JAIA – Journal Of Artificial Intelligence And Applications
Journal Homepage : http://jurnal.sar.ac.id/index.php/JAIA/

Product Classification Based on Categories and Customer Interests on the Shopee Marketplace Using the Naïve Bayes Method

Muhammad Oase Ansharullah[1], Wirta Agustin [2], Lusiana[3], Junadhi[4], Susi Erlinda[5], Fransiskus Zoromi [6]
[1,2,3,4,5,6]Program Studi Teknik Informatika STMIK Amik Riau, Pekanbaru

Article Info

Received : Jan 20, 2022
Revised : Feb 20, 2022
Accepted : Apr 10, 2022

Keywords :
Marketplace,
classification,
Naïve Bayes,
Shopee,
Weka.

Abstract

Marketplace is an electronic product marketing platform that brings together many sellers and buyers to transact with each other. The large variety of products sold on Shopee is one of the reasons this application is in great demand by all walks of life. However, the weakness of the large variety of products sold in a marketplace causes buyers who have no potential to buy these products. To overcome this problem, it is necessary to do a classification to determine which products are most in demand by customers. Product categories consist of: Clothing, Beauty Products, Daily Goods, Electronics, and Accessories. The classification method used is Naïve Bayes and the software used is WEKA. The next data collection is done by distributing questionnaires to the existing customers on social media namely, Whatsapp and Instagram, the distribution of the questionnaire is conducted through Google form. There are 90 questionnaires that will be distributed in this study. Some of the indicators asked in the questionnaire namely, do you like shopping online? And what marketplaces are commonly used. These results will be the training data. Interest categories are divided into 4 categories, namely: Very interested, Interested, Not interested, Very not interested. The results obtained in this study are clothing products (72 respondents) are products that are in great demand, daily goods products (7 respondents) are products of interest, beauty and electronic products (5 respondents) are products that are not in demand, and accessories (1 respondents) is a product that is not very attractive to customers on the Shopee marketplace

Susi Erlinda
Email: susierlinda@stmik-amik-riau.ac.id

1. Introduction

Marketplace is an electronic product marketing platform that brings together many sellers and buyers to transact with each other [1]. The thing that underlies the formation of the marketplace is because of the large number of internet users from all walks of life who create innovations in conducting business processes using the internet. So, everyone can carry out buying and selling activities easily, quickly and cheaply because there are no limits on space, distance and time.

Marketplace has more or less the same concept as traditional markets. Basically, the marketplace owner is not responsible for the goods being sold because their job is to provide a place for sellers who want to sell and help them to meet customers and make transactions more simply and easily. The transaction is indeed regulated by the marketplace then after making the payment the seller will send the goods to the buyer. One of the reasons why the marketplace is famous is because of the ease and convenience of use. Many describe the marketplace as a department store[2] . Based on the calculation of statistical data from iprice, the most visited marketplace in Indonesia in the fourth quarter of 2020 was Shopee with 129,320,800 clicks per month.

The large variety of products sold on Shopee is one of the reasons this application is in great demand by all walks of life. However, the weakness of the large variety of products sold in a marketplace This causes buyers who do not have the potential to buy the product. Therefore, to increase sales results, appropriate marketing strategies are needed so that sellers can determine which products are of interest to customers.

The marketing strategy that can be done is by classifying customers to determine which products are most interested by customers by distributing questionnaires to a group of people who use the Shopee marketplace. Then, the collected data will be classified to see customer tendencies towards a product in the marketplace. According to Widaningsih & Suheri [3], classification is used to group a category into certain predetermined groups. There are many methods that can be used to classify data, including Support Vector Machine, K-Nearest Neighbor, and Naive Bayes. In this study, the method that will be used is Naïve Bayes. Naïve Bayes is a type of classification method based on Bayes' theory, namely making a prediction in the future based on statistical results from evidence and previous data that have been collected [4] and based on simple probabilities based on the application of Bayes' theorem with a strong assumption of independence. In other words, in Naïve Bayes using the independent feature model, a strong independent intention on features is that the data is not related to other data in the same case or other attributes [5] . This can be proven by the results of research by Devita et al, [6], the use of the Naïve Bayes method is considered better than other classification methods, because it can produce maximum accuracy by using less data, and shows a high level of accuracy, namely as much as 70%.

The purpose of this study is to classify customer interests to the shopee marketplace that has potential and does not have the potential for certain products by distributing questionnaires of products of interest to a group of people who use the marketplace. Then the data will be classified and processed using WEKA software with the Naïve Bayes method to classify customer interest in buying a product based on the data that has been collected

2. Research Method

The flow of the research carried out can be seen in the following diagram:

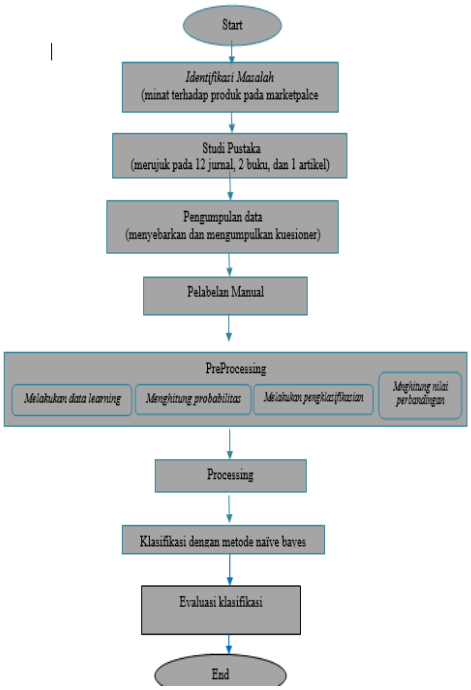

Figure 1. Research Flow

2.1. Study and Data Collection

This research was started by conducting a case study related to classification, Naïve Bayes and WEKA software. Then, data collection is done in two ways, namely

a. Observation, researchers make direct observations of. which product categories in the marketplace are the customers most interested in. The data that will be used in this research is data from Shopee marketplace users.

b. Dissemination of questionnaires, the distribution of this questionnaire is done through the google form. There are 90 questionnaires that will be distributed in this study. Some of the indicators asked in the questionnaire include: Do you like shopping online? and commonly used marketplaces. These results will be the data training

2.2. Manual Labeling

Manual labeling is carried out on respondent data collected through the google form. In this manual labeling, normalization of the data is carried out, the normalization meant by the author is to carry out the data justification process in accordance with the data that can be processed by WEKA 3.9.5. The data obtained through the distribution of questionnaires are then collected so that the manual labeling process can immediately be carried out by normalizing the data in accordance with the provisions.

2.3. Pre-Processing

The processes that run in this stage are tokenization, stopword deletion, stemming, and calculating the weights for each word in the document collection (input data). After this stage is complete, then proceed with the processing stage, namely data normalization

2.4. Classification by Naïve Bayes Method

Naïve Bayes calculation is done by determining attribute data, attribute value data, training data, test data, determining probability and determining the calculation of data.test

2.5. Classification Evaluation

Cross Validation is a method that can be used to evaluate the predictive performance of the model. The data is usually divided into two parts and based on this separation in one part, the training is carried out while the predictive is tested in the other part In general.

3. Results and Discussion

3.1. Results

The results of this study were tested using WEKA software version 3.9.5. with 64 bits. The software requirements needed are as follows.

a. Software (software)

Software to support WEKA 3.9.5 performance. at least use some software (software) used in testing the data are:

1. Minimum operating system Microsoft Windows 8 64 bit
2. Google Chrome version > 52
3. Mozilla Firefox
4. Visual Code Version 1.59.0
5. Microsoft Excel 2019

a. Hardware (hardware)

Hardware (hardware). The hardware used to run WEKA 3.9.5. this minimum is:

1. *Prosessor Intel Core i3-5200U @2.53 GHz*
2. *Hardisk Seagate 500 GB*
3. *Memory 4096 MB RAM SSD*

3.1.1. Data collection

Data collection in this study began with distributing questionnaires to the public through social media (Whatsapp and

Instagram) using Google forms. The link for the questionnaire is: *https://docs.google.com/forms/d/1fVlh5wLd9SODSDIQdJdxWKysyFtQ5ro0nN0rIqkFugQ/edit*

In the questionnaire, several questions were asked to be filled out by the community, so that the results of the data would be processed and grouped. The questions asked in the questionnaire are as follows:

a. Gender
b. Age
c. Ever shopped online
d. Marketplace used
e. Purchased product

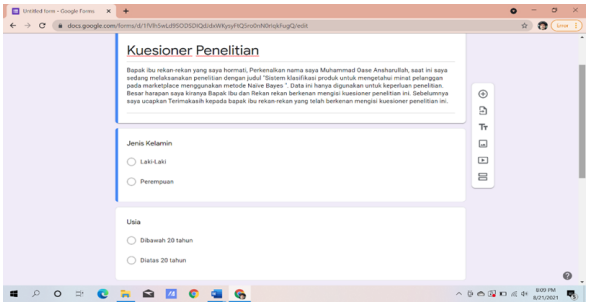

Figure 2: Research questionnaire form

Results of distributing the questionnaire obtained 90 respondents, with the number of male respondents being more than female with a value of 55.7% and 44.4%.

Attribute	Class Pakaian (0.55)	Produk Kecantikan (0.14)	Barang Harian (0.17)	Elektronik (0.13)	Aksesoris (0.02)
No					
mean	45.7843	35.1667	60.4	38	14
std. dev.	23.2913	23.1942	29.0168	26.5672	0.1667
weight sum	51	12	15	11	1
precision	1	1	1	1	1
Jenis Kelamin					
Laki-Laki	35.0	2.0	6.0	11.0	2.0
Perempuan	18.0	12.0	11.0	2.0	1.0
[total]	53.0	14.0	17.0	13.0	3.0
Usia					
Diatas 20 tahun	43.0	9.0	11.0	10.0	2.0
Dibawah 20 tahun	10.0	5.0	6.0	3.0	1.0
[total]	53.0	14.0	17.0	13.0	3.0

Figure 3. Number of Respondents by Age and Interest in Online Shopping

Figure 3. Shows the age of respondents who filled out the questionnaire above 20 years more, namely 77.5% with all respondents who stated that they had shopped online After the results are obtained, the initial data display will be obtained, namely data in the form of excel files and separate data files per file.

Based on transactions per month, transactions that have been successfully collected starting from August 13, 2021 to August 24, 2021, with details of the total capacity of the 12 kb file and the number of rows after being combined in an excel file totaling 91 rows and has a total of 5 columns, each column consisting of No, Gender, Age, Shopping, Marketplace used, Product Category. Each column heading has a function according to the column name. The first column contains the transaction number which is useful for calculating the amount of data that has been obtained through 91 respondents. The second column is Gender which is useful as a separator between male and female respondents. The third column is the Shopping column which is useful as a marker whether the respondent has ever shopped online or not. The fourth column is the Marketplace column, this column is useful for finding out what Marketplace respondents often use in online shopping, and the fifth column is the Product Category column, this column aims to find out the product what respondents often use in online shopping. The next stage is to normalize the data by carrying out the data justification process in accordance with the data that can be processed

by WEKA 3.9.5. and from the results of combining data that has been carried out in the previous step, the purpose of normalizing this data is to justify the name of each Marketplace which is the same but the writing of the Marketplace name varies so that it can create and affect the results processed by WEKA 3.9.5. Adjustment of data in a format that is acceptable and processed by WEKA 3.9.5. This is done by changing the format, namely from an excel file which initially has an extension (xls) or (xlsx) then it will be converted into an extension (csv). This can be seen in Figure 4. what respondents often use in online shopping. The next stage is to normalize the data by carrying out the data justification process in accordance with the data that can be processed by WEKA 3.9.5. and from the results of combining data that has been carried out in the previous step, the purpose of normalizing this data is to justify the name of each Marketplace which is the same but the writing of the Marketplace name varies so that it can create and affect the results processed by WEKA 3.9.5.

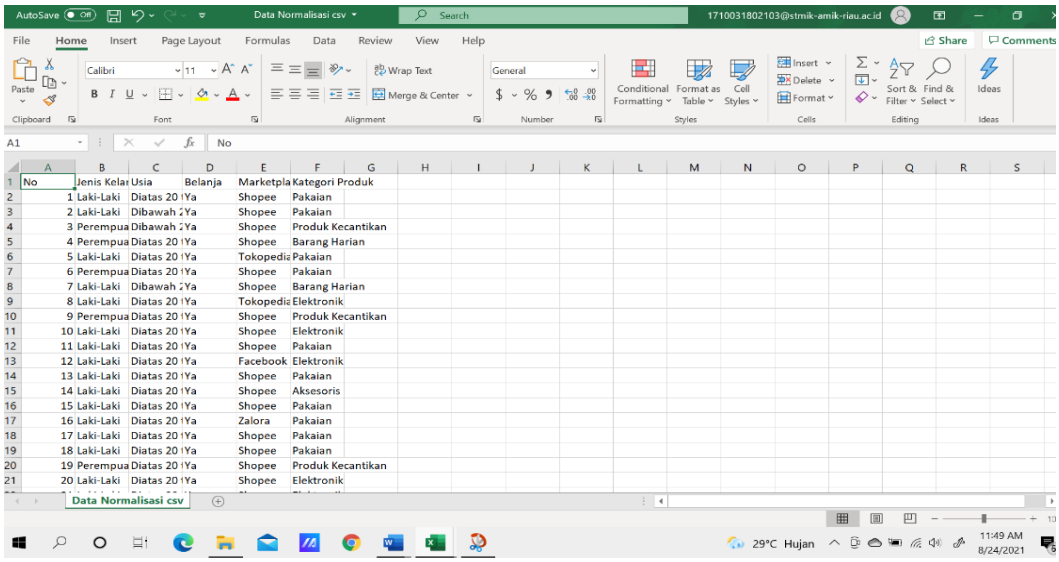

Figure 4. Converting Excel to CSV Format (MS-DOS)

3.1.2. Manual Calculation

To find the class average / class probability according to the 5 existing categories
then it can be calculated the total value of the product divided by the total number of products, then the value of clothing products 53: 90 is 0.58, the value of daily goods products 15: 90 is 0.16, the value of beauty products 12: 90 is 0.13, the value of electronic products 11: 90 is 0.12, and the value of the accessories product 1: 90 is 0.01. And to calculate the probability of each attribute from a data set that has 5 data attributes, namely: Gender, Age, Ever shopped online, Marketplace used, and Products that are often purchased. For further calculations, calculations will be carried out using the WEKA software with the Naïve Bayes method.

Table 1. Results of Classification Manual Calculations

Kategori Produk	Jumlah dari Setiap Produk	Rata-Rata Class
Pakaian	51	0,566666667
Produk Kecantikan	12	0,133333333
Barang Harian	15	0,166666667
Elektronik	11	0,122222222
Aksesoris	1	0,111111111

3.1.3. Naïve Bayes Classification.

The data is imported and classified by selecting the Classify menu in the WEKA 3.9.5 software, then choosing a classification method, namely Naïve Bayes, the next step is choosing the test method used, where in the test option there are Supplied Test, Cross Validation, and Percentage Split options, The

test method used is Cross Validation and pressing the Start button.

3.2. Discussion

This study uses WEKA 3.9.5 software, to process data from respondents who have filled out questionnaires that have been distributed by the author. Results Interest results can be seen in the image below:

Kategori Produk	Jumlah Produk
Pakaian	51
Produk Kecantikan	12
Barang Harian	15
Elektronik	11
Aksesoris	1

Figure 5. Classification results from various attributes

In Figure 5. there are five classification results with values a, b, c, d, and e, which are explained as follows: a : Clothing, b : Beauty Products, c : Daily Items,
d: Electronics and e : Accessories. For Gender Men are more likely to shop for clothes compared to women, where the interest rate for men to buy clothing products is 35.0 while women are only 18.0. However, in the attribute of beauty products, women are more interested than men, with a value of 12.0 for women and 2.0 for men. And in daily goods, women are still superior compared to men with a value of 11.0 for women and 6.0 for men. However, in the Electronic Attribute, men tend to excel with 11.0 compared to women only having a value of 2.0, and Back in the Accessories Attribute, men again excel with a value of 2.0 and women only get a value of 1.0.

Attribute	Pakaian Produk (0.55)	Kecantikan (0.14)	Barang Harian (0.17)	Elektronik (0.13)	Aksesoris (0.02)
No					
mean	45.7843	35.1667	60.4	38	14
std. dev.	23.2913	23.1942	29.0168	26.5672	0.1667
weight sum	51	12	15	11	1
precision	1	1	1	1	1
Jenis Kelamin					
Laki-Laki	35.0	2.0	6.0	11.0	2.0
Perempuan	18.0	12.0	11.0	2.0	1.0
[total]	53.0	14.0	17.0	13.0	3.0

Figure 6. Attributes by gender

In Figure 6. from the values obtained in the classification by gender, it can be said that male customers' interest in clothing products = very interested, in the electronic product category = interest, in
Beauty Products, and Accessories, = Very Disinterested. For female customers of clothing products = very interested, in the category of beauty products, daily goods = interested, in electronic products = not interested, and in the accessories category = very disinterested

Attribute	Class: Pakaian (0.55)	Produk Kecantikan (0.14)	Barang Harian (0.17)	Elektronik (0.13)	Aksesoris (0.02)
No					
mean	45.7843	35.1667	60.4	38	14
std. dev.	23.2913	23.1942	29.0168	26.5672	0.1667
weight sum	51	12	15	11	1
precision	1	1	1	1	1
Jenis Kelamin					
Laki-Laki	35.0	2.0	6.0	11.0	2.0
Perempuan	18.0	12.0	11.0	2.0	1.0
[total]	53.0	14.0	17.0	13.0	3.0
Usia					
Diatas 20 tahun	43.0	9.0	11.0	10.0	2.0
Dibawah 20 tahun	10.0	5.0	6.0	3.0	1.0
[total]	53.0	14.0	17.0	13.0	3.0

Figure 7. Attributes by age

In Figure 7. The results of the classification based on age above 20 years old customers' buying interest in clothes is worth 43.0 while those under 20 years old are only 10.0, in beauty products, those over 20 years old are still superior with a value of 9.0 while those under 20 years old are only 5.0, on goods daily age above 20 years is still superior with a value of 11.0 and the value of age under 20 years is only 6.0, as well as Electronic products aged 20 years are superior compared to those under 20 years with a value of 10.0 compared to 3.0, and for accessories products aged over 20 years is superior to those under 20 years with a value of 2.0 and 1.0.
In the data that has been obtained, it can be seen that the shopping interest of 20 year olds in clothing products = very interested, in beauty products = not interested, in daily goods, and electronics = interested, and in accessories products = very disinterested
In Figure 8. there is a confusion matrix value, which is a table of several product categories. For clothing products the value is 72, beauty products are worth 5, daily items are worth 7, electronics are worth 5 and accessories are only worth 1, and if you add up the values in the confusion matrix, they will match the total number of respondents.

And it can be seen that there are 5 alphabets in the confusion matrix consisting of a, b, c,
d, and e, the following are the meanings of the 5 alphabets, namely a: Clothing, b: Beauty
And it can be seen that there are 5 alphabets in the confusion matrix consisting of a, b, c confusion matrix yang terdiri dari a, b, c,
d, and e, the following are the meanings of the 5 alphabets, namely a: Clothing, b: Beauty Products, c : Daily Items, d : Electronics and e : Accessories

```
 a  b  c  d  e   <-- classified as
41  3  2  4  1 |  a = Pakaian
 8  1  3  0  0 |  b = Produk Kecantikan
12  1  2  0  0 |  c = Barang Harian
10  0  0  1  0 |  d = Elektronik
 1  0  0  0  0 |  e = Aksesoris
```

Figure 8. Confusion Matrix

The value of the clothing product category is 72, beauty products are worth 5, daily items are worth 7, electronics are worth 5 and accessories are only worth 1, and if you add up the values in the confusion matrix, they will match the total number of respondents.

4. Conclusion

Marketplace is an electronic product marketing platform that brings together many sellers and buyers to transact with each other. The large variety of products sold on Shopee is one of the reasons this application is in great demand by all walks of life. However, the weakness of the large variety of products sold in a marketplace causes buyers who have no potential to buy these products. To overcome this problem, it is necessary to do a classification to determine which products are most in demand by customers. Product categories consist of: Clothing, Beauty Products, Daily Goods, Electronics, and Accessories. The classification method used is Naïve Bayes and the software used is WEKA. Interest categories are divided into 4 categories, namely: Very interested, Interested, Not interested, Very not interested. The results obtained in this study are clothing products (72 respondents) are products that are in great demand, daily goods products (7 respondents) are products of interest, beauty and electronic products (5 respondents) are products that are not in demand, and accessories (1) is a product that is not very attractive to customers on the Shopee marketplace

5. Reference

[1] D.Apriadi, and A. Y. Saputra, "E-Commerce Berbasis Marketplace Dalam Upaya Mempersingkat Distribusi Penjualan Hasil Pertanian," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi, pp. 1-8, 2017

[2] E. R. Widyayanti, "Pengaruh Marketplace Terhadap Peningkatan Pendapatan Pada Ukm (Studi Pada Ukm Di Daerah Istimewa Yogyakarta," Jurnal Optimum, vol 9, 2019

[3] S. Widaningsih and A. Suheri, "Klasifikasi Jurnal Ilmu Komputer Berdasarkan Pembagian Web Of Science Dengan Menggunakan Text Mining,"Jurnal SENTIKA, pp.320–328, 2018

[4] Z. Zhang, "Naïve Bayes Classification In R. Annals Of Translational Medicine," Annals of Translational Medicine, vol 4, 2016

[5] C. Fadlan, S. Ningsih and A. P. Windarto, "Penerapan Metode Naïve Bayes Dalam Klasifikasi Kelayakan Keluarga Penerima Beras Rastra,". Jurnal Teknik Informatika Musirawas, vol 3, 2018

[6] R. N. Devita, H. W. Herwanto and A. P. Wibawa, "Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa Indonesia," Jurnal Teknologi Informasi Dan Ilmu Komputer, vol. 5, no. 4, pp. 427, 2018

[7]B.T. Pham, D. Tien Bui, H.R. Pourghasemi, P. Indra, M. B. D, "Landslide Susceptibility Assesssment In The Uttarakhand Area (India) Using GIS: A Comparison Study Of Prediction Capability Of Naïve Bayes, Multilayer Perceptron Neural Networks, And Functional Trees Methods," Springer, 2017

[8] D. T. Bui, H. Shahabi, A. Shirzadi, K. Chapi, M. Alizadeh, W. Chen, A. Mohammadi, Ahmad, B. Bin, M. Panahi, H. Hong and Y. Tian, "Landslide Detection And Susceptibility Mapping By AIRSAR Data Using Support Vector Machine And Index Of Entropy Models In Cameron Highlands, Malaysia," Jurnal

MDPI (Multidisiplin Digital Publishing Institute), vol. 10, no. 10, 2018
[9] N. Dicky, E. Kamil and M. Ramadhan, "Penerapan Data Mining dengan Algoritma Naive Bayes Clasifier untuk Mengetahui Minat Beli Pelanggan terhadap Kartu Internet XL," Jurnal Ilmiah Saintikom, 2017
[10] S. Nugroho, Adi and Y. A. Sari, "Implementasi Data Mining Menggunakan Weka," Malang, UB press, 2018
[11]E. Syahputra, "Snowball Throwing Tingkatkan Minat dan Hasil Belajar" Sukabumi, Haura, 2020
[12] N. I. Widiastuti, E. Rainarli and K. E. Dewi, "Peringkasan dan Support Vector Machine pada Klasifikasi Dokumen" Jurnal Infotel(Informasi Telekomunikasi dan Elektronika), vol. 9, no. 4, pp. 416-421, 2017
[13] S. Yadav, and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification". Semantic Scholar, vol. 6, 2016.