e-ISSN: 2614-8404
p-ISSN:2776-3234

Application of Information Gain to Select Attributes in Improving Naive Bayes Accuracy in Predicting Customer's Payment Capability

Herfandi[1*], Mohammad Taufan Asri Zaen[2], Yuliadi[3], M. Julkarnain[4], Fahri Hamdani [5]
[1,3,4,5] Teknik Informatika, Universitas Teknologi Sumbawa, Sumbawa, Indonesia
[2] Sistem Informasi STMK Lombok, Lombok Tengah, Indonesia
Email: [1]herfandi@uts.ac.id, [2]opanzain@gmail.com, [3,]yuliadi@uts.ac.id, [4]m.julkarnain@uts.ac.id, [5]fahri.hamdani@uts.ac.id

Abstract – The customer is the main factor in the running of PT. XYZ. A good understanding of customers is very important for predicting the capability of customers to pay. The implementation of credit collectibility is used to determine the quality of customer credit, one of which is the customer's capability to pay interest and principal on time. While manually, it is very difficult to accurately predict the capability of customer credit payments. Data mining techniques with the Naïve Bayes algorithm were chosen to classify customers to be able to find patterns, analyze and predict, because they have good performance, are efficient, and simple. The Naïve Bayes algorithm has a weakness in terms of sensitivity to many attributes, so the accuracy is low. Based on the problem stated, his study will apply the Information Gain method to select the most influential attribute on the label in order to increase the accuracy of the Naïve Bayes algorithm. This research produces a new dataset with seven attributes: TENOR, SALARY, DOWN PAYMENT, INSTALLMENT, APPROVAL, OTR CLASS, AGE with Labels: Status and Id: Id number based on the Information Gain method. The dataset comparison process with 995 data records showed an increase in accuracy, precision, and AUC using the new dataset compared to the old dataset, but in the t-Test test with an alpha value = 0.05 there is a difference but not significant. In the evaluation process, performance experienced a significant increase in the use of new datasets with the following percentages of performance improvement: accuracy = 8%, precision = 18.42%, recall = 17.65% and AUC= 0.057%. The results of this study obtained AUC of 0.876, accuracy of 87.88%, precision of 61.90%, and recall of 76.47%, and classified into good classification.

Keywords: credit collectibility, customer prediction, Data Mining, Naive Bayes, Information Gain

I. INTRODUCTION

In the present era, business strategy has developed very significantly. Customers occupy a very important position in this regard. Customers are also a major factor in the running of PT. XYZ. A good understanding of customers is very important for predicting the capability of customers to pay in the future. The implementation of credit collectibility is used to determine the quality of customer credit[1], one of which is the customer's capability to pay interest and principal on time that has been mutually agreed upon[2]. The problem currently being faced is that manually it is very difficult to predict the capability of customer credit payments accurately based on the dataset they have[3], and many companies have difficulty identifying customers who are able to pay on time[4].

The data mining technique using the customer classification approach is an approach that is widely used to find patterns, analyze and predict[5]. With a customer classification approach based on existing datasets, we can predict a customer's payment capability[6]. Meanwhile, manually, it is very difficult to accurately predict the capability of customer credit payments based on the dataset they have.

Currently, there are many data mining techniques with classification approach algorithms that have been used to find patterns, analyze and predict customer behavior, such as the Decision Tree Algorithm[7], Neural Network[8], Support Vector Machine[9], Naive Bayes[10], and K-Nearest Neighbor [11]. The Naïve Bayes algorithm is the algorithm with the most widely used classification approach and was chosen to classify customers because it has good performance, is efficient, and is simple in terms of finding patterns, analyzing and predicting[12]. The Naïve Bayes algorithm has one drawback, namely that it is sensitive to many attributes, so the accuracy is low. Selection or choosing the attribute that has the most influence on the label is very important for the Nave Bayes algorithm to be able to increase the accuracy of the algorithm[13].

One of the methods for determining the best attribute or selecting the attribute that has the most influence on the label is the Information Gain method. The Information Gain method is superior to other methods because the Information Gain method will measure how much absence and presence of an attribute that plays a role in making good classification decisions in any class or label. The Information Gain method is one of the successful attribute selection approaches in classification[14].

e-ISSN: 2614-8404
p-ISSN:2776-3234

In this study, the authors apply the Information Gain method to select the most influential attribute on the label to be applied to the new dataset in an effort to improve the accuracy of the Naïve Bayes Algorithm in predicting the payment capability of customers at PT. XYZ.

1.1 Collectibility of Credit

Credit collectibility is a credit quality status that is taken from a person's score or track record in the banking world[15]. This quality is based on 3 main standards, one of which is the customer's capability to pay principal and interest on time that has been mutually agreed upon[16]. Low collectibility or bad loans can affect the economic condition of a business and worsen the trickle down effect on the overall economy, where this has an impact on the company's growth and income in the future[17]. There are several important elements in the provision of a credit facility, namely Trust, Agreement, Term, Risk and Rewards[18]. To find out how to provide credit to customers based on good credit quality, it is necessary to accurately predict the capability of customer credit payments as a reference for management in making decisions to improve credit quality and collectability[19].

1.2 Data Mining

Data mining is a process of using existing or past data, then processing it so that it finds patterns, meaningful relationships, and trends by examining a set of stored data using statistical and mathematical techniques[20]. Data mining became popular in the 1990s as a solution for extracting previously unknown patterns and information based on a set of data[21]. Data mining can complete several jobs and is divided into four groups, namely prediction modeling, cluster analysis, association analysis, and anomaly detection[22]. However, data mining techniques can also be applied to other data representations, such as spatial, text-based, and multimedia (image) data domain[23]. Data Mining can also be defined as the process of extracting information from large data sets through the use of algorithms with techniques taken from the field of statistics and Database Management Systems[24].

II. RESEARCH METHODOLOGY

The object of this research is the Population History database of customer payments in the New Motorcycle (NMC) program at PT. XYZ year 2019-2020. The use of population data is expected to be able to find out what kind of customer criteria can complete credit on time. Software instrument used are:

1) Delphi 7 : Programming language
2) MySQL : Database system software
3) Rapid Miner 9.9 : Tools that help in testing
4) SQLYog : Database administrator application

2.1 Research Process Flow

The research process flow adopts the CRISP-DM (Cross Standard Industries Process for Data Mining) model, where the CRIPS-DM model is modified according to the research stages[25]:

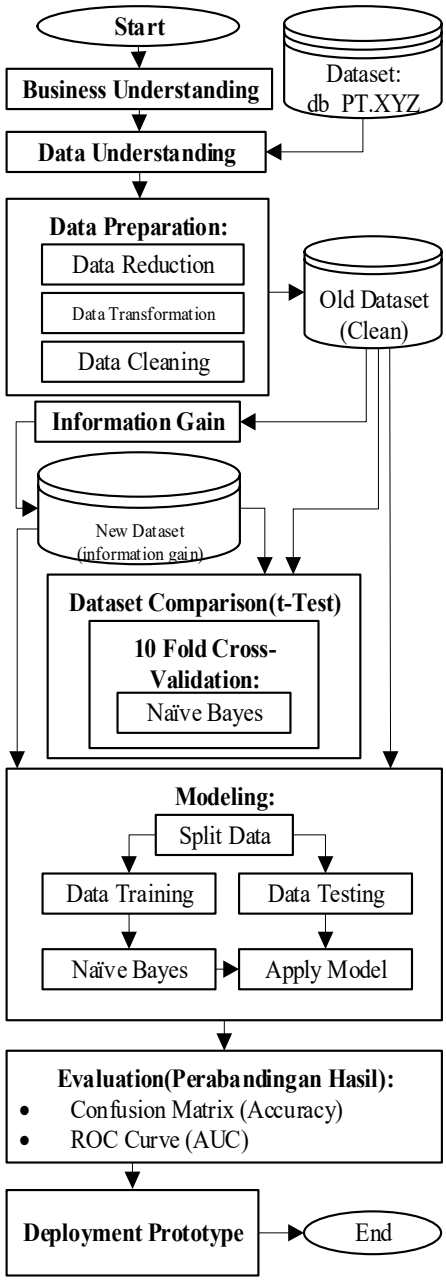

Figure 1. Research Process Flow

e-ISSN: 2614-8404
p-ISSN:2776-3234

The flow of the research process described can be explained as follows:

a. Business Understanding: at this stage, the researcher understands the problem in the research object and then looks for solutions and goals to solve it.

b. Understanding data: at this stage, the researcher determines and collects what data is needed and then defines it according to the solution and research objectives.

c. Data Preparation: at this stage, the researcher cleans the data so that he gets a clean dataset that will be used as a classification model.

d. Information gain: researchers will choose the best or most influential attribute on the label and then make it a new dataset.

e. Dataset Comparison: at this stage, the researcher compares the old dataset and the new dataset on the Naïve Bayes algorithm by means of a different test using t-Test.

f. Modeling: at this stage, the researcher will model the data based on the dataset and a new dataset with a 90% data split (Training) : 10% (Testing)

g. Evaluation: at this stage, the researcher will compare the results of measuring accuracy, precision, recall, and AUC on the old dataset and the new dataset.

h. Development: this stage, the researcher builds a prototype that will be used to predict the customer's payment capability.

2.2 Research Formulas

a. Naïve Bayes Algorithm

The Naive Bayes algorithm is an algorithm with a classification approach for predicting a simple probabilistic based that was put forward by the English scientist Thomas Bayes which is based on the application of Bayes' theorem or rules with the assumption of strong independence on features, meaning that a feature in a data is not related to the existence or the absence of other features in the same data[26].

The advantage of using the Nave Bayes algorithm classification approach is that the Nave Bayes algorithm only requires a small amount of training data needed for the classification process[27]. The Naïve Bayes algorithm classification approach has been proven to be applicable in real and complex situations[27].

The Naïve Bayes algorithm classification approach can be defined as follows[28]:

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

where

P(c|x) : posterior probability of the label (target) to the predictor (attribute)

P(x|c) : likelihood, which is the probability of the predictor on the label

P(c) : prior probability of the label

P(x) : prior probability of the predictor.

b. The Information Gain Method

The Information Gain method is a method of selecting or selecting attributes in the simplest dataset by only ranking the attributes. The Information Gain method is widely used in the application of data analysis categorization, text, microarray, and image data analysis[29]. In the selection of dataset attributes, the pattern classification approach plays a very important role[30]. The Information Gain method can help reduce noise caused by irrelevant attributes[31]. The initial step that must be done is to determine the best attribute value by calculating the entropy value. Entropy is the process of using the probability of certain events or attributes to measure class uncertainty[32]. After calculating the entropy value, then we can only calculate the Information Gain method[33].

Calculating entropy is defined as follows[33]:

$$\boldsymbol{Entropy(S) = \sum_{i}^{c} - P_i \log_2 P_i} \quad (2)$$

where c is the number of values on the classification label and Pi is the number of samples for class i.

The information gain method is defined as follows [23]:

$$\boldsymbol{Gain\ (S, A) = Entropy(S) - \sum_{Values\ (A)} \frac{|S_v|}{S} Entropy(S_v)} \quad (3)$$

where A is an attribute, v is a possible value for attribute A, Values(A) is the set of possible values for A, |Sv| is the number of samples for the value of v, |S| is the sum of all data samples and Entropy (Sv) is the entropy for samples that have a value of v.

2.3 Prototype Development Design

The following is the flow of the Flowchart which is implemented into the prototype payment capability prediction:

e-ISSN: 2614-8404
p-ISSN:2776-3234

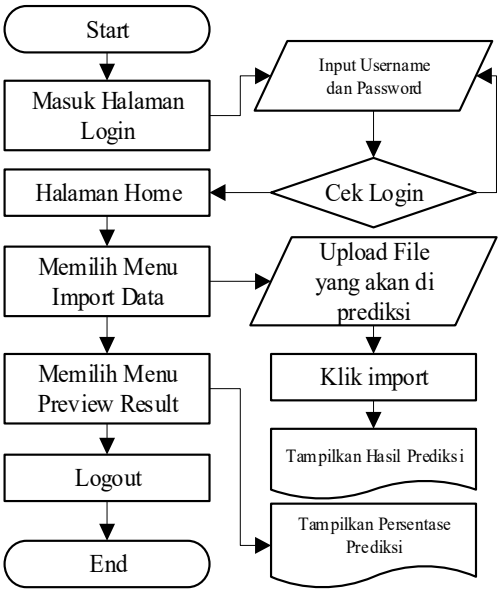

Figure 2. Flowchart Prototype

The user opens the application where the application will directly direct the user to the login page menu, the user enters the username and password to enter the home page, to predict the new data, the user selects the import data menu and uploads the data to be predicted. The user clicks import to see the prediction result display. After that, the user can choose the preview result menu to see the percentage of the predicted results from the data, to exit the user logs out.

III. RESULTS AND DISCUSSION

This research begins with the business understanding stage, where researchers find the problem manually is very difficult to accurately predict the capability of customer credit payments based on the dataset they have, as well as the Naïve Bayes algorithm which is sensitive to many attributes. The next stage is data understanding. At this stage, the researcher combines data from 4 tables, namely active management summary contracts, order management application hiders, order management applications, and acctmgmt.ar contracts. From this process, researchers get 995 data from 2019-2020. The data taken is contract data from 3 branch offices, namely Serang Branch, Bandung Branch, and Tasik Branch, with a total of 27 attributes and 1 label. Furthermore, data preparation, at this stage, the researcher performs data reduction, namely the selection of attributes that are relevant to the target to be achieved. The selected attributes are expected to be determinants of the information to be processed. After data reduction is done, 13 attributes are generated, 1 ID and 1 label. The data transformation stage here is used to obtain a suitable representation for the specific task to be performed. For example: the age value of 17-25 becomes the new value for late teenagers, as can be seen in Table 1.

Table 1. Data Transformation

Atribut	Value	New Value	Atribut	Value	New Value
STATUS (LABEL)	WO	BAD DEBT	PEKERJAAN	Wiraswasta	Wiraswasta
	PT & CL	CLEAR		Petani/Pekebun	Petani/Pekebun
ANGSURAN	3,000,000 - 5,000,000	Tinggi		Pelajar/Mahasiswa	Pelajar/Mahasiswa
	1,700,000 - 2,500,000	Menengah		Pegawai Negeri Sipil	Pegawai Negeri Sipil
	500,000 - 1,500,000	Rendah		Mengurus Rumah Tangga	Mengurus Rumah Tangga
APPROVAL	Grey	BM		Karyawan Swasta	Karyawan Swasta
	White	CAH		Buruh	Buruh
CABANG	32003	Bandung		Belum/Tidak Bekerja	Belum/Tidak Bekerja
	32002	Tasik		Lainnya	Lainnya
	32001	Serang	STTS TINGGAL	H02	TUMPANG
GAJI	> 10000000	SANGAT TINGGI		H01	PRIBADI
	7000000 - 10000000	TINGGI	TENOR	>31	Sangat Lama
	4000000 - 6000000	MENENGAH		25 - 30	Lama
	2500000 - 3500000	RENDAH		19 - 24	Sedang
	500000 - 2300000	SANGAT RENDAH		13 -18	Singkat
JK	F	Wanita		09 -12	Sangant Singkat
	M	Pria	TOTAL DP	>13,000,000	Sangat Tinggi
KELAS OTR	> 40,000,000	Premium		9,000,000 - 12,000,000	Tinggi
	30,000,000 - 40,000,000	Sport		5,000,000 - 8,000,000	Menengah
	25,000,000 - 30,000,000	Matic150		2,500,000 - 4,500,000	Rendah
	19,000,000 - 24,000,000	Bebek2		1,350,000 - 2,000,000	Sangat Rendah
	13,000,000 - 18,000,000	Bebek1	UMUR	56-65	Lansia Akhir
KODE PRODUK	KPM	Non Rek		46-55	Lansia Awal
	KSM	Rek		36-45	Dewasa Akhir
STATUS NIKAH	D	Duda/Janda		26-35	Dewasa Awal
	M	Menikah		17-25	Remaja Akhir
	S	Single	Id	Nomor Kontrak	id_number

Then the data cleaning stage is done by filling in the blank data based on the average value, and replacing the data values that do not match. The results of all the stages above produce a dataset (Clean) or old dataset, which can be seen in table 2.

Table 2. Dataset Clean or Old Dataset

Id	UMUR	STTS TINGGAL	GAJI	PEKERJAAN	STTS NIKAH	.....	ANGSURAN	KELAS OTR	KELAS STATUS
1	DEWASA AKHIR	PRIBADI	MENENGAH	WIRASWASTA	MENIKAH	....	RENDAH	MATIC150	CLEAR
2	DEWASA AKHIR	PRIBADI	MENENGAH	KARYAWAN SWASTA	MENIKAH	....	RENDAH	BEBEK2	CLEAR
3	DEWASA AWAL	TUMPANG	RENDAH	KARYAWAN SWASTA	DUDA/JANDA	....	RENDAH	BEBEK1	CLEAR
4	DEWASA AKHIR	PRIBADI	RENDAH	BURUH	MENIKAH	....	RENDAH	BEBEK1	CLEAR
5	DEWASA AWAL	TUMPANG	MENENGAH	KARYAWAN SWASTA	MENIKAH	....	RENDAH	BEBEK1	CLEAR
6	DEWASA AKHIR	PRIBADI	RENDAH	KARYAWAN SWASTA	MENIKAH	....	RENDAH	BEBEK1	CLEAR
7	REMAJA AKHIR	TUMPANG	SANGAT RENDAH	PELAJAR/MAHASISWA	SINGLE	....	RENDAH	BEBEK1	CLEAR
8	DEWASA AWAL	TUMPANG	RENDAH	KARYAWAN SWASTA	MENIKAH	....	RENDAH	BEBEK2	CLEAR
9	DEWASA AWAL	TUMPANG	RENDAH	KARYAWAN SWASTA	MENIKAH	....	RENDAH	MATIC150	CLEAR
.....	....	....	....	....	....	....	....	....	.....
993	DEWASA AKHIR	TUMPANG	RENDAH	BURUH	DUDA/JANDA	....	RENDAH	BEBEK1	BAD DEBT
994	REMAJA AKHIR	TUMPANG	SANGAT RENDAH	KARYAWAN SWASTA	SINGLE	....	RENDAH	MATIC150	BAD DEBT
995	LANSIA AWAL	PRIBADI	RENDAH	BURUH	MENIKAH	....	RENDAH	BEBEK1	BAD DEBT

3.1 The experimental process

Researchers conducted an experimental process using Rapidminer 9.9 tools to perform Data Cleaning, Information Gain, Dataset Comparison, Modeling, and Evaluation. The selection of Rapidminer tools is considered capable of being used for research, prototyping, and supporting all steps of the data mining process such as data preparation, result visualization, validation, and optimization[34]. The experimental process can be seen in Figure 3.

e-ISSN: 2614-8404
p-ISSN:2776-3234

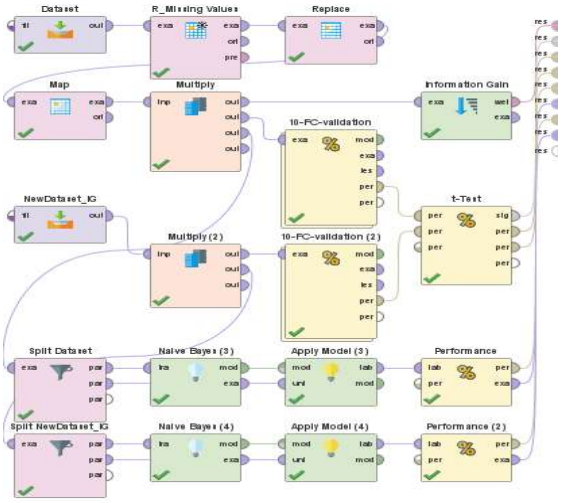

Figure 3. The experiment process

3.2 The Information Gain

The Information Gain method process uses an old dataset of 995 data with specifications of 13 attributes, 1 ID and 1 label to get the best or most influential attribute on the label. The results of the Information Gain Method can be seen in figure 4.

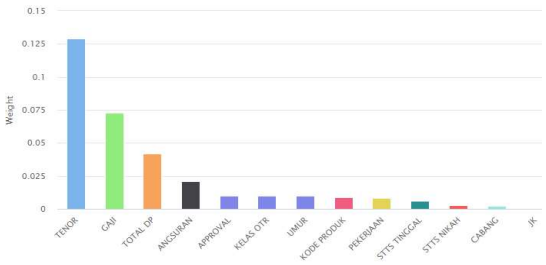

Figure 4. Results of the Information Gain method

The researcher will use the attribute that has the most influence on the label. This attribute will later become an attribute in the new dataset based on the calculation results of the Information Gain method. New attribute by weight method Information Gain can be seen in table 3.

Table 3. New Attributes

No	Attribut	Weight	Status
1	TENOR	0.12898411785188602	new atribut
2	GAJI	0.07251636940706196	new atribut
3	TOTAL DP	0.04205947913917707	new atribut
4	ANGSURAN	0.020693432559648506	new atribut
5	APPROVAL	0.010477564333977618	new atribut
6	KELAS OTR	0.009832172681986773	new atribut
7	UMUR	0.00970525033831704	new atribut
8	KODE PRODUK	0.008562723828346774	delete
9	PEKERJAAN	0.00809223987708052	delete
10	STTS TINGGAL	0.006496154668806264	delete
11	STTS NIKAH	0.0026098786500883264	delete
12	CABANG	0.0020787262163179943	delete
13	JK	0	delete

Based on the calculation results of the Information Gain method in table 3, the new dataset will have the following seven attributes: TENOR, SALARY, DOWN PAYMENT, INSTALLMENT, APPROVAL, OTR CLASS, AGE with Label: Status and Id: Id_number.

3.3 Dataset Comparison

In this process, the researcher compares the old dataset and the new dataset on the Naïve Bayes algorithm using 10 fold cross-validation by dividing the dataset into 10 parts, of the 10 parts of the data, 9 parts are used as training data, and the remaining 1 part is used as testing data. Then a different test was performed using t-Test to determine the significant difference in the dataset used in the Naïve Bayes algorithm. The results can be seen in the table below.

Table 4. Comparison of 10 fold cross-validation

	Dataset+NB	New_DatasetIG+NB
Accuracy	78.89% +/- 3.70%	81.20% +/- 3.64%
Precision	41.40% +/- 9.38%	45.85% +/- 9.58%
Recall	45.92% +/- 10.39%	44.77% +/- 10.35%
AUC	0.804 +/- 0.066	0.822 +/- 0.055

Table 5. Different Test (t-Test)

A	B	C
	0.789 +/- 0.037	0.812 +/- 0.036
0.789 +/- 0.037		0.178
0.812 +/- 0.036		

Based on the results of the 10-fold cross-validation test, there is an increase in accuracy, precision, and AUC in the Naïve Bayes algorithm using the new dataset compared to the old dataset, but in the t-Test test with an alpha value of 0.05 there is a difference but not significant.

3.4 Modeling

The researcher will model the old dataset and new dataset using the Naïve Bayes algorithm based on the split data operator in rundom subsets with a ratio of 90% (Training): 10% (Testing) as shown in figure 3.

3.5 Evaluation.

This process will compare the model testing of the old dataset and the new dataset that have been determined in the modeling process by measuring their performance. The results of the performance comparison can be seen in table 6.

e-ISSN: 2614-8404
p-ISSN:2776-3234

The following are the results of model testing based on performance measurements and AUC on the old dataset. The results can be seen in figures 5 and 6.

PerformanceVector

```
PerformanceVector:
accuracy: 79.80%
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  69      7
BAD DEBT:       13      10
precision: 43.48% (positive class: BAD DEBT)
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  69      7
BAD DEBT:       13      10
recall: 58.82% (positive class: BAD DEBT)
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  69      7
BAD DEBT:       13      10
AUC (optimistic): 0.819 (positive class: BAD DEBT)
AUC: 0.819 (positive class: BAD DEBT)
AUC (pessimistic): 0.819 (positive class: BAD DEBT)
```

Figure 5. Performance Vector Old Dataset

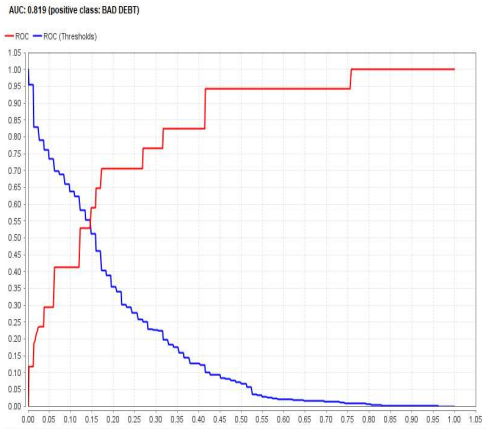

Figure 6. ROC Curve (AUC) Old Dataset

While the results of model testing based on performance measurements and AUC on the new dataset can be seen in figures 7 and 8.

PerformanceVector

```
PerformanceVector:
accuracy: 87.88%
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  74      4
BAD DEBT:       8       13
precision: 61.90% (positive class: BAD DEBT)
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  74      4
BAD DEBT:       8       13
recall: 76.47% (positive class: BAD DEBT)
ConfusionMatrix:
True:   CLEAR   BAD DEBT
CLEAR:  74      4
BAD DEBT:       8       13
AUC (optimistic): 0.881 (positive class: BAD DEBT)
AUC: 0.876 (positive class: BAD DEBT)
AUC (pessimistic): 0.872 (positive class: BAD DEBT)
```

Figure 7. Performance Vector New Dataset

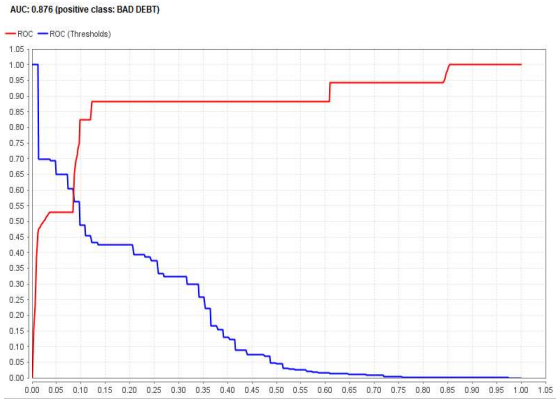

Figure 8. ROC Curve (AUC) New Dataset

Based on the results of the performance and AUC above, the result of the comparison performance is as follows:

Table 6. Result Comparison Performance

Dataset + Algoritma Naïve Bayes			
	true CLEAR	true BAD DEBT	class precision
pred. CLEAR	69	7	90.79%
pred. BAD DEBT	13	10	43.48%
class recall	84.15%	58.82%	
accuracy	79.80%		
precision	43.48% (positive class: BAD DEBT)		
recall	58.82% (positive class: BAD DEBT)		
AUC	0.819 (positive class: BAD DEBT)		
New Dataset (Information Gain) + Algoritma Naïve Bayes			
	true CLEAR	true BAD DEBT	class precision
pred. CLEAR	74	4	94.87%
pred. BAD DEBT	8	13	61.90%
class recall	90.24%	76.47%	
accuracy	87.88%		
precision	61.90% (positive class: BAD DEBT)		
recall	76.47% (positive class: BAD DEBT)		
AUC	0.876 (positive class: BAD DEBT)		

Based on the measurement of the performance model with the operator split data in a rundom subset with a comparison of 90% (Training): 10% (Testing) which can be seen in table 6. The results show that by using a new dataset based on the calculation of the Information Gain method on the Naïve algorithm Bayes is much better than the old dataset, both in terms of Accuracy, Precision, Recall, and AUC. The percentage increase in performance using the new dataset is as follows: Accuracy = 8%, Precision = 18.42%, Recall = 17.65% and AUC = 0.057%.

3.6 Development

This process is based on the results of the dataset comparison and evaluation of the above model testing. It is known that the Naïve Bayes algorithm has a good level of accuracy and performance by using a new

e-ISSN: 2614-8404
p-ISSN:2776-3234

dataset based on the calculation of the Information Gain method, so that the rules generated by the Nave Bayes algorithm can be used as rules for making prototypes. The researcher hopes that this prototype can make it easier for PT. XYZ in predicting the capability of customers to pay. for the flowchart prototype can be seen in figure 3.

The prototype used in this study was made desktop-based with programming language using Delphi 7.0 and database using MySQL. The display for the main form of the Graphical User Interface (GUI) prototype predicting the capability of customer credit payments can be seen in the image below.

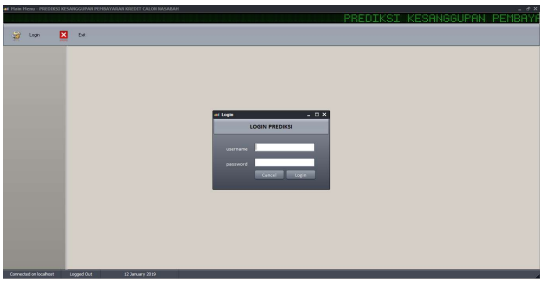

Figure 9. Login Form

When the application is running, the first form will be displayed, which is the login form. For security of access, the user must have a username and password.

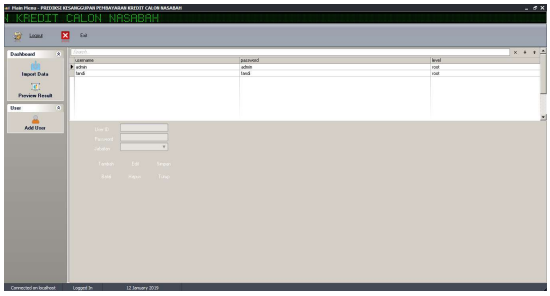

Figure 10. Form for Managing Users

The Manage User form is used to view information about registered users, and admins can add, edit, delete users. User accounts can predict new data based on models that have been deployed in the application.

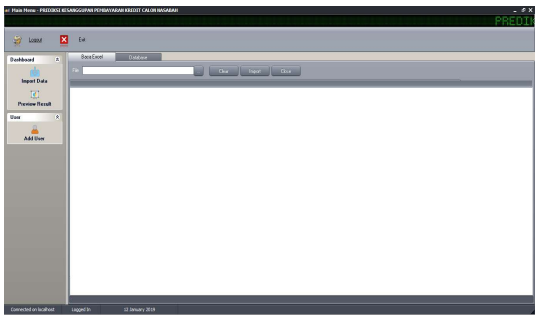

Figure 11. Form File Upload

The import menu is used by the user to make predictions on new data by uploading the data to be predicted, after uploading the user clicks Import to see the display of the prediction results and the user can select the preview result menu to see the percentage number of results from the predicted data.

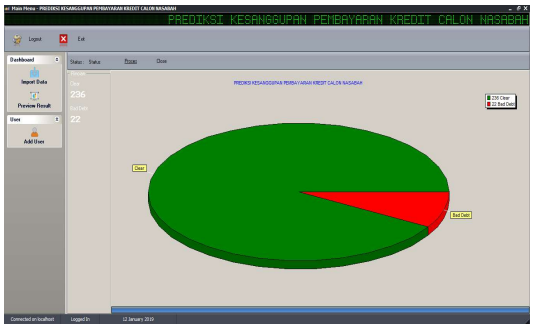

Figure 12. Prediction Result Percentage Form

Testing is carried out with the aim of knowing whether the application is built according to the expected functionality.

Table 7. The Black Box Test

Kelas Uji	Butir Uji	Jenis Pengujian
File Upload	Pilih File	Black Box
	Upload File	Black Box
Dashboard	Lihat Grafik	Black Box

In testing, the upload file is divided into two parts, namely, selecting the file and uploading the file.

Table 8. File Upload Test

Kasus dan Hasil Uji (Data Sesuai)			
File yang di Upload	File dalam bentuk etc.xls	Dapat melakukan pilih file upoload	Diterima
Klick Upload	File berhasil di import oleh sistem	Dapat melakukan upload data	Diterima
Kasus dan Hasil Uji (Data salah)			
Data Masukan	Yang Diharapkan	Pengamatan	Kesimpulan
Extension berbeda, tidak sesuai dengan yang di inginkan sistem	Data Tidak Tersimpan	Data tidak tersimpan dan menampilkan kesalahan	Diterima

In testing the results, the user just clicks on the results.

Table 9. Testing Results

Kasus dan Hasil Uji (Data sesuai)			
Data Masukan	Yang Diharapkan	Pengamatan	Kesimpulan
File yang telah di upload	Berhasil menampilkan Grafik persentase	Dapat menampilkan grafik presentase	Diterima
Klik Tombol Detail	Dapat menampilkan detail customer baik yang disetujui ataupun ditolak	Dapat menampilkan sesuai yang diharapkan	Diterima

Based on the Black Box testing that has been done, it explains that the application that was built has been running well and as expected and The results of this study obtained AUC of 0.876, accuracy of 87.88%, precision of 61.90%, and recall of 76.47%, and classified into good classification.

e-ISSN: 2614-8404
p-ISSN:2776-3234

IV. CONCLUSION

Based on the Research Process Flow that has been carried out by the researchers, it can be concluded:

1. The data preparation process produces 13 attributes, 1 ID and 1 label with a total of 995 data from 2019-2020.
2. The process of calculating the Information Gain method on the old dataset produces a new dataset with seven attributes: TENOR, SALARY, DOWN PAYMENT, INSTALLMENT, APPROVAL, OTR CLASS, AGE and Label: Status and Id: Id_number.
3. In the dataset comparison process, there is an increase in accuracy, precision, and AUC using the new dataset compared to the old dataset, but in the t-Test test with an alpha value of 0.05, there is a difference but not significant.
4. In the evaluation process, performance experienced a significant increase in the use of new datasets with the following percentage increases in performance: Accuracy = 8%, Precision = 18.42%, Recall = 17.65% and AUC = 0.057%.
5. The development process Based on Black Box testing, the application that was built was running well and as expected and The results of this study obtained AUC of 0.876, accuracy of 87.88%, precision of 61.90%, and recall of 76.47%, and classified into good classification.

This study has not been able to provide good t-Test test results because there is no significant difference in the t-Test test results. Based on the findings of this study, it is suggested that further researchers can use chi square, log likelihood ratio or others to select the most influential attribute in order to get good t-test results on the Naïve Bayes algorithm.

REFERENCES

[1] R. Parvizi and M. A. Adibi, "Assessing and Validating Bank Customers Using Data Mining Algorithms for Loan Home," *Int. J. Ind. Eng. Oper. Res.*, vol. 2, no. 1, 2020.

[2] L. N. Rani, "Klasifikasi Nasabah Menggunakan Algoritma C4.5 Sebagai Dasar Pemberian Kredit," *INOVTEK Polbeng - Seri Inform.*, vol. 1, no. 2, p. 126, 2016, doi: 10.35314/isi.v1i2.131.

[3] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," *Heliyon*, vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[4] A. Raorane and R. V Kulkarni, "Data Mining Techniques: A Source for Consumer Behavior Analysis," *Int. J. Database Manag. Syst.*, vol. 3, no. 3, pp. 45–56, 2011, doi: 10.5121/ijdms.2011.3304.

[5] A. U. Khasanah and Harwati, "A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 215, no. 1, 2017, doi: 10.1088/1757-899X/215/1/012036.

[6] M. Ala'raj, M. F. Abbod, and M. Majdalawieh, "Modelling customers credit card behaviour using bidirectional LSTM neural networks," *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00461-7.

[7] M. Sudhakar, C. V. K. Reddy, and A. Pradesh, "TWO STEP CREDIT RISK ASSESMENT MODEL FOR RETAIL BANK LOAN APPLICATIONS USING DECISION TREE DATA MINING TECHNIQUE Research Scholar , D epartment of Computer Science and Technology Professor , D epartment of Physics , Rayalaseema University Kurnool , Andhra," vol. 5, no. 3, 2016.

[8] P. M. Addo, D. Guegan, and B. Hassani, "Credit risk analysis using machine and deep learning models," *Risks*, vol. 6, no. 2, pp. 1–20, 2018, doi: 10.3390/risks6020038.

[9] J. Shi and B. Xu, "Credit Scoring by Fuzzy Support Vector Machines with a Novel Membership Function," *J. Risk Financ. Manag.*, vol. 9, no. 4, p. 13, 2016, doi: 10.3390/jrfm9040013.

[10] A. Krichene, "Using a naive Bayesian classifier methodology for loan risk assessment: Evidence from a Tunisian commercial bank," *J. Econ. Financ. Adm. Sci.*, vol. 22, no. 42, pp. 3–24, 2017, doi: 10.1108/JEFAS-02-2017-0039.

[11] Jamaluddin and R. Siringoringo, "Improved Fuzzy K-Nearest Neighbor Using Modified Particle Swarm Optimization," *J. Phys. Conf. Ser.*, vol. 930, no. 1, 2017, doi: 10.1088/1742-6596/930/1/012024.

[12] F. Harahap, A. Y. N. Harahap, E. Ekadiansyah, R. N. Sari, R. Adawiyah, and C. B. Harahap, "Implementation of Naïve Bayes Classification Method for Predicting Purchase," *2018 6th Int. Conf. Cyber IT Serv. Manag. CITSM 2018*, no. April, 2019, doi: 10.1109/CITSM.2018.8674324.

[13] H. Muhamad, C. A. Prasojo, N. A. Sugianto, L. Surtiningsih, and I. Cholissodin, "Optimasi Naïve Bayes Classifier Dengan Menggunakan Particle Swarm Optimization Pada Data Iris," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 3, p. 180, 2017, doi: 10.25126/jtiik.201743251.

[14] A. Harris and A. E. Mintaria, "Komparasi

e-ISSN: 2614-8404
p-ISSN:2776-3234

Information Gain , Gain Ratio , CFs-Bestfirst dan CFs-PSO Search Terhadap Performa Deteksi Anomali," vol. 5, pp. 332–343, 2021, doi: 10.30865/mib.v5i1.2258.

[15] D. Dwihandayani, "Analisis Kinerja Non Performing Loan (Npl) Perbankan Di Indonesia Dan Faktor-Faktor Yang Mempengaruhi Npl," *J. Ilm. Ekon. Bisnis*, vol. 22, no. 3, p. 228985, 2017.

[16] U. Al Qoroni, "PROFITABILITAS (Studi pada PT . Federal International Finance Rangkasbitung)," vol. 26, no. 1, pp. 1–5, 2015.

[17] S. Rusnaini, H.- Hamirul, and A. M, "Non Performing Loan (Npl) Dan Return on Asset (Roa) Di Koperasi Nusantara Muara Bungo," *J. Ilm. Manajemen, Ekon. Akunt.*, vol. 3, no. 1, pp. 1–18, 2019, doi: 10.31955/mea.vol3.iss1.pp1-18.

[18] M. Agustiningtyas, "Analisis Faktor-Faktor Yang Mempengaruhi Non Performing Loans Kredit Pada Bank Umum di Indonesia," vol. 1, no. September, pp. 120–133, 2018.

[19] T. T. Muryono and I. Irwansyah, "Implementasi Data Mining Untuk Menentukan Kelayakan Pemberian Kredit Dengan Menggunakan Algoritma K-Nearest Neighbors (K-Nn)," *Infotech J. Technol. Inf.*, vol. 6, no. 1, pp. 43–48, 2020, doi: 10.37365/jti.v6i1.78.

[20] S. Agarwal, *Data mining: Data mining concepts and techniques*. 2014.

[21] E. Knowledge, "Data Mining : Extracting Knowledge from Data."

[22] J. Sinaga and B. Sinaga, "Data Mining Classification Of Filing Credit Customers Without Collateral With K-Nearest Neighbor Algorithm (Case study: PT. BPR Diori Double)," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 2, no. 2, pp. 204–210, 2020, doi: 10.47709/cnapc.v2i2.401.

[23] M. Otivation and I. Ntroduction, "M Achine L Earning and D Ata M Ining," no. September, pp. 1–21, 2012, doi: 10.13140/RG.2.2.20395.49446/1.

[24] A. J. Chamatkar and P. Butey, "Importance of Data Mining with Different Types of Data Applications and Challenging Areas," *J. Eng. Res. Appl. www.ijera.com*, vol. 4, no. 5, pp. 38–41, 2014.

[25] Konacaklı Enis and KARAARSLAN ENİS, "Artificial Intelligence and Applied Mathematics in Engineering Problems," *Artif. Intell. Appl. Math. Eng. Probl. - Proc. Int. Conf. Artif. Intell. Appl. Math. Eng. (ICAIAME 2019)*, vol. 43, no. January, 2020, doi: 10.1007/978-3-030-36178-5.

[26] S. Karthika and N. Sairam, "A Naïve Bayesian classifier for educational qualification," *Indian J. Sci. Technol.*, vol. 8, no. 16, 2015, doi: 10.17485/ijst/2015/v8i16/62055.

[27] A. P. Wibawa *et al.*, "Naïve Bayes Classifier for Journal Quartile Classification," *Int. J. Recent Contrib. from Eng. Sci. IT*, vol. 7, no. 2, p. 91, 2019, doi: 10.3991/ijes.v7i2.10659.

[28] A. W. Syaputri, E. Irwandi, and M. Mustakim, "Naïve Bayes Algorithm for Classification of Student Major's Specialization," *J. Intell. Comput. Heal. Informatics*, vol. 1, no. 1, p. 17, 2020, doi: 10.26714/jichi.v1i1.5570.

[29] S. Chormunge and S. Jena, "Efficient feature subset selection algorithm for high dimensional data," *Int. J. Electr. Comput. Eng.*, vol. 6, no. 4, pp. 1880–1888, 2016, doi: 10.11591/ijece.v6i4.9800.

[30] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, "Variable selection for Naïve Bayes classification," *Comput. Oper. Res.*, vol. 135, p. 105456, 2021, doi: 10.1016/j.cor.2021.105456.

[31] H. Sulistiani and A. Tjahyanto, "Comparative Analysis of Feature Selection Method to Predict Customer Loyalty," *IPTEK J. Eng.*, vol. 3, no. 1, p. 1, 2017, doi: 10.12962/joe.v3i1.2257.

[32] N. A. Shaltout, M. El-Hefnawi, A. Rafea, and A. Moustafa, "Information gain as a feature selection method for the efficient classification of influenza based on viral hosts," *Lect. Notes Eng. Comput. Sci.*, vol. 1, no. July, pp. 625–631, 2014.

[33] A. A. Prasetyo and B. Kristianto, "Integration of Iterative Dichotomizer 3 and Boosted Decision Tree to Form Credit Scoring Profile," *Sisforma*, vol. 7, no. 2, p. 58, 2020, doi: 10.24167/sisforma.v7i2.2659.

[34] Dr.J.Arunadevi, S.Ramya, and M. R. Raja, "A study of classification algorithms using Rapidminer," *Int. J. Pure Appl. Math.*, vol. Volume 119, no. 12, pp. 15977–15988, 2018.