JOURNAL OF DATA SCIENCE | Vol.2023:11
eISSN:2805-5160

Utilizing Classifier Algorithms to Analyze Lending Activity at The Sumsel Babel Bank's Pagar Alam Branch

Muhammad Ichsan[1], Tri Basuki Kurniawan[1*]

[1] Magister of Information Technology, University of Bina Darma, Palembang, Indonesia

***Email:** tribasukikurniawan@binadarma.ac.id

Abstract

Everyday living requires accurate information, and knowledge will play a significant role in civilization's present and future growth. It is not sufficient to rely solely on operational data when using existing data in information systems to assist decision-making activities; data analysis is necessary to fully realize the potential of the information already available. The government and banking currently work together to distribute foreign exchange credit, which helps MSMEs who want to expand their businesses by providing additional capital. Bad credit cannot be separated from lousy credit when granting bank credit, one of the issues that banks nowadays frequently face. Additionally, a credit analyst must conduct manual research and analysis to evaluate the business circumstances of potential debtors that are anticipated to affect their capacity to perform their obligations to the Bank while reviewing the distribution of foreign credit to MSMEs. In this research, a classifier algorithm was applied to create a prediction model to predict the customer before the lending application was used and process to pass the lending process in Bank Sumsel, branch Pagar Alam. The experiment was conducted, and based on our data and model, the result obtained 85.54% accuracy based on the Random Forest classifier model. The result shows the algorithm is entirely reasonable in predicting customer data.

Keywords

Classifier Algorithms, Classification Algorithms, Lending Activity

Introduction

Everyday living requires accurate information, and knowledge will play a significant role in civilization's present and future growth. It is not sufficient to rely solely on operational data when using existing data in information systems to assist decision-making activities; data analysis is necessary to fully realize the potential of the information already available. Decision-makers attempt to use the data warehouse they already possess in their decision-making, which has sparked the development of a new scientific field known as data mining to address the challenge of extracting significant and intriguing information or patterns from enormous amounts of data. Data

Submission: 9 September 2023; **Acceptance:** 4 October 2023

mining techniques are anticipated to reveal knowledge kept secret in the data warehouse and turn it into useful information.

Banking is a financial organization whose job is to gather money from the general population and disperse it as loans. Every service that banks offer is convenient, such as the service of supplying MSMEs with foreign exchange loans. The government and banking currently work together to distribute foreign exchange credit, which helps MSMEs who want to expand their businesses by providing additional capital. Because foreign exchange credit is a government-issued credit product, the government in Permenko Regulation No.2 of 2021 provides a policy for banks to offer cheap credit interest, specifically 6% per year, free bank admin fees, and insurance guarantee fees. Banks distribute the more foreign exchange credit, the better it is for the government because it increases economic growth, particularly in the small and medium businesses and the employment sector.

In contrast, the benefit of foreign exchange credit is a profit for the banking companies—the business, as the higher the loan disbursement, the greater the gain for the business. But by offering foreign exchange credit, the Bank not only gives a debtor a loan but also takes on most of the company's risk.

One of the problems that banks nowadays regularly have is that poor credit cannot be distinguished from bad credit when issuing bank loans. The Bank must determine which potential borrowers are qualified because there are so many people who potentially use loans. When deciding whether potential debtors are eligible, the Bank applies the prudence principle.

Additionally, a credit analyst must conduct manual research and analysis to evaluate the business circumstances of potential debtors that are anticipated to affect their capacity to perform their obligations to the Bank while reviewing the distribution of foreign credit to MSMEs. To solve this problem, in this research, classification, one of the methods in data mining, is proposed. Many other researchers successfully solved similar issues based on the classification approach.

Dhea et al. (2022) used Naïve Bayes, Random Forest, and SVM to solve the MTI student thesis documents. Students who plan to conduct research first choose the theme or topic of the study they will perform and then look for references to support the chosen theme or topic. Classification of thesis documents will result in a classification of thesis themes or issues, whose findings can assist students in finding precise thesis references following the subject, which is anticipated to assist students in finding their connections more quickly. The experiments show the result from Random Forest gives the best performance. Another researcher, Misinem et al. (2022), used Linear Regression, Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine (SVM) to deal with users' satisfaction with mobile banking apps.

Haryati, Kurniawan, and Negara (2023) in their research used Naïve Bayes, Random Forest and SVM approaches to solve the achievement scholarship award in high school. The results show Random Forest gives the best performance.

Methodology

This paper's process to solve the problem includes data mining or machine learning research. Data mining is the process of finding knowledge in databases known as data mining. Data mining involves identifying and extracting relevant information and related knowledge from massive databases using mathematical, statistical, artificial intelligence, and machine learning approaches (Turban. E., et al., 2005).

The process of looking for hidden patterns in the form of previously undiscovered knowledge from a set of data—which can be in databases, data warehouses, or other information storage media—is known as data mining. The following are crucial factors in data mining:

1. The automated method of mining existing data.
2. The data that needs to be processed is quite vast.
3. Data mining aims to find links or patterns that could yield helpful indicators (Kusrini and Emha, 2009).

With the aid of specialized tools, data mining procedures designed based on the analytical model are carried out. Data mining is analyzing data to uncover hidden information in vast amounts of data kept for use in corporate operations.

The following five kinds of data mining tasks can be separated based on their functionality (Mints, Aleksey, 2017):

1. Classification
   The goal of classification is to group categorical variables. For instance, there are three categories for categorizing the collectability level of foreign exchange credit: Current, Under Consideration, and Rejected.
2. Clustering (Clustering)
   It is gathering information or records, careful observation, and creating groups of items with shared characteristics. Data within the group's similarity threshold will be added; if it does not, it will be split into a different group.
3. Estimation
   The best approach for resolving issues involving output estimation typically uses attributes and classes represented as numeric data.
4. Associative
   In data mining, associations are attributing discoveries in the data that appear simultaneously. In association, methods like FP-Growth, A Priori, Correlation Coefficient, and Chi-Square can be employed.
5. Predictive
   In data mining, predictive analysis involves estimating the value of a specific property based on the importance of other qualities. There are two variables: the independent and dependent variables, which represent the traits utilized to generate predictions and the predicted attributes, respectively.

The classification technique involves studying data to develop rules to classify or recognize fresh data that has never been studied, a crucial component of data mining (Suyanto, 2017). There are two basic jobs in classification, namely:

1. Creation of the prototype model to be kept in memory.

2. Applying recognition, classification, and prediction to another data object using the model to determine which class it belongs to in the model it has stored.

According to Sugiono (2014), the research framework is a conceptual model of how the theory relates to various factors identified as essential problems. Based on the identification of the situation that has been described in the previous section, the final goal that will be achieved in this research is to analyze the bad credit at the time of the submission of credit KUR (people's business credit) that will be submitted by UMKM (micro small and medium enterprises) customers using classifier algorithms method.

The European Commission designed the Cross Industry Standard Process for Data Mining (CRISP-DM) data mining process model (data mining framework) in 1996 as a uniform data mining procedure that may be used in various industrial and banking sectors. Data processing through CRISP-DM must go through several steps to generate data output results that can boost business value and prevent losses to industrial and banking firms. According to (Mustafa & Sarah, 2021), there are 6 phases of CRISP-DM, as shown in Figure 1.

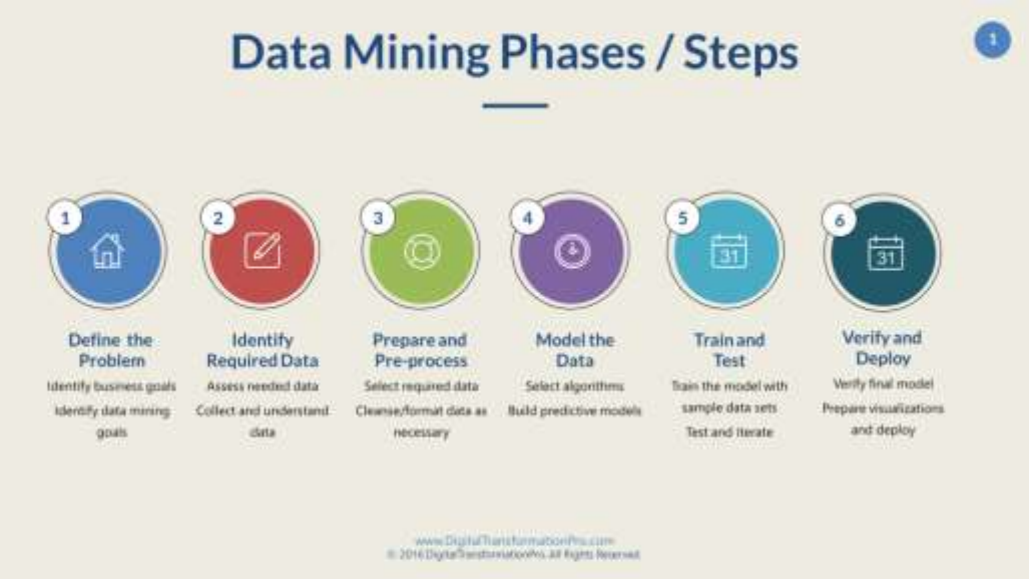

Data Mining Phases / Steps
1
1
Define the Problem
Identify business goals
Identify data mining goals
2
Identify Required Data
Assess needed data
Collect and understand data
3
Prepare and Pre-process
Select required data
Cleanse/format data as necessary
4
Model the Data
Select algorithms
Build predictive models
5
31
Train and Test
Train the model with sample data sets
Test and iterate
6
31
Verify and Deploy
Verify final model
Prepare visualizations and deploy
www.DigitalTransformationPro.com
© 2016 DigitalTransformationPro. All Rights Reserved

Figure 1. Data Mining Phases/Steps (source: Ramesh, 2016)

a. Define the Problem

The first step in this data mining phase is to define the problem. In this research, the problem is how to classify the data to use the prediction of future data. How well the model can predict future data depends on how the dataset is used to create the model. How good the data was pre-processing and the quality of the training dataset.

A few next steps must be performed carefully and in good condition to obtain good models, including data collection, splitting the dataset into training, testing data, and performing feature selection processes.

b. Data Collection

The sample data used in this research is credit report customer data for 2012 to 2022 at Bank Sumsel Babel Pagar Alam Branch. The following is the data used in this study, as shown in Table 1.

Table 1. Data Collection

No	Field	Description
1	Gender	This attribute explains the gender that does a lot of lousy credit (female or male).
2	Date of birth	This attribute explains the customer's date of birth.
3	Age	This attribute explains the age of customers with bad, semi-bad and lancer credit. Age is categorized as follows: Teenagers (17 - 25) with unmarried status, Adults (26 - 35) or age (17 - 25) with married status, Late Adults (36 - 45), Early Elderly (46 - 55), Late Elderly (56 - 65).
4	Marital status	This attribute explains the status of customers with bad credit (unmarried, married or widow/widower).
5	Work Category	This attribute explains the jobs with bad credit (entrepreneurs, civil servants, non-teachers, private employees, students, BUMN/BUMD or others).
6	KUR Credit Type	This attribute defines the type of KUR credit
7	Collectability	This attribute is a tool for measuring the status of harmful credit levels (smooth, still under consideration or rejected); collectability is categorized into three: 1 and 2 = soft, 3=semi smooth, 4 and 5=soaked
8	Credit Opening Date	This attribute explains the date the customer entered a credit agreement.
9	The tenor	This attribute explains how many months the credit instalments are; in the tenor attribute, it will be categorized into 3 categories as follows: short (until 12 months), medium (until 36 months) and long (until 60 months.)
10	Loan Amount	This attribute explains how much the customer's loan amount is; the loan amount is divided into 3 categories: Super Micro Loan (1 to 10 million), Micro (11 to 50 million) and KUR (people's business credit) Micro (51 to 500 million).
11	Net Income	This attribute explains the net income from the customer's opinion for one month, according to the BPS classification of income into 4 groups as follows: the first group is very high (>= Rp 3.500.000), the second group is high (Rp 2.500.000 to 3.500.000), the

		third group medium (Rp 1.500.000 to Rp 2.500.000), the fourth group is low (<= Rp 1.500.000).

The data that can be collected in this process is as much as 2055. However, the data is still in the form of raw data, which needs to be processed again. For that, the following process needs to be done: data pre-processing.

c. Pre-Processing Data

Data preparation must be done before data mining to be effective. Suyanto (2017) lists the following methods for pre-processing data:

1. Data cleansing (data cleaning). Data are cleaned of noise at this stage, such as redundant, inconsistent, and missing value data. Duplicate attributes and attributes with empty values will be eliminated.
2. Data Integration. This stage involves gathering data from many sources, aggregating diverse datasets, and performing data integration using a primary key.
3. Data selection. This process involves deleting attributes not essential for the research and choosing the relevant features from the dataset.
4. Data transformation. In this stage, the data chosen to produce the exact data for the mining process is changed in appearance.
5. Normalization. The process aims to give each attribute the same weight.
6. Data discretization. It is the procedure of dividing the values of a characteristic into intervals.

From the original data collected around 4155 data, after pre-processing the data by cleaning and transforming the data, 3624 data lines were obtained that were clean and ready to be formed into a model.

Data cleaning is correcting wrongly typed data, such as writing errors of the type of work that needs to be fixed by programming or manually (Rahm, 2000). In addition, some data needs to be transformed, such as age, changed into category data based on age range, marital status changed to their respective codes, and the number of loans changed to categories based on the loan size. Table 2 shows the transformation process, as shown below.

Table 2. Transformation process

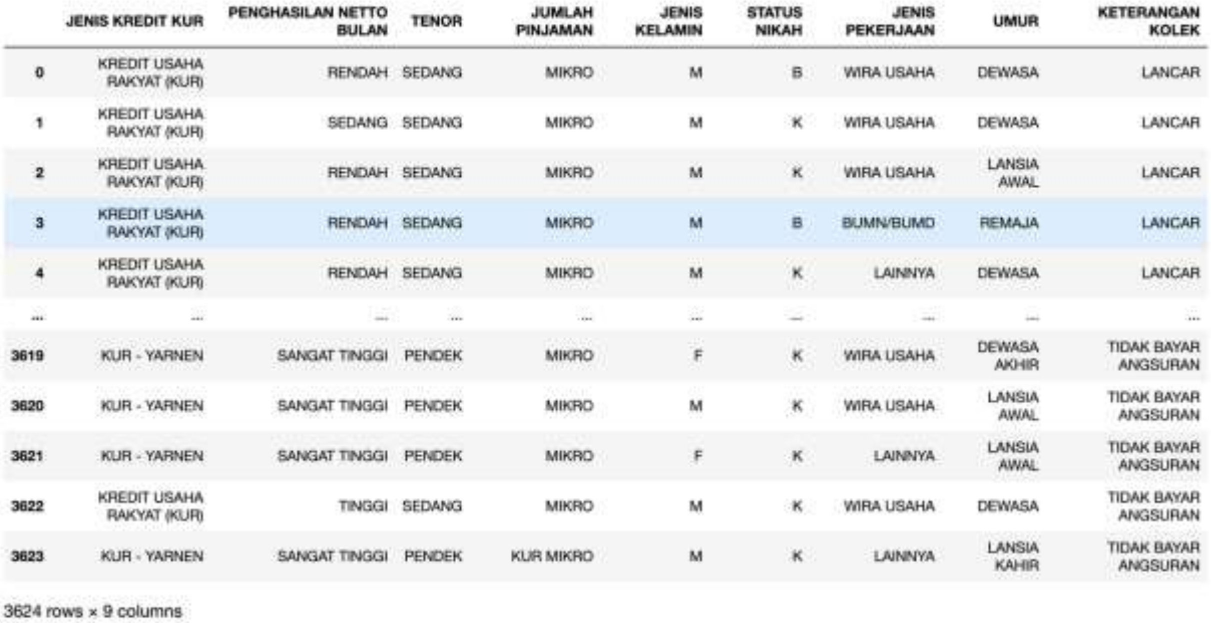
No	Field	Description
1	Gender	M = Male, F=Female
2	Age	Teenagers (17 - 25) with unmarried status, Adults (26 - 35) or age (17 - 25) with married status, Late Adults (36 - 45), Early Elderly (46 - 55), and Late Elderly (56 - 65).
3	Loan Amount	Super Micro Loan (1 to 10 million), Micro (11 to 50 million) and KUR (people's business credit) Micro (51 to 500 million).
4	Net Income	first group is very high (>= Rp 3.500.000), the second group is high (Rp 2.500.000 to 3.500.000), the third group medium

	(Rp 1.500.000 to Rp 2.500.000), the fourth group is low (<= Rp 1.500.000).

Another process includes removing the properties because they can be represented by other properties, like date of birth, which can be described by age, so the [date of birth] feature can be removed. Also, the feature [credit opening date] was removed because it should not be relevant to affect whether the load will be bad or good.

d. Data Modeling Process

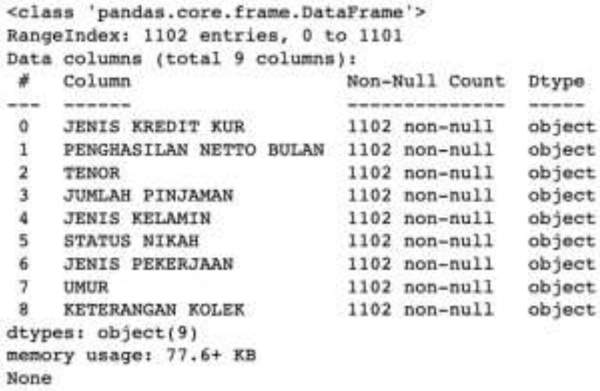
Before the modeling data was processed, the analysis data was conducted. In this process, we first count how much data will be used in our model and what features will be included. Figure 2 shows the data that will be processed, and Figure 3, shows all the features.



	JENIS KREDIT KUR	PENGHASILAN NETTO BULAN	TENOR	JUMLAH PINJAMAN	JENIS KELAMIN	STATUS NIKAH	JENIS PEKERJAAN	UMUR	KETERANGAN KOLEK
0	KREDIT USAHA RAKYAT (KUR)	RENDAH	SEDANG	MIKRO	M	B	WIRA USAHA	DEWASA	LANCAR
1	KREDIT USAHA RAKYAT (KUR)	SEDANG	SEDANG	MIKRO	M	K	WIRA USAHA	DEWASA	LANCAR
2	KREDIT USAHA RAKYAT (KUR)	RENDAH	SEDANG	MIKRO	M	K	WIRA USAHA	LANSIA AWAL	LANCAR
3	KREDIT USAHA RAKYAT (KUR)	RENDAH	SEDANG	MIKRO	M	B	BUMN/BUMD	REMAJA	LANCAR
4	KREDIT USAHA RAKYAT (KUR)	RENDAH	SEDANG	MIKRO	M	K	LAINNYA	DEWASA	LANCAR
...	...	...	...	...	...	...	...	...	...
3619	KUR - YARNEN	SANGAT TINGGI	PENDEK	MIKRO	F	K	WIRA USAHA	DEWASA AKHIR	TIDAK BAYAR ANGSURAN
3620	KUR - YARNEN	SANGAT TINGGI	PENDEK	MIKRO	M	K	WIRA USAHA	LANSIA AWAL	TIDAK BAYAR ANGSURAN
3621	KUR - YARNEN	SANGAT TINGGI	PENDEK	MIKRO	F	K	LAINNYA	LANSIA AWAL	TIDAK BAYAR ANGSURAN
3622	KREDIT USAHA RAKYAT (KUR)	TINGGI	SEDANG	MIKRO	M	K	WIRA USAHA	DEWASA	TIDAK BAYAR ANGSURAN
3623	KUR - YARNEN	SANGAT TINGGI	PENDEK	KUR MIKRO	M	K	LAINNYA	LANSIA KAHIR	TIDAK BAYAR ANGSURAN

3624 rows × 9 columns

Figure 2. The summary data will be processed to create the model.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1102 entries, 0 to 1101
Data columns (total 9 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   JENIS KREDIT KUR         1102 non-null   object
 1   PENGHASILAN NETTO BULAN  1102 non-null   object
 2   TENOR                    1102 non-null   object
 3   JUMLAH PINJAMAN          1102 non-null   object
 4   JENIS KELAMIN            1102 non-null   object
 5   STATUS NIKAH             1102 non-null   object
 6   JENIS PEKERJAAN          1102 non-null   object
 7   UMUR                     1102 non-null   object
 8   KETERANGAN KOLEK         1102 non-null   object
dtypes: object(9)
memory usage: 77.6+ KB
None
```

Figure 3. All features’ data

There are 13 features in our dataset, and it consists of 1102 rows of data. The last feature, “KETERANGAN KOLEK”, is used as the label. The analysis will examine whether the data is balanced for each label. Figure 4 shows the amount of data for each label that the data is balanced. The balanced data gives the same chance for each label to learn the data pattern.

```
KETERANGAN KOLEK
GAGAL BAYAR              999
LANCAR                  1034
TELAT BAYAR ANGSURAN     865
TIDAK BAYAR ANGSURAN     726
dtype: int64
```

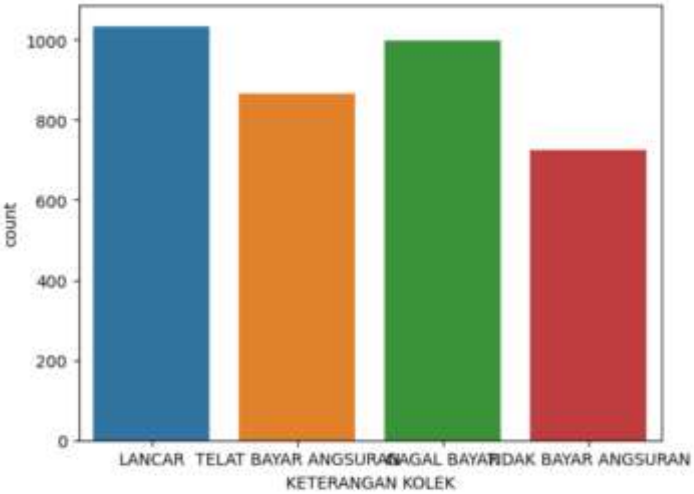
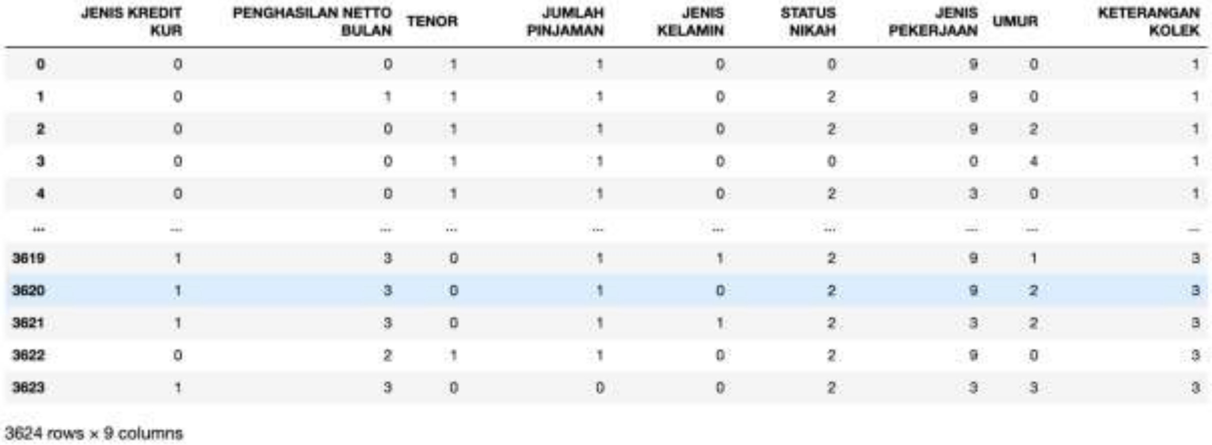


Figure 4. Amount of data for each label

The last process before we split data, the converted data from string categorical into numerical, needed to be performed. This process is necessary because our sklearn library needs these value types to allow the process to function, as shown in Figure 5.



	JENIS KREDIT KUR	PENGHASILAN NETTO BULAN	TENOR	JUMLAH PINJAMAN	JENIS KELAMIN	STATUS NIKAH	JENIS PEKERJAAN	UMUR	KETERANGAN KOLEK
0	0	0	1	1	0	0	9	0	1
1	0	1	1	1	0	2	9	0	1
2	0	0	1	1	0	2	9	2	1
3	0	0	1	1	0	0	0	4	1
4	0	0	1	1	0	2	3	0	1
...	...	...	...	...	...	...	...	...	...
3619	1	3	0	1	1	2	9	1	3
3620	1	3	0	1	0	2	9	2	3
3621	1	3	0	1	1	2	3	2	3
3622	0	2	1	1	0	2	9	0	3
3623	1	3	0	0	0	2	3	3	3

3624 rows × 9 columns

Figure 5. Data after conversion is already done.

In the following process, the dataset will be divided into two groups in this data modeling process. The first is the training dataset; the second group tests data with 70% and 30%. The sklearn library in Python is used, and a few classifier algorithms, such as Naïve Bayes, Support Vector Machine, and Random Forest, are applied to create the model. The comparison results are shown and discussed in the next chapter.

After getting the confusion matrix for each classifier, the following step process is conducted: feature selection (FS). This step is proposed to know which feature contributes most to our labels. The feature selection process uses a filter methods approach: chi-square and gain information. The result will be compared from two filters with the result without the feature selection. The results and discussion will be shown in the next chapter as well.

Results and Discussion

Based on the previous process, the dataset is divided into two groups: 70% for the training dataset and 30% for the testing dataset. Next, the training dataset is applied to three (3) different classifier algorithms named Naïve Bayes, Support Vector Machine, and Random Forest. The results are shown in Table 3 and Figure 6.

Table 3. Comparison result between Naïve Bayes, Support Vector Machine and Random Forest classifier.

No	Classifier	Accuracy %	
1	Naïve Bayes	On training dataset	57.36
		On testing dataset	54.42
2	Support Vector Machine	On training dataset	70.42
		On testing dataset	66.23
3	Random Forest	On training dataset	87.49
		On testing dataset	85.54

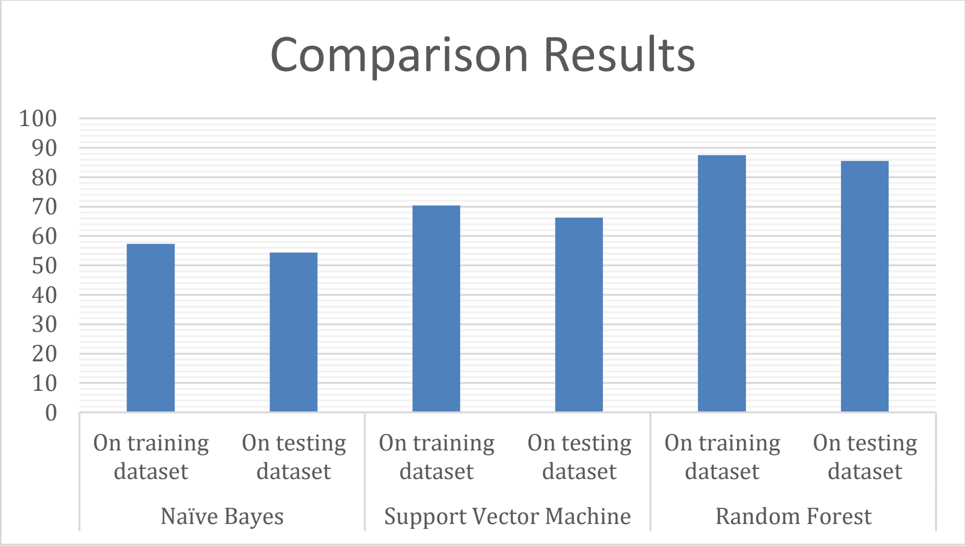


Figure 6. Comparison results

Based on Table 3 and Figure 6, The results show the trend the training dataset gives higher accuracy than the testing dataset. It is logical since the model learns from the training dataset. In this comparison, we focus more on testing the dataset. The results show Random Forest gives the best accuracy on the testing dataset, which is 85.54%, compared to the others. Further analysis, when we reached the training dataset with the testing dataset, the difference in accuracy was only slight. Our model is relatively stable, and our data is balanced between the underfitting and overfitting trade-off process. Figure 7 shows the confusion matrix for the Random Forest algorithm.

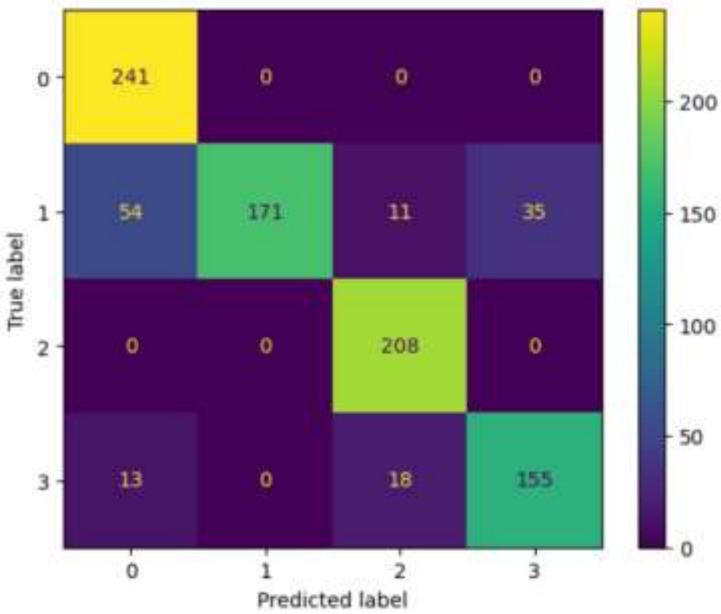
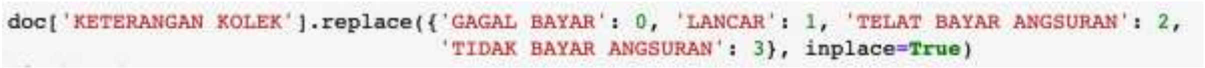


Figure 7. Confusion matrix results from the Random Forest algorithm.

Figure 7 shows label 0, ‘GAGAL BAYAR’, (Fail to Pay) and label 2, ‘TELAT BAYAR ANGSURAN’ (Payment is Late) (please refer to Figure 8); all data was predicted correctly. On the other hand, label 1 and label 3 get some data that was expected wrongly.



```
doc['KETERANGAN KOLEK'].replace({'GAGAL BAYAR': 0, 'LANCAR': 1, 'TELAT BAYAR ANGSURAN': 2,
                                 'TIDAK BAYAR ANGSURAN': 3}, inplace=True)
```

Figure 8. Mapping value for Label results.

Further analysis was conducted. Cross-validation is applied for each classifier to get an understanding of our model. In this process, the cross_validation function from sklearn.model_selection module is used. The parameter cv is set to 10, creating 10 different splitting datasets. Table 4 and Figure 9 show the cross-validation process results.

Table 4. Cross-validation comparison results

No	Classifier	10 Cross-validation	Max Accuracy %	Average Accuracy %
1	Naïve Bayes	60.06; 58.678; 57.30; 53.19; 55.80; 5 3.59; 55.52; 58.29; 55.80; 57.46;	60.06	56.57
2	Support Vector Machine	78.24; 72.18; 66.39; 65.01; 66.30; 63.26; 66.02; 75.14; 67.68; 68.23;	78.24	68.84

3	Random Forest	90.91; 91.46; 82.09; 77.69; 84.25; 84.81; 86.46; 87.57; 88.95; 86.19;	91.46	86.04

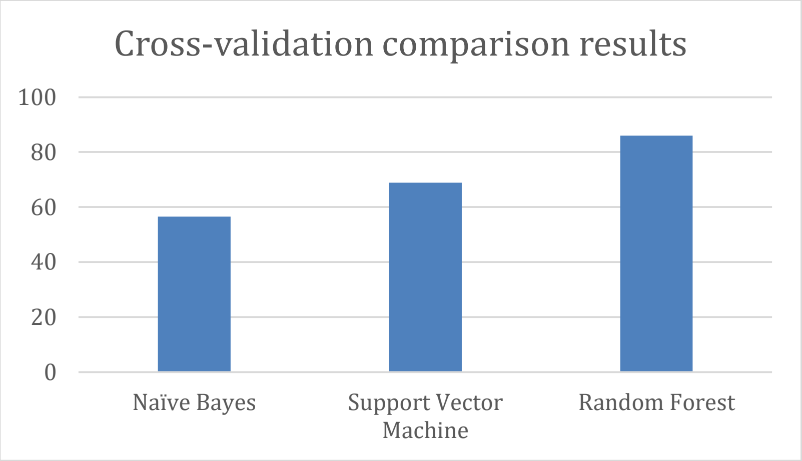


Figure 9. Cross-validation comparison results

Table 4 and Figure 9 show that the Random Forest still performs best. The max accuracy gives quite a high value of 91.46%. The average accuracy, 86.04%, obtained better than single running in the previous experiment, is only 85.54%. It means our model is very stable, and it is easy to get higher accuracy results.

The last experiment was done by applying the feature selection approach to analyze further which features affect our label more. Two methods used in feature selection experiments are chi-square and information gain, based on sklearn python library. The experiment was set by using a different number of features selected based on chi-square and gain information results suggestion. We started with all features and then continued with fewer features established. We show which features are used for each process with selected characteristics and report the result's accuracy, as shown in Table 5 and Figure 10.

Table 5. Comparison Feature Selection for Chi-square and Information Gain.

Chi-Square Method		
	Selected Features	Accuracy %
Naïve Bayes		
8	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JUMLAH PINJAMAN', 'JENIS KELAMIN' 'STATUS NIKAH', 'JENIS PEKERJAAN', 'UMUR']	54.42
7	'JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JUMLAH PINJAMAN', 'JENIS KELAMIN', 'JENIS PEKERJAAN', 'UMUR']	54.30
6	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JENIS KELAMIN', 'JENIS PEKERJAAN', 'UMUR']	54.64
5	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JENIS PEKERJAAN', 'UMUR']	54.64
4	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR'， 'JENIS PEKERJAAN']	54.64
3	['JENIS KREDIT KUR', 'TENOR', 'JENIS PEKERJAAN']	54.53
2	['TENOR', 'JENIS PEKERJAAN']	54.53

1	['TENOR']	49.56
Support Vector Machine		
8	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JUMLAH PINJAMAN', 'JENIS KELAMIN' 'STATUS NIKAH', 'JENIS PEKERJAAN', 'UMUR']	66.23
7	'JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JUMLAH PINJAMAN', 'JENIS KELAMIN', 'JENIS PEKERJAAN', 'UMUR']	66.23
6	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JENIS KELAMIN', 'JENIS PEKERJAAN', 'UMUR']	64.57
5	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JENIS PEKERJAAN', 'UMUR']	63.02
4	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JENIS PEKERJAAN']	59.93
3	['JENIS KREDIT KUR', 'TENOR', 'JENIS PEKERJAAN']	55.30
2	['TENOR', 'JENIS PEKERJAAN']	54.19
1	['TENOR']	49.56
Random Forest		
8	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JUMLAH PINJAMAN', 'JENIS KELAMIN', 'STATUS NIKAH', 'JENIS PEKERJAAN', 'UMUR']	85.43
7	'JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JUMLAH PINJAMAN', 'JENIS KELAMIN', 'JENIS PEKERJAAN', 'UMUR']	82.34
6	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JENIS KELAMIN', 'JENIS PEKERJAAN', 'UMUR']	79.47
5	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JENIS PEKERJAAN', 'UMUR']	74.06
4	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JENIS PEKERJAAN']	66.34
3	['JENIS KREDIT KUR', 'TENOR', 'JENIS PEKERJAAN']	57.17
2	['TENOR', 'JENIS PEKERJAAN']	56.51
1	['TENOR']	49.56
Gain Information		
Naïve Bayes		
8	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JUMLAH PINJAMAN', 'JENIS KELAMIN', 'STATUS NIKAH', 'JENIS PEKERJAAN', 'UMUR']	54.42
7	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JUMLAH PINJAMAN', 'STATUS NIKAH', 'JENIS PEKERJAAN', 'UMUR']	54.30
6	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'STATUS NIKAH', 'JENIS PEKERJAAN', 'UMUR']	54.30
4	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JENIS PEKERJAAN', 'UMUR']	54.64
4	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JENIS PEKERJAAN']	54.64
3	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR']	49.56
2	['JENIS KREDIT KUR', 'TENOR']	49.56
1	['TENOR']	49.56
Support Vector Machine		
8	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JUMLAH PINJAMAN', 'JENIS KELAMIN', 'STATUS NIKAH', 'JENIS PEKERJAAN', 'UMUR']	66.23

7	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JUMLAH PINJAMAN', 'STATUS NIKAH', 'JENIS PEKERJAAN', 'UMUR']	66.23
6	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'STATUS NIKAH', 'JENIS PEKERJAAN', 'UMUR']	65.01
5	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JENIS PEKERJAAN', 'UMUR']	63.02
4	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JENIS PEKERJAAN']	59.93
3	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR']	55.30
2	['JENIS KREDIT KUR', 'TENOR']	50.22
1	['TENOR']	49.56
Random Forest		
8	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JUMLAH PINJAMAN', 'JENIS KELAMIN', 'STATUS NIKAH', 'JENIS PEKERJAAN', 'UMUR']	85.76
7	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JUMLAH PINJAMAN', 'STATUS NIKAH', 'JENIS PEKERJAAN', 'UMUR']	80.79
6	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'STATUS NIKAH', 'JENIS PEKERJAAN', 'UMUR']	77.04
5	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JENIS PEKERJAAN', 'UMUR']	74.06
4	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR', 'JENIS PEKERJAAN']	66.34
3	['JENIS KREDIT KUR', 'PENGHASILAN NETTO BULAN', 'TENOR']	57.06
2	['JENIS KREDIT KUR', 'TENOR']	50.22
1	['TENOR']	49.56

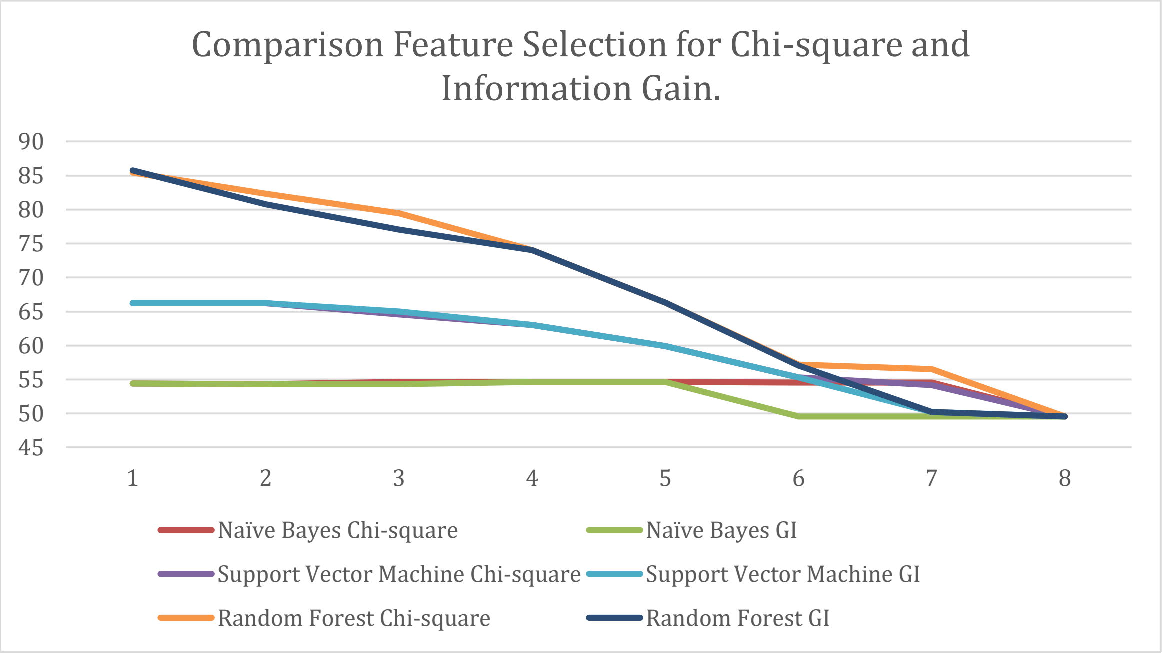


Figure 10. Comparison Feature Selection for Chi-square and Information Gain.

Table 5 and Figure 10 show some trends that the accuracy will be higher when using all features. When the features used are fewer, the accuracy is reduced. That indicates all properties are essential to influence our label. Although the feature selection did not successfully reduce the number of properties of the dataset, the results from the 3 classifier algorithms were consistent.

Chi-square and Gain Information also give the different selected features, but accuracy still shows similar trends. The most important feature is 'TENOR', in which two approaches, chi-square, and Gian information, suggest the same properties.

Conclusion

In this paper, the data mining process was done following the best practices in data mining procedure. First, the data was collected and used pre-processing already prepared carefully. The experiments were conducted comprehensively, the results were compared, and the analysis was explained in detail. After the dataset was ready, the training process model was achieved by splitting the dataset into two parts, the training and testing sets, by 70% and 30% based on 3 classifier algorithms: Naïve Bayes, Support Vector Machine, and Random Forest. The results show that Random Forest obtained the best accuracy performance, 85.54%.

Next, the cross-validation was conducted, in which cv is set to 10. The model shows very stable, and it is easy to get higher accuracy results. The average accuracy, 86.04%, was better than single running in the previous experiment, which was only 85.54%. The last experiment was conducted, the feature selection approach, in which the result shows that all dataset properties are important to influence our label.

References

Haryati, N.T., Negara, E.S. and Kurniawan, T.B. (2023) *Klasifikasi Pemberian Beasiswa Berprestasi Manggunakan Perbadingan Tiga Algoritma.* Jurnal Tekno Kompak, 2023, Vol.17, No. 1.

Kusrini and Emha, T, 2009, *Algoritma Data Mining*, Yogyakarta, Andi.

Mints, A., (2017), *Classification of Tasks of Data Mining and Data Processing in The Economy.* Baltic Journal of Economic Studies, 3(3): 47-52.

Misinem., Kurniawan, T.B., Zakaria, M.Z. and Uzailee, M.A.A (2022) *Sentiment Analysis on Users' Satisfaction for Mobile Banking Apps in Malaysia.* Journal of Data Science, 2022 (11). pp. 1-15

Mustafa & Sarah, (2021), *Data Mining preparation: Process, Techniques and Major Issues in Data Analysis*, IOP Conference Series: Materials Science and Engineering, 1090 (1).

Putri D. N., Kurniawan, TB, Negara, E.S., Kunang, Y.N., and Misinem, (2022) *Classification of MTI Student Thesis Documents at Bina Darma University Palembang Using Naïve Bayes.* Journal of Data Science, 2022 (17). pp. 1-12.

Rahm, E., (2000), *Data Cleaning: Problems and Current Approaches.* IEEE Data Eng. Bull.

Ramesh Dontha, (2018), *Data Mining Step*, (Online, access on 2022, July 20, www.digitaltransformationpro.com)

Sugiyono, (2014), *Metode Penelitian Pendidikan Pendekatan Kuantitatif, Kualitatif, dan R&D*, Alfabert, Bandung.

Suyanto, (2017), *Data Mining Untuk Klasifikasi dan Klasterisasi Data*, Bandung, Informatika.

Turban. E., (2005), *Decision Support System and Intelligent Systems*, Yogyakarta. Andi.