International Journal of Information System & Technology
Akreditasi No. 158/E/KPT/2021 | Vol. 6, No. 5, (2023), pp. 590-602

Data Mining to Predict Gojek's Consumer Satisfaction Level Using Naive Bayes Algorithm

Rudika Rahman[1], Felix Andreas Sutanto[2]*
[1.2] Informatics Engineering, Universitas Stikubank, Semarang, Indonesia
E-mail: dikarudika@gmail.com

Abstract

Gojek is an application that is very popular and in demand as a means of transportation because it is practical and fast. Consumer satisfaction is where the expectations, desires and needs of consumers are met. To assess whether the company provides quality service to consumers, it is necessary to evaluate consumers to determine the level of consumer satisfaction when using the Gojek application. This study aims to build a system for predicting satisfaction levels from Gojek Driver services to consumers using the Naive Bayes algorithm, as well as to determine the level of accuracy in classifying customer satisfaction using Gojek services. questionnaire is the method used in collecting data on Gojek consumer satisfaction. In this study, 120 questionnaires were distributed to respondents, namely Gojek service users, and these questionnaires would later become training data. Researchers use the survey method as a direct observation of the process of using Gojek services to identify the services provided to consumers. Researchers use the waterfall method as a system development model. This model is the oldest software development model paradigm, and the most widely used. The process of calculating the accuracy of the system uses the Naive Bayes method by testing based on training data taken from the questionnaire. The calculation results on the level of accuracy obtained from the training data are equal to 88.9%. The calculation is processed and divided by the system as much as 70% training data and 30% testing data or as many as 84 training data and 36 testing data. This consumer satisfaction prediction system can assist an admin in determining the classification of consumer satisfaction with web-based Gojek services by applying the Naive Bayes method. In this study, researchers only calculated the level of accuracy and predictive value, for further research it is hoped that they can try to calculate the precision value and recall value calculations.

Keywords: *Gojek, Naive Bayes ,Classification, Data Mining*

1. Introduction

The development of technology today occurs very rapidly and provides social changes for society. Many businesses have emerged by utilizing this communication technology, including the emergence of a business providing online motorcycle taxi transportation services, namely Gojek [1]. Gojek is a very popular application and is in demand as a means of transportation because it is practical and fast. Consumersatisfaction is where the expectations, desires and needs of consumers are met. Each consumer compares the expected performance with the received performance. Customer satisfaction is very important for companies to improve their business because it can be translated positively in the form of increased profits and positive evaluation of the services provided [2]. To assess whether the company provides quality service to consumers, it is necessary to evaluate from consumers to find out the level of consumer satisfaction when using the Gojek application. Judging from previous research, there are cases related to using the Naive Bayes method. The previous research used the Naive Bayes method with the title "Prediction of satisfaction levels in online learning using the Naive Bayes algorithm" from as many as 110 respondents' data with a questionnaire-shaped data collection method, there were 80 training data and 30 testing data and obtained an accuracy rate of 100% [3]. This study aims to build a system for predicting the level of satisfaction from

ISSN : 2580-7250

Gojek Driver services to consumers using the Naive Bayes algorithm, as well as to determine the level of accuracy in classifying customer satisfaction using Gojek services. Based on the description above, researchers are interested in conducting research by building a system with the Naive Bayes method, where the system can help in improving the quality of Gojek services and can become a trustworthy shuttle service.

2. Research Methodology

2.1. Data Mining

Data mining is defined as the process of finding patterns in data. This process is automatic or often semi-automatic. The pattern found must be meaningful and the pattern provides an advantage, usually an economic advantage. Large amounts of data are needed [4]. Data mining is a series of processes of extracting added value in the form of information that has not been manually known from a database. Data mining is mainly used to search for knowledge contained in large databases so it is often called the Knowledge Discovery Database (KDD) [5]. The Knowledge Discovery Database (KDD) is a knowledge-seeking process that benefits from a data set. The KDD process is interactive and iterative, includes a number of steps involving use in making decisions and can be repeated between two steps. Data mining is one of the core processes contained in the Knowledge Discovery Database (KDD). Many people treat data mining as a synonym for KDD, as most of the work in KDD is focused on data mining [6].

2.2. Prediction

Prediction is a process of systematically estimating what is most likely to happen in the future based on information possessed in the past and present, so that the error (the difference between what happened and the forecast result) can be reduced. Predictions do notnecessarily givea definitive answer to what happened, instead predictions try to get an answer that is as close as possible [7].

2.3. Data Mining Engineering

Some of the techniques and properties of data mining are as follows:

a) Clustering, (data clustering) considers an important approach to finding similarities in data and placing the same data into groups. Clustering divides data sets into groups where the similarity in a group is greater than among groups [8].
b) Regression, is to predict the value of a given continuous variable based on the values of another variable, consuming a linear and nonlinear dependency model.
c) Classification, is the evaluation of data objects to assign them to a certain category with the number of categories that have been available. Classification creates a model based on existing training data and uses the model to classify new data. Classification can be defined as work that performs training/learning from each set of goal functions that will map attributes (features) to the number of classes available [9].
d) Association rule, is to detect a collection of attributes that appear together (co-occur) in frequent frequency and form a number of rules from those groups.

2.4. Naive Bayes

The Naïve Bayes Bayesian classification algorithm is a statistical classification that can be used to predict the probability of membership of a class. Bayesian classification is based on the Bayes Theorem which has classification capabilities similar to decesion tree and neural networks. Bayesian classificationter is proven to have a high level of accuracy and speed that can be applied to databases with big data [10].

2.5. Customer Satisfaction

Customer satisfaction is one of the most important factors of improving the company's marketing efficiency. Customer satisfaction can increase the intensity of those customers

purchases . Creating a level of satisfaction optimal pelanggan then encourages the creation of loyalty earlier in the mind satisfied customers. Satisfaction is measured by how well customer expectations are met. Although customer loyalty is a measure does the customer want to make another purchase[11].

2.6. Visual Studio Code

Visual Studio Code is an open source code editor application developed by Microsoft for Windows, Linux, and MacOS operating systems. Visual Code makes it easy to write code that supports several types of programming, such as C++, C#, Java, Python, PHP, GO. Visual Code can detect the type of programming language used and provide color variations based on the capabilities of the coder. Visual Studio Code has also been integrated into Github. In addition, another feature is the ability to add plugins, where developers can add plugins to add functionality not included in Visual Studio Code[12].

2.7. XAMPP

XAMPP is software designed to run PHP-based websites and process MySQL data on a local computer. XAMPP runs as a web server on the local computer. XAMPP is also known as CPanel virtual server, which can help with preview, allowing websites to be modified without having to go online or have access to the internet[13].

2.8. Data Collection Techniques

2.8.1. Questionnaire

Questionnaire is a method used in collecting Gojek consumer satisfaction data. In this study, 120 questionnaires were distributed to respondents, namely Gojek service users, and the questionnaire will later become training data.

2.8.2. Survey

This method is carried out by conducting a survey by observing the object directly where the object certainly supports and is related to the research. With this survey method, the author will observe the process of using Gojek services directly to identify the services provided to consumers.

2.8.3. Literature Study

The literature study in this research is to study reference books or sources related to this research, both from books and the internet.

2.9. System Design Method

Researchers use the waterfall method as a model for system development. This model is the oldest, and most widely used software development model paradigm.

2.10. Design Stage

a) Use Case Diagram

A use case is a form of diagram that illustrates the expected functionality of a system viewed from the perspective of a user outside the system. Use cases can also be used to represent interactions that occur between actors and system processes created.

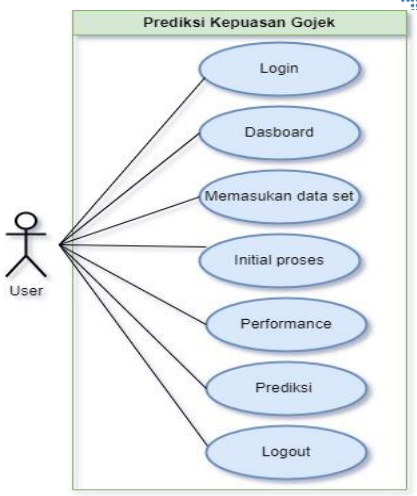

Figure 1. Use Case Diagram

b) Activity Diagram

The activity diagram describes the workflow or activity of a system, but not the activity of an actor. The activity diagram also illustrates how the system flow starts, the possible choices, and how the system flow ends.

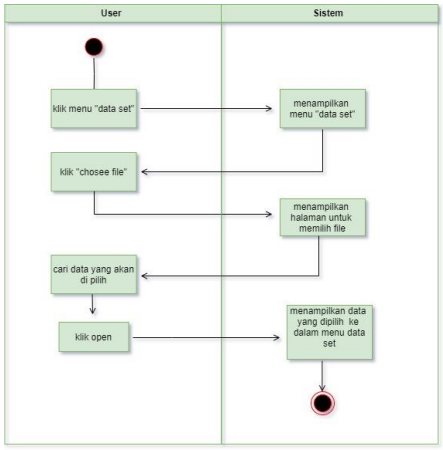

Figure 2. Activity Diagram

c) Interface Design

Interface design is carried out to design the system menu structure so that it can match the menu on the system to be implemented.

Figure 3. Login

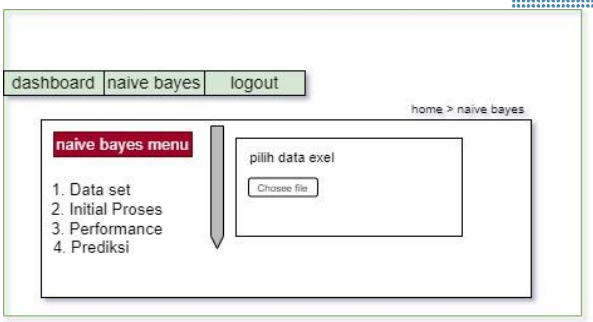

Figure 4. Proses Page

d) Data Preparation

The dataset to be used has 1 variable as a class, namely the Gojek application status of "satisfied" and "dissatisfied" and 5 variables as attributes.

Tabel 1. Variables and Categories

Variable	Information	Scale	Category
X1	Application	Real	1. (Strongly disagree) 2. (Disagree) 3. (Agree) 4. (Strongly agree)
X2	Timeliness	Real	1. (Strongly disagree) 2. (Disagree) 3. (Agree) 4. (Strongly agree)
X3	Ride comfort	Real	1. (Strongly disagree) 2. (Disagree) 3. (Agree) 4. (Strongly agree)
X4	Hospitality	Real	1. (Strongly disagree) 2. (Disagree) 3. (Agree) 4. (Strongly agree)
X5	Price	Real	1. (Strongly disagree) 2. (Disagree) 3. (Agree) 4. (Strongly agree)

e) Use of Naive Bayes

Accuracy is defined as the degree of proximity between the predicted value and the actual value. Measurement of accuracy to the model using a confusion matrix that focuses on its class. A confusion matrix is an array for recording the results of classification work. At this stage the Confusion matrix tests the data then finds the best level of accuracy with the model used.

Tabel 2. Accuracy Calculation

Clasification Naive Bayes	Satisfied (+)	Disgruntled (-)
Satisfied (+)	*True positives* (TP)	*False negatives* (FN)
Disgruntled (-)	*False positives* (FS)	*True negatives* (TN)

This naïve bayes model has a high degree of accuracy when implemented on large data sets in databases. The basic concept is the *Bayes Theorem* which is to calculate the probability to perform the classification. Confusion matrix formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FS+TN} \quad (1)$$

Information:

a) Accuracy is the proposition of the number of correct predictions.
b) TP (True positive) is the number of positive records classified as *positive* by the classifier.
c) TN (True negative) is the number of negative records classified as *negative* by the clasifier.
d) FP (False positive) is the number of negative records classified as *positive* by the classifier.
e) FN (False negative) is the number of positive records classified as *negative* by the classifier.

Naive bayes formula :

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (2)$$

Tabel 3. Formula Description

Information	
X	Data with unknown classes
Y	The data hypothesis X is a specific class
P(Y| X)	Probability hypothesis Y based on condition X
P(Y)	Probability hypothesis Y
P(X|Y)	Probability X based on the condition at the time of hypothesis Y
P(X)	Probability X

3. Result and Discusstion

3.1. Naive Bayes Calculation

The dataset that will be used as *training* data in this study is 120 data. As data *testing* data that processes as many as 36 data.

3.2. Class Probability Calculation

Tabel 4. Class Probability

Determining Class Probability			
Label	**Amount of Data**	**Total Amount of Data**	**Result**
Satisfied	98	120	0,816666667
Dissatisfied	22	120	0,183333333

3.3. Category Probability Calculation

Determining the probability in each category can be described in the following table:

Tabel 5. Application Probability

Determining Application Probability (X1)				
Application	**Satisfied**	**Unsatisfied**	**P(Satisfied)**	**P(Unsatisfied)**
Totally Agree	35	0	0,4	0
Agree	59	12	0,602040816	0,545454545
Disagree	4	7	0,040816327	0,318181818

Determining Application Probability (X1)				
Application	**Satisfied**	**Unsatisfied**	**P(Satisfied)**	**P(Unsatisfied)**
Strongly disagree	0	3	0	0,136363636
sum	98	22	1,0	1,0

Tabel 6. Timeliness Probability

Determining the Probability of Time Tag(X2)				
Timeliness	**Satisfied**	**Unsatisfied**	**P(Satisfied)**	**P(Unsatisfied)**
Totally Agree	31	0	0,3	0
Agree	61	8	0,62244898	0,363636364
Disagree	5	9	0,051020408	0,409090909
Strongly disagree	1	5	0,010204082	0,227272727
sum	98	22	1,0	1,0

Tabel 7. Probability of Driving Comfort

Determining the Probability of Driving Comfort (X3)				
Driving Comfort	**Satisfied**	**Unsatisfied**	**P(Satisfied)**	**P(Unsatisfied)**
Totally Agree	36	1	0,4	0,045454545
Agree	54	8	0,551020408	0,363636364
Disagree	7	11	0,071428571	0,5
Strongly disagree	1	2	0,010204082	0,090909091
sum	98	22	1,0	1,0

Tabel 8. Probability of Hospitality

Determining the Probability of Hospitality (X4)				
Hospitality	**Satisfied**	**Unsatisfied**	**P(Satisfied)**	**P(Unsatisfied)**
Totally Agree	32	0	0,3	0
Agree	57	6	0,581633	0,272727
Disagree	9	9	0,091837	0,409091
Strongly disagree	0	7	0	0,318182
sum	98	22	1,0	1,0

Tabel 9. Price Probability

Determining Pricing Probability (X5)				
Price	**Satisfied**	**Unsatisfied**	**P(Satisfied)**	**P(Unsatisfied)**
Totally Agree	40	0	0,4	0
Agree	53	9	0,540816	0,409091
Disagree	4	9	0,040816	0,409091
Strongly disagree	1	4	0,010204	0,181818
sum	98	22	1,0	1,0

3.4. Naive Bayes Manual Calculation

The following is a manual calculation using snippets from the available test data:

Tabel 10. Testing Data Pieces

Application	Accuracy Time	Comfort	Hospitality	Price	Predictions
Agree	Strongly Agree	Disagree	Agree	Strongly Agree	**?**

a) Prediction calculation
P (X|Prediction = satisfied)
$0{,}602040816 * 0{,}3 * 0{,}071428571 * 0{,}581633 * 0{,}4 = 0{,}003229$
P (X|Prediction = dissatisfied)
$0{,}545454545 * 0 * 0{,}5 * 0{,}272727 * 0 = 0$

b) Calculation of the probability of the predicted result with the probability class
P (X|Prediction = satisfied) P(class = satisfied)
$0{,}003229 * 0{,}816666667 = 0{,}002637$
P (X|Prediction = dissatisfied) P(class = dissatisfied)
$0 * 0{,}183333333 = 0$

Judging from the calculations above, the largest probability value is in P (X|Prediction = satisfied), so the conclusion for calculations in the testing data and will be inputted is a satisfied prediction.

3.5. Calculation of Accuracy Rate

From the results of the calculation of the data testing on the system, the values obtained are:

TP = 26, TN = 6, FN = 3, FS = 1.

Accuracy = $\frac{26+6}{26+3+1+6} = 0888889$

The accuracy calculation process in the system uses the *Naive Bayes* method by testing based on training data taken from a questionnaire of 120 respondents' data. The calculation results on the accuracy level obtained from the training data are 88.9%. The calculation is processed and divided by the system as much as 70% training data and 30% testing data or as many as 84 training data and 36 testing data.

3.6. System Display

3.6.1. Login page

This page is the first page that will appear before entering the application. By entering the user name and password that the user already has.

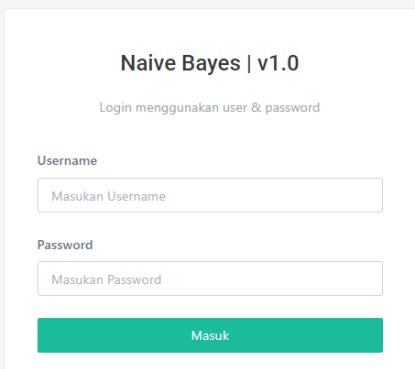

Figure 5. Login Page

3.6.2. Naive Bayes Page

This page is the main page on the application, which is to process data mining. In this page there are several menus available such as dataset menus, initial processes, performance and predictions.

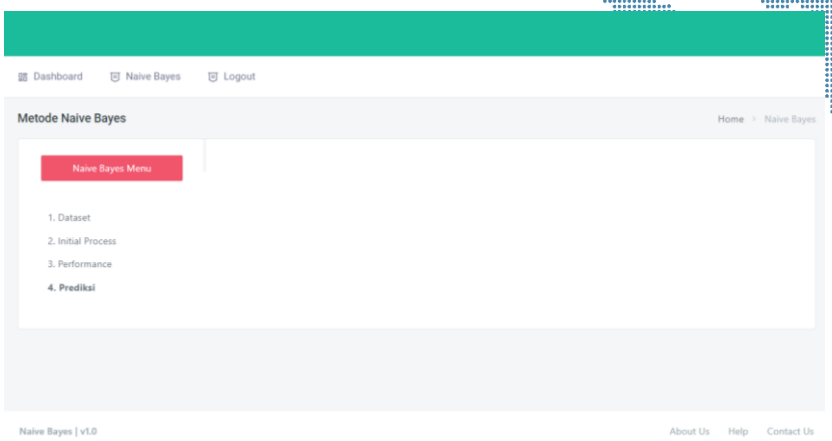

Figure 6. Naive Bayes Page

3.6.3. Dataset Page

The dataset page is a page to enter the training data to be processed and display it on this page. This page provides a menu to select file or choose file when clicked on the letter dataset.

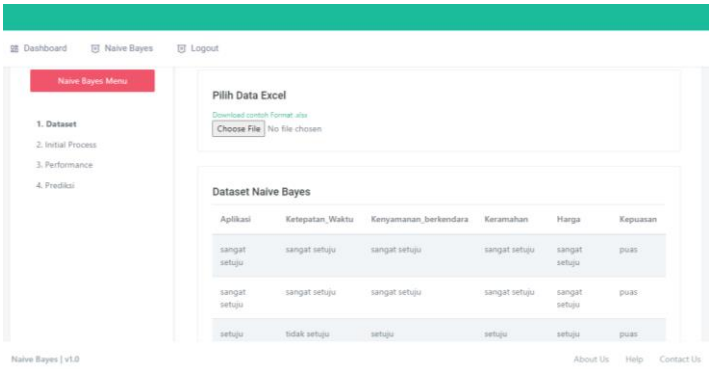

Figure 7. Dataset Page

3.6.4. Initial Proses Page

The initial process page contains datasets. The initial process then initializes the data by providing supporting attributes and target labels to the dataset as a form of format that will be processed on the next page.

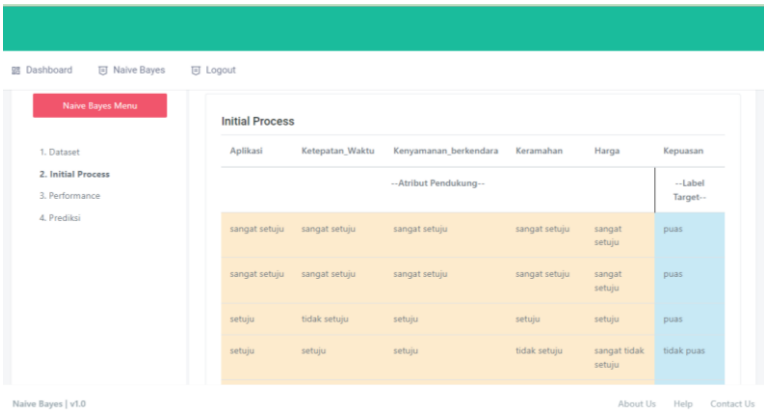

Figure 8. Initial Proses Page

3.6.5. Peformance Page

This page is a calculation page for calculating the accuracy test of the method by processing the dataset that has been selected. In this menu, there is a percentage menu for data processing ranging from 10% to 90%. For example, there is a dataset of 100 data and choose a percentage of 70%, then the data to be processed is data of 70 training data and 30 testing data.

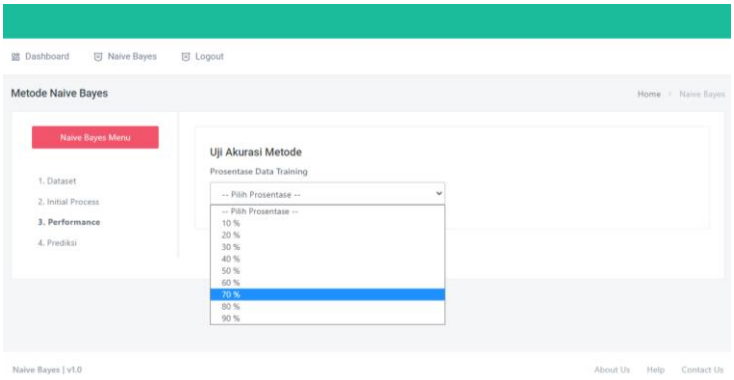

Figure 9. Performance Page

After clicking the 70% percentage, the dataset will separate training data and testing data.

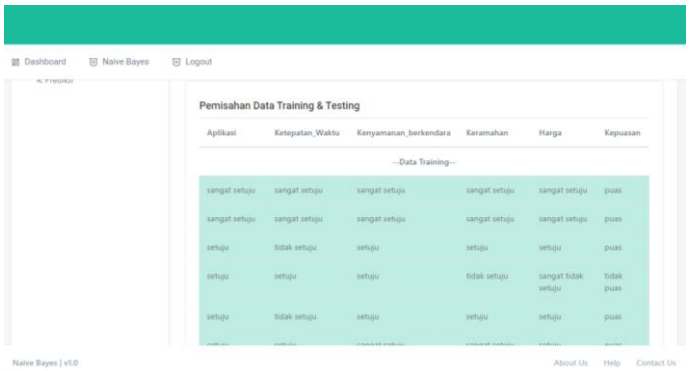

Figure 10. Training Data

After the data is divided, it can be tested data and will be processed.

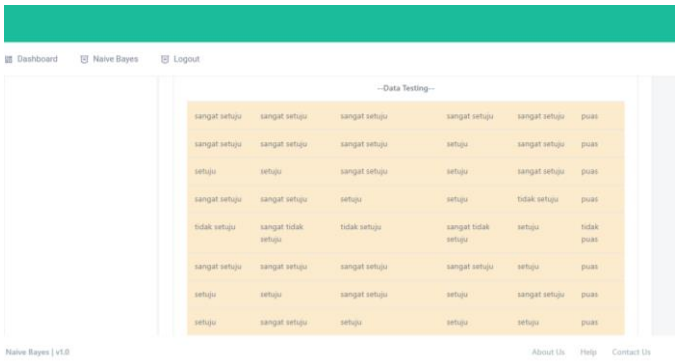

Figure 11. Testing Data

Data Testing Figure 12 is a display of the testing data process which contains the data of the respondents to be tested. In the data testing process, the data will get the results of the previous data testing process in figure 11.

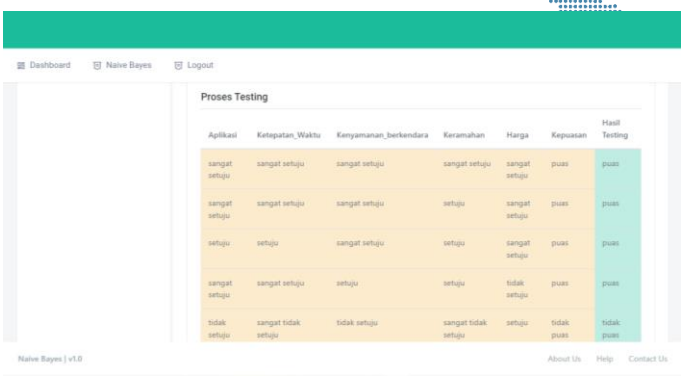

Figure 12. Data Testing Proses

After the data was tested, the data obtained an accuracy result of 88.9% can be seen in figure 13.

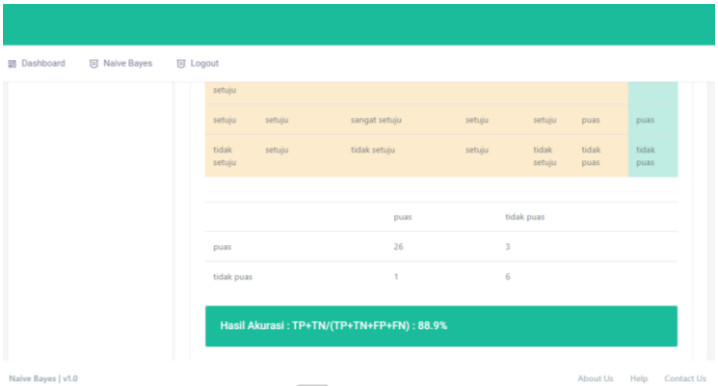

Figure 13. Accuracy Results

3.6.6. Prediction Page

The prediction page contains attribuats according to the attributes in the dataset selected earlier. In accordance with the dataset, there are application attributes, punctuality, driving comfort, friendliness and price. There are 4 variables used, namely strongly agree, agree, disagree, strongly disagree. In figure 14 the user inputs the prediction form. Figure 15 shows the results of the naïve bayes prediction calculation from the user's input.

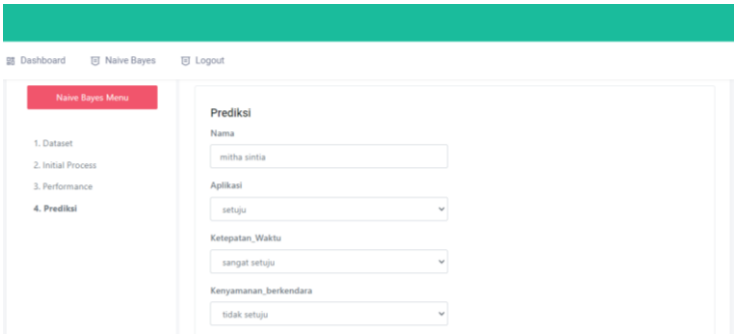

Figure 14. Prediction Variables

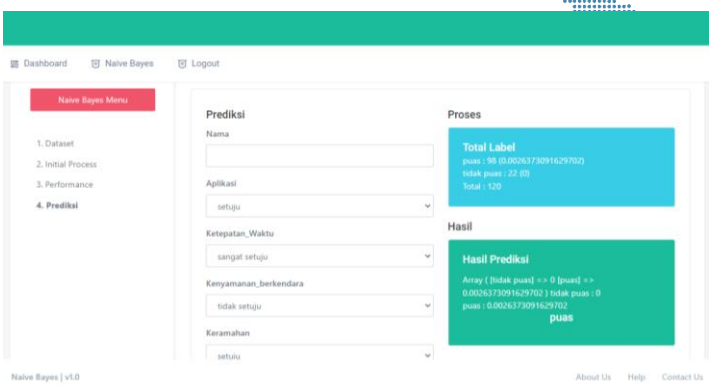

Figure 15. Predicted Results

4. Conclusion

Conclusion Based on the results of the research conducted, researchers can be drawn conclusions as A consumer satisfaction prediction system can help an admin in determining the classification of customer satisfaction with Gojek's web-based services by applying the Naive Bayes method. The Naive Bayes method by utilizing training data obtains the result of classifying the probability value of the "satisfied" class greater than the probability value of the "dissatisfied" class. After testing using the application that has been created, an accuracy rate of 88.9% was obtained.

In this study the researcher only calculated at the level of accuracy value and prediction value, for subsequent studies it is expected to try to calculate on the calculation of precission value and recall value.Researchers expect from this study to be able to apply using other algorithm methods, in order to be able to develop different applications in the next research.

References

[1] F. Setiawan, S. W. K. Dewi, and Musafa, "Pengaruh kualitas pelayanan dan harga terhadap kepuasan pelanggan Gojek Di Kota Bandung," J. Econ. Bus. UBS, vol. 8, no. 1, pp. 1–17, 2022, doi: 10.52644/joeb.v8i1.13.

[2] F. Martiningsih, "Sistem evaluasi kepuasaan pelanggan go-jek menggunakan metode naïve bayes," p. 20, 2018.

[3] A. R. Damanik, S. Sumijan, and G. W. Nurcahyo, "Prediksi tingkat kepuasan dalam pembelajaran daring menggunakan algoritma naïve bayes," J. Sistim Inf. dan Teknol., vol. 3, pp. 88–94, 2021, doi: 10.37034/jsisfotek.v3i3.49.

[4] B. D. Meilani and N. Susanti, "Akurasi data mining untuk menghasilkan pola kelulusan mahasiswa dengan metode naive bayes," J. Sist. Inf. Univ. Suryadarma, vol. 3, no. 2, pp. 182–189, 2014, doi: 10.35968/jsi.v3i2.66.

[5] N. D. Sari, "Penerapan kelasifikasi kepuasan pelanggan go-jek menggunakan metode algoritma naïve bayes," p. 60, 2018.

[6] S. L. B. Ginting and R. P. Trinanda, "Teknik data mining menggunakan metode bayes classifier untuk optimalisasi pencarian aplikasi perpustakaan," J. Tek. Komput., vol. 4, no. 2, pp. 17–20, 2016.

[7] D. S. O. Panggabean, E. Buulolo, and N. Silalahi, "Penerapan data mining untuk memprediksi pemesanan bibit pohon dengan regresi linear berganda," JURIKOM (Jurnal Ris. Komputer), vol. 7, no. 1, p. 56, 2020, doi: 10.30865/jurikom.v7i1.1947.

[8] S. M. Sinaga, J. T. Hardinata, and M. Fauzan, "Implementasi data mining clustering tingkat kepuasan konsumen terhadap pelayanan Go-Jek," Kesatria J. Penerapan Sist. Inf. (Komputer dan Manajemen), vol. 2, no. 2, pp. 118–124, 2021.

[9] D. P. Utomo and M. Mesran, "Analisis komparasi metode klasifikasi data mining dan reduksi atribut pada data set penyakit jantung," J. Media Inform. Budidarma,

vol. 4, no. 2, p. 437, 2020, doi: 10.30865/mib.v4i2.2080.

[10] H. D. Wijaya and S. Dwiasnati, "Implementasi data mining dengan algoritma naïve bayes pada penjualan obat," J. Inform., vol. 7, no. 1, pp. 1–7, 2020, doi: 10.31311/ji.v7i1.6203.

[11] P. E. Setyo, "Pengaruh kualitas produk dan harga terhadap kepuasan konsumen 'Best Autoworks,'" PerformA J. Manaj. dan Start-Up Bisnis, vol. 1, no. 6, pp. 755–764, 2017.

[12] N. A. Ramdhan and D. A. Nufriana, "Rancang bangun dan implementasi sistem informasi skripsi oline berbasis WEB," J. Ilm. Intech Inf. Technol. J. UMUS, vol. 1, no. 02, pp. 1–12, 2019, doi: 10.46772/intech.v1i02.75.

[13] E. W. Fridayanthie and T. Mahdiati, "Rancang bangun sistem informasi permintaan ATK berbasis intranet (studi kasus: KEJAKSAAN NEGERI RANGKASBITUNG)," J. khatulistiwa Inform., vol. 4, no. 02, pp. 1–138, 2016, [Online]. Available: http://dspace.unitru.edu.pe/bitstream/handle/UNITRU/10947/Miñano Guevara%2C Karen Anali.pdf?sequence=1&isAllowed=y%0Ahttps://repository.upb.edu.co/bitstream/handle/20.500.11912/3346/DIVERSIDAD DE MACROINVERTEBRADOS ACUÁTICOS Y SU.pdf?sequence=1&isAllowed=