going forward. For this reason, companies need to learn from past data to find customer payment patterns. So this study aims to predict the ability of prospective customers to support decision making in an effort to increase credit collectibility so that they get good credit quality with data mining implementation [7].

In an effort to increase the percentage of prospective customers' payment ability is to analyze the patterns in the History database of customer payments, to predict the level of ability of prospective customer payments that are difficult to analyze manually [8].

II. Related Work

In order to make it easier to understand the material related to the writing of scientific articles, the writer presents it simply as follows.

A. Data Mining

Data mining is defined as a set of techniques that are used automatically to explore thoroughly and bring to the surface complex relationships on very large data sets. The data set referred to here is a data set in the form of tabulation, as is widely implemented in relational database management technology. However, data mining techniques can also be applied to other data representations, such as spatial data domains, text based, and multimedia (images) [9].

Data Mining can also be interpreted as the process of extracting information from large data sets through the use of algorithms and techniques taken from the fields of statistics and Database Management Systems [3]:

B. Naïve Bayes

The Naive Bayes algorithm (NB) is a simple method in classification based on probability theory proposed by British scientist Thomas Bayes. Naive because it simplifies problems that depend on two important assumptions [10].

The advantage of Naïve Bayes calcification is that this algorithm only requires a small amount of training data needed for the classification process [11]. The Naïve Bayes classification has been proven to be applied in real and complex situations.

Naïve bayes can be defined as follows:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \qquad (1)$$

Note:

- P (c | x) is posterior probability of the class (target) against predictor (attribute).
- P (x | c) is likelihood which is a predictor of the class.
- P (c) is the prior probability of the class.
- P (x) is a prior probability from a predictor.

C. Gain Ratio Attributes

In the task the pattern classification feature plays a very important role [12]. Therefore, the selection of appropriate features is needed because most raw data may be excessive or irrelevant to pattern recognition. In some cases, classifiers cannot function properly because of the many redundant features [13]. Different features play different roles in grouping datasets. Unwanted features will produce error information during classification which will reduce the precision of classification. Most selection of traditional features can eliminate this interference to improve classification performance.

Gain Ratio is a modification of information gain to reduce the bias attribute that has many branches. Gain ratio has properties:

a. Great value if the data spread evenly.
b. Small value if all data entered in one branch.

Gain ratio has a formula:

$$GainRatio\ (A) = \frac{Gain(A)}{SplitInfo(A)} \qquad (2)$$

Where the info split formula as in the formula above with m states the number of splits. The type of split chosen is split which has the largest Gain Ratio value.

You could say the value of split information represents the potential information generated by dividing the training data set D to v partition, according to the result of v in attribute A.

$$SplitInfo\ \ _A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} x log_2 \left(\frac{|D_j|}{D}\right) \qquad (3)$$

III. RESEARCH METHOD

A. *CRISP-DM*

The steps used in this study the author adopts the CRISP-DM (Cross Standard Industries Process for Data Mining) model, where there are 6 stages [14], namely:

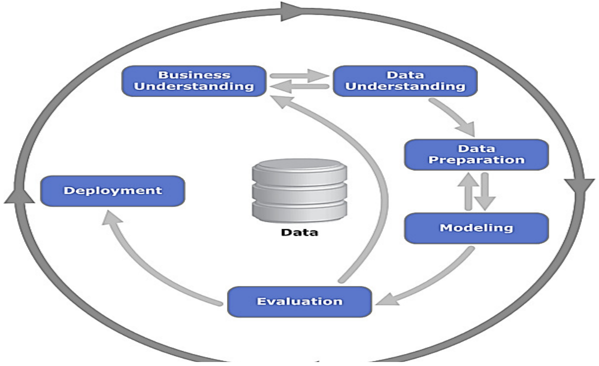


Fig1. CRISP-DM Cycle

1. *Business understanding*
2. *Data understanding*
3. *Data preparation*
4. *Modelling*
5. *Evaluation*
6. *Deployment*

Meanwhile, the conceptual framework in research can be described as follows:

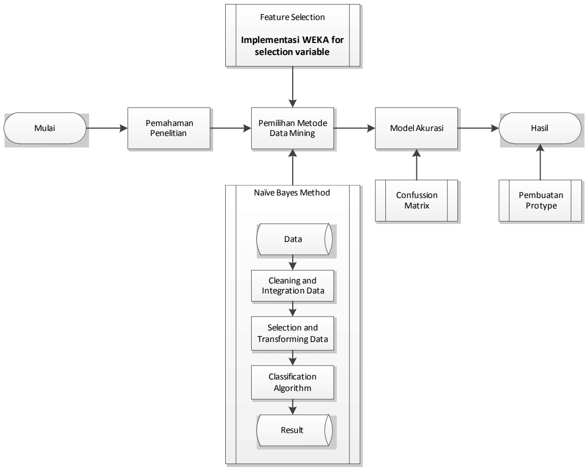


Fig2. Conceptual framework

The mindset illustrated above can be explained as follows:

1. The process of understanding this research is a phase in collecting data that supports research. Data taken is data from Customer's payment history
2. The classification process, in this phase the algorithms of data mining classification are selected and applied. Then it is compared with the algoirtma attribute selector using help from WEKA tools.
3. The validation process, this phase serves to measure the level of accuracy of a model that has been made. Receiver Operating Characteristic (ROC) curves will be used to measure AUC (Area Under Curve) originating from the original dataset attribute with the dataset attribute derived from the attribute selector algorithm by WEKA tools.
4. Making a prototype, this phase builds a prototype that will be used to predict the potential payment ability of the prospective customer

B. *Design Prototype System*

Next is the flow of Activity Diagrams, Use Cases, Class Diagrams and Flow Charts that are implemented into prediction prototypes.

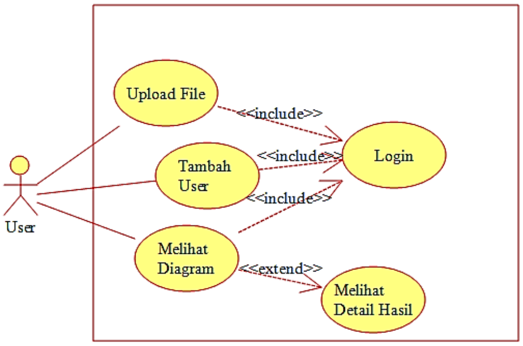


Fig3. Usecase system

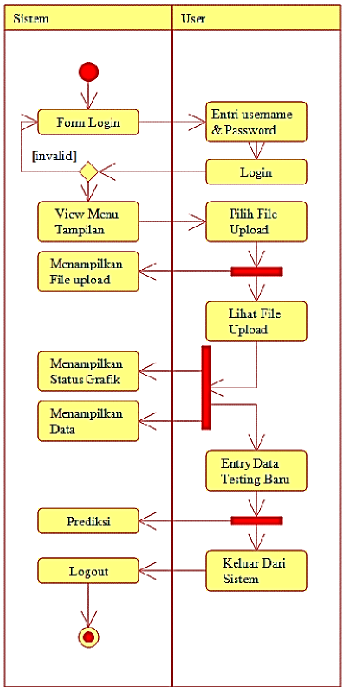


Fig4. Activity Diagram System

The user opens the application by entering a login on the system and will display the display menu and displays the option to upload files selected by the user at the same time the system can display data from the uploaded file selected by the user, then the system will import upload files and provide a summary of data upload. The user will upload a selection of testing data files and display a prediction graph from the selected data. After completion the user will log out of the application.

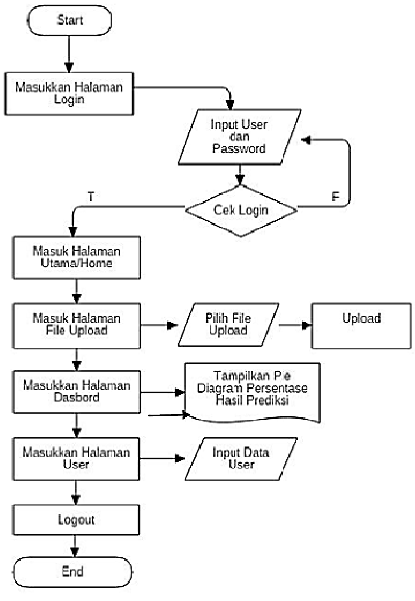


Fig5. *Flowchart Prototype System*

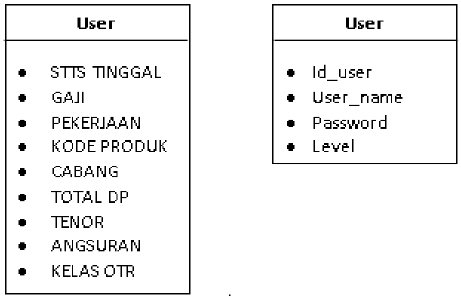


Fig 6. Class Diagram

IV. RESULT AND DISCUSSION

A. Implementation System

To apply the method used, 2 (two) datasets which are the initial dataset and the dataset obtained from the selection process use the Attribute Gain Ratio algorithm. The data used is generally divided into 2 (two) types of data, namely:

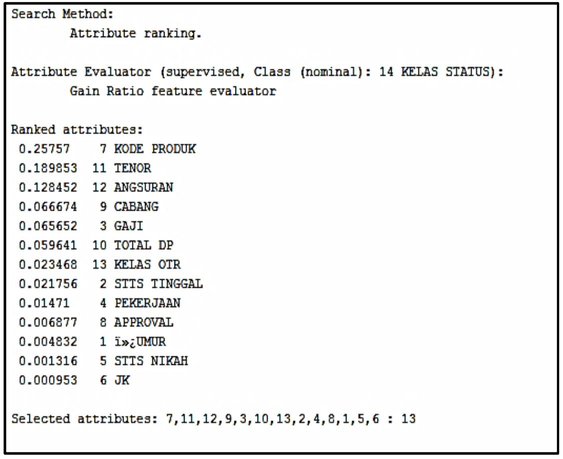
1. Contract data
2. Application data

The initial dataset has 14 (fourteen) attributes which consist of 13 (thirteen) predictor attributes and 1 result attribute. Each attribute and value can be seen in the table below:

Table I.Initial dataset attribute

No	Atribut	Nilai	Nilai Baru
1	UMUR	17-25	Remaja Akhir
		26-35	Dewasa Awal
		36-45	Dewasa Akhir
		46-55	Lansia Awal
		56-65	Lansia Akhir
2	STATUS TINGGAL	H01	PRIBADI
		H02	TUMPANG
3	GAJI	500000 - 2300000	SANGAT RENDAH
		2500000 - 3500000	RENDAH
		4000000 - 6000000	MENENGAH
		7000000 - 10000000	TINGGI
		> 10000000	SANGAT TINGGI
4	PEKERJAAN	Belum/Tidak Bekerja	Belum/Tidak Bekerja
		Buruh	Buruh
		Karyawan Swasta	Karyawan Swasta
		Lainnya	Lainnya
		Mengurus Rumah Tangga	Mengurus Rumah Tangga
		Pegawai Negeri Sipil	Pegawai Negeri Sipil
		Pelajar/Mahasiswa	Pelajar/Mahasiswa
		Petani/Pekebun	Petani/Pekebun
		Wiraswasta	Wiraswasta
5	STATUS NIKAH	S	Single
		M	Menikah
		D	Duda/Janda
6	JK	M	Pria
		F	Wanita
7	KODE PRODUK	KSM	Rek
		KPM	Non Rek
8	APPROVAL	White	CAH
		Grey	BM
9	CABANG	32001	Serang
		32002	Tasik
		32003	Bandung
10	TOTAL DP	1,350,000 - 2,000,000	Sangat Rendah
		2,500,000 - 4,500,000	Rendah
		5,000,000 - 8,000,000	Menengah
		9,000,000 - 12,000,000	Tinggi
		>13,000,000	Sangat Tinggi
11	TENOR	9 -12	Sangant Singkat
		13 -18	Singkat
		19 - 24	Sedang
		25 - 30	Lama
		>31	Sangat Lama
12	ANGSURAN	500,000 - 1,500,000	Rendah
		1,700,000 - 2,500,000	Menengah
		3,000,000 - 5,000,000	Tinggi
		9,000,000 - 12,000,000	Tinggi
13	KELAS OTR	13,000,000 - 18,000,000	Bebek1
		19,000,000 - 24,000,000	Bebek2
		25,000,000 - 30,000,000	Matic150
		30,000,000 - 40,000,000	Sport
		> 40,000,000	Premium
14	KELAS STATUS	PT & CL	CLEAR
		WO	BAD DEBT

In testing sample data containing 3,085 data with 13 predictor attributes and 1 result attribute, using the Selector Gain Ratio Attribute attribute algorithm through WEKA tools, 9 attributes with the highest ranking were taken, namely:



```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 14 KELAS STATUS):
        Gain Ratio feature evaluator

Ranked attributes:
 0.25757    7 KODE PRODUK
 0.189853  11 TENOR
 0.128452  12 ANGSURAN
 0.066674   9 CABANG
 0.065652   3 GAJI
 0.059641  10 TOTAL DP
 0.023468  13 KELAS OTR
 0.021756   2 STTS TINGGAL
 0.01471    4 PEKERJAAN
 0.006877   8 APPROVAL
 0.004832   1 ï»¿UMUR
 0.001316   5 STTS NIKAH
 0.000953   6 JK

Selected attributes: 7,11,12,9,3,10,13,2,4,8,1,5,6 : 13
```

Fig7. Output Selector attribute from WEKA tools

The selected attributes are presented in the following table form:

Table 2. Result Attributes from WEKA Tools

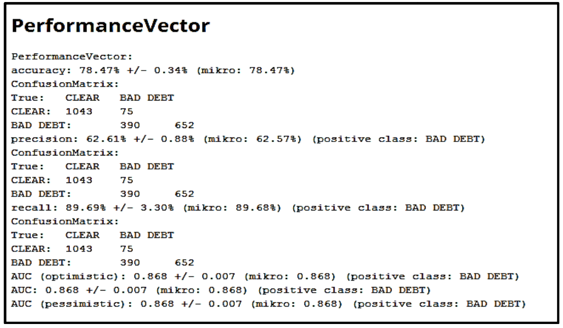
No	Ranking	Atribut
1	0.25757	KODE PRODUK
2	0.189853	TENOR
3	0.128452	ANGSURAN
4	0.066674	CABANG
5	0.065652	GAJI
6	0.059641	TOTAL DP
7	0.023468	KELAS OTR
8	0.021756	STTS TINGGAL
9	0.01471	PEKERJAAN

Testing dataset models using k-fold cross validation. The results of the testing methods that have been done are Naïve Bayes algorithm, testing the level of accuracy by using the help of Rapid Miner software to find confusion matrix and ROC / AUC (Area Under Cover) curves. After that, the results of attribute selection were tested by using the Gain Ratio Attribute Evaluator algorithm from WEKA tools.

Table 3. Confusion Matrix of Early dataset

	true CLEAR	true BAD DEBT	class precision
pred. CLEAR	1043	75	93.29%
pred. BAD DEBT	390	652	62.57%
class recall	72.78%	89.68%	

From the model testing that has been carried out using the rapid miner software, we obtain the following performance vectors and ROC curves.



```
PerformanceVector

PerformanceVector:
accuracy: 78.47% +/- 0.34% (mikro: 78.47%)
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  1043    75
BAD DEBT:       390     652
precision: 62.61% +/- 0.88% (mikro: 62.57%) (positive class: BAD DEBT)
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  1043    75
BAD DEBT:       390     652
recall: 89.69% +/- 3.30% (mikro: 89.68%) (positive class: BAD DEBT)
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  1043    75
BAD DEBT:       390     652
AUC (optimistic): 0.868 +/- 0.007 (mikro: 0.868) (positive class: BAD DEBT)
AUC: 0.868 +/- 0.007 (mikro: 0.868) (positive class: BAD DEBT)
AUC (pessimistic): 0.868 +/- 0.007 (mikro: 0.868) (positive class: BAD DEBT)
```

Fig8. *Performance Vector* First Dataset

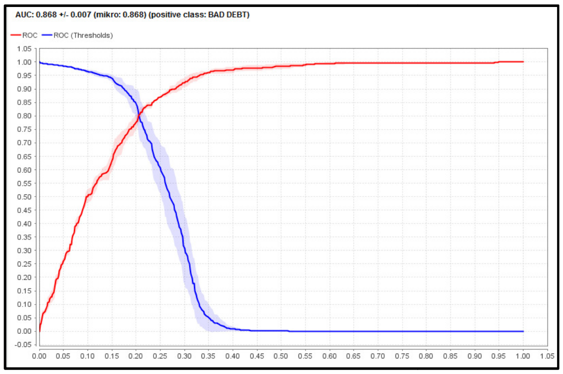


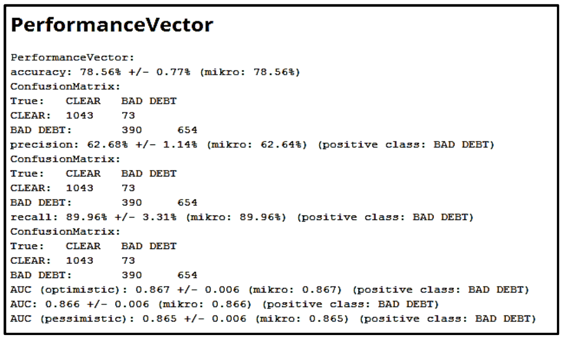
Fig9. ROC Curve First Dataset

Confusion matrix testing for dataset attributes from WEKA tools that are processed using the Naive Bayes algorithm with 70% training data and 30% testing data can be seen in the table below:

Table 4. Confusion Matrix dataset Feature Selection

	true CLEAR	true BAD DEBT	class precision
pred. CLEAR	1043	73	93.46%
pred. BAD DEBT	390	654	62.64%
class recall	72.78%	89.96%	

From the model testing that has been done using rapid miner software, the verctor and ROC curves are obtained as follows:



```
PerformanceVector

PerformanceVector:
accuracy: 78.56% +/- 0.77% (mikro: 78.56%)
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  1043    73
BAD DEBT:       390     654
precision: 62.68% +/- 1.14% (mikro: 62.64%) (positive class: BAD DEBT)
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  1043    73
BAD DEBT:       390     654
recall: 89.96% +/- 3.31% (mikro: 89.96%) (positive class: BAD DEBT)
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  1043    73
BAD DEBT:       390     654
AUC (optimistic): 0.867 +/- 0.006 (mikro: 0.867) (positive class: BAD DEBT)
AUC: 0.866 +/- 0.006 (mikro: 0.866) (positive class: BAD DEBT)
AUC (pessimistic): 0.865 +/- 0.006 (mikro: 0.865) (positive class: BAD DEBT)
```

Fig 10. *Performance Vector* Dataset *Feature Selection*

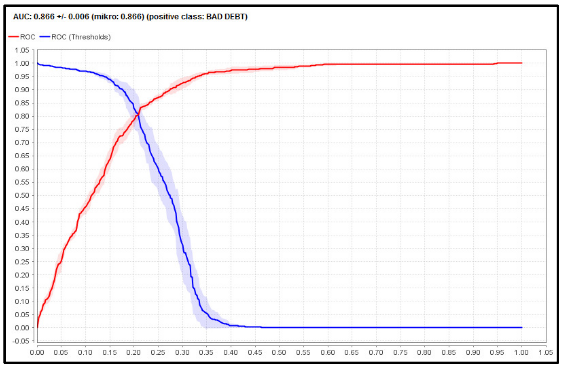


Fig 11. ROC Curve Dataset *Feature Selection*

The dataset used has 3,085 instances divided into 70% training data and 30% testing data with the initial model 14 attributes get precision values of 93.29%, recall 72.78%, accuracy 78.47%, and ROC values of 0.868. While models with datasets that use attributes resulting from the feature selection Gain Ratio get 10 attributes with 1 predictor attribute. The value of precision achieved is 93.46%, recall 72.78%, accuracy 78.56%, and ROC value of 0.866. So that the dataset model is classified as good classification.

V. CONCLUSIONS

In this paper, the proposed method using a selection feature with the Attribute Gain Ratio algorithm gets good results and has increased accuracy. In further research it is possible to implement a feature selection into other classification algorithms, such as the Neural Network algorithm, K-Means or SVM (Support Vector Machine), etc.

REFERENCES

1. Melissa, Ira, and Raymond S Oetama. “Analisis Data Pembayaran Kredit Nasabah Bank Menggunakan Metode Data Mining.” ULTIMA InfoSys IV, no. 1 (2013): 18–27.
2. Ying, Li. “Research on Bank Credit Default Prediction Based on Data Mining Algorithm.” The International Journal of Social Sciences and Humanities Invention 5, no. 6 (2018): 4820–23. https://doi.org/10.18535/ijsshi/v5i6.09.
3. V. S. Moertini, “Data Mining Sebagai Solusi Bisnis,” *Integral*, vol. 7, no. 1, pp. 44–56, 2002.
4. E. Osmanbegović and M. Suljić, “Data mining approach for predicting student performance,” *J. Econ. Bus.*, vol. X, no. 1, pp. 3–12, 2012.
5. B. K. Bhardwaj and S. Pal, “Data Mining : A prediction for performance improvement using classification,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 9, no. 4, pp. 136–140, 2011.
6. H. Wang, “An Empirical Study on the Stability of Feature Selection for Imbalanced Software Engineering Data,” *Int. J. Adv. Comput. Res.*, vol. 2, no. 3, pp. 1–5, 2012.
7. A. L. Blum and P. Langley, “Selection of relevant features and examples in machine learning,” *Artif. Intell.*, vol. 97, no. 1–2, pp. 245–271, 1997.
8. P. Chapman *et al.*, “Crisp-Dm 1.0,” *Cris. Consort.*, p. 76, 2000.