Classification Analysis of Product Sales Results at Alfamart Using the Naïve Bayes Method

Yuyun Yusnida Lase[1*], Citra Wasti Silaban[1], Alex Sander Sitepu[1] , Reza Kavarin Telaumbanua[1]

[1] *Department of Computer Engineering and Informatics, Politeknik Negeri Medan,* Indonesia

ABSTRACT

This research focuses on the analysis of the number of products sold, especially stock items from the distribution center to Alfamart stores. The main problem discussed in this study is the result of the number of unsold and sold products, which causes overstocking in the warehouse area. To overcome this problem, it will be solved using the Naive Bayes classification method. This research uses sample data of 100 products and uses data collection techniques such as observation and interviews. The collected data is analysed through a classification approach. This research aims to predict goods that sell and do not sell using Rapidminer using the NaïveBayes method. And to produce more accurate data for the product sales process. The reason for using this naïve bayes algorithm in the process of processing and analysing data is because the way this algorithm works uses statistical methods and probability in predicting future results. The validation results show that the Naive Bayes classification method implemented through Rapidminer provides a significant explanation with a fairly high accuracy and a positive effect on the prediction of sales of goods based on consumer demand and needs.

1.   INTRODUCTION

With the development of the times, competition in the world of business and technology is increasing rapidly and business people are competing to develop their business by utilizing existing technology. Alfamart Simpang Pos needs to improve product quality and increase the variety of products most needed by consumers, the rapid development of the business world and this economic system. This causes economic actors to face obstacles in selling their products in the market. This causes competition for companies to improve product quality and find out which products sell the most and are in demand by consumers. In every region in Indonesia, there are many stores that sell various types of products, especially food variations, both light and heavy. Medan City has many Alfamart stores in various places, including the Alfamart Simpang Pos branch [1][2]. Alfamart is a company engaged in the industry of selling food, drinks, or other products. This of course increases the number of other stores operating in the same field, business competition between Alfamart. It is expected that product sales continue to increase, can also increase profits from sales, one of which is by providing products according to market demand so that there are no losses, improving product quality and can adjust product prices to maximise sales. So far, the business has often suffered losses because Alfamart does not have a system that can predict sales because the products provided are lacking or not selling [3].

According to research, many areas of computer science can solve complex problems. Systematic and computerised use of technology is essential as it enables solutions to be found [4][5][6]. Many problems are statistical, group-related, complex and interconnected. Whether with decision enthusiasts, artificial intelligence, or data mining [7][8]. This last science will be proposed as a solution to address this research problem. Data mining is an analytical method designed for this purpose, analysing data sets to find valuable and consistent hidden knowledge [9][10]. Data mining can process large amounts of data and therefore search data automatically[11], data mining plays a very important role in several fields of life, including in the industrial field at Alfamart Simpang Pos. Naive Bayes is a classification using probability and statistical methods based on the British scientist Thomas Bayes, which is to predict the future chances of past experience [12][13][14]. Good potential is also ascribed to the Naive Bayes method of classification, determining the accuracy and efficiency of product data [15][16].

There are several previous studies that look at Naive Bayes in solving problems [17][18][19]. One of them is research that discusses the use of data mining to predict the number of best selling products carried out using the Naive Bayes algorithm with the conclusion that the probability of each attribute is calculated in Quarter 1, Quarter 2, Quarter 3, and Quarter 4 [12][20]. The relevance of an item purchased by consumers can be calculated using the naive Bayes classifier algorithm technique. The accuracy of product data using RapidMiner with an accuracy value of 95.00%, precision sells 100.00%, precision less sells 92.31% [21]. Based on this description, research was conducted to assist Alfamart employees in the application of easy problem solving of sales patterns in stores using the NaiveBayes algorithm.

2.   THEORETICAL FOUNDATION

*Corresponding Author Email: yuyunlase@polmed.ac.id

Naive bayes is the simplest Bayesian network classifier, which is based on the assumption that the attributes of each instance differ from each other depending on the context class. The conditional independence assumption is set as the attribute for this part. While this may be a common misconception, naive Bayes classifiers are often used. Domingos and Pazzani even compete with advanced classifiers in many real-world applications. Many studies have tried to explain this paradox. For many attributes, the assumption of conditional independence of these attributes is very useful to simplify learning, as the parameters for each attribute can be easily estimated separately [22][23][24].

3. RESEARCH METHODS

3.1 Sales Data Collection

Methods used in this study, the data used is research data obtained at Alfamart Simpang Pos. The data is processed using the Naive Bayes Clacification method. The following are the stages in conducting data mining research. Data collection To get an overview of a good system structure, intensive communication with store employees is required. Researchers made direct observations to Alfamart Simpang Pos to get relevant data, and also researchers collected data by taking data samples directly from daily sales records. The data source in this researcher was obtained from observations made by researchers at the Alfamart Simpang Pos store. Then the results of these observations are input into Microsoft Excel to perform manual calculations and are temporary [25].

3.2 Naïve Bayes

The method used in this research is the data mining method, namely Naive Bayes. The Naïve Bayes algorithm is a method that effectively classifies data by optimizing control Estimates in the right probability with the assumption of simplifying conditional attribute values that do not depend on each other if an output value is determined. The following are the completion steps that the author takes in predicting sales of the Naïve Bayes method.
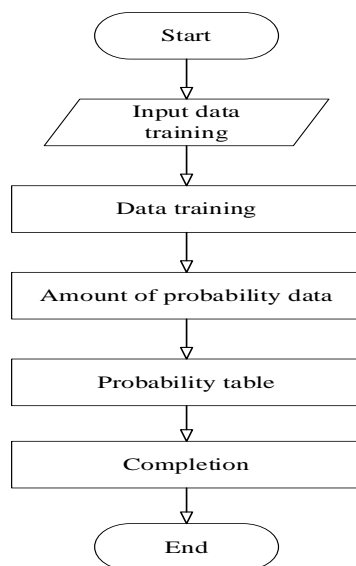

Figure 1. Naïve Bayes model

The description of the figure is :
1. Data entry
   Determine the data to be used in classifying and predicting sales where the data is available at Alfamart Simpang Pos.
2. Data Training
   Calculation of sales amount from the received data.
3. Amount of probability data
   Calculation of sales quantities for each type based on individual data obtained.
4. Probability table
   Calculates the product of one of the variables used to predict sales.
5. Completion
   Naive Bayes algorithm to predict product sales in order to produce products that can be accepted in other Alfamart stores or even in the outside market.
6. Finish
   Draw conclusions from the data that has been classified by Naïve Bayes.

4 RESULTS AND DISCUSSIO

This research is a study that describes and explains the conditions that exist in the field. This research cannot only be explained in the form of words or descriptions and graphs, therefore this research in RapidMiner is based on the Naive Bayes method. In addition, the method used is also looking for data as a tool for the results of this research, namely the observation method. The data used is based on the actual situation in the market business area with a total of 100 datasets. Consists of 7 attributes namely product name, size, variant, product type, product stock, sales, output. The main objective of the research is to obtain the probability amount of product sales data. The calculation process uses an equation to find the probability value of each variable. The data will be divided into 2 namely training data and testing data. The variables that will be used to predict product sales at the Alfamart Simpang Pos store are:

Table 1. Training data

Product name	size	Variants	Product type	Product stock	Sales	Output
niacinamide lab SNPs	32 ml	orange	serum	Low	Slow	Less sellable
Jumbo peeled white bread cider	555 grams	Bargain	Bread	Low	Quick	Apply
....	....	....	....	....	....	....
....	....	....	....	....	....	....
Relaxa	125 grams	Barley mint	Candy	Low	Slow	Less sellable
Kis mint	125 grams	Grapes	Candy	Low	Quick	Apply
Milkita	125 grams	Milk candy	Candy	Low	Slow	Less sellable

Table 2. Testing data table

Product name	size	Variants	Product type	Product stock	Sales	Output
Stick noodles	200 grams	Spicy	Snacks	High	Quick	Apply
Orange leaf spicy basreng	500 grams	Spicy	Snacks	High	Quick	Apply
Fitbar	1 box (12 pcs)	original	Snacks	High	Slow	Less sellable
Mini pocky	22 grams	Chocolate	Snacks	High	Slow	Less sellable
Cadbury dairymilk	62 grams	Chocolate	Snacks	Low	Slow	Less sellable
....	....	....	....	....	....	....
....	....	....	....	....	....	....
Mujigai	80 grams	Creamy rose	Instant noodles	High	Slow	Less sellable
Almona cookies	90 grams	Cookies & cream	Snacks	Low	Slow	Less sellable

The following are the results of the testing process with Rapid Miner software against manual calculations and testing the performance of the Naïve Bayes classification algorithm on the case under study. Where the data connects the testing data excel operator and Naïve Bayes with Apply Model and Performance which serves to produce how good the results of the performance of the data that has been input.
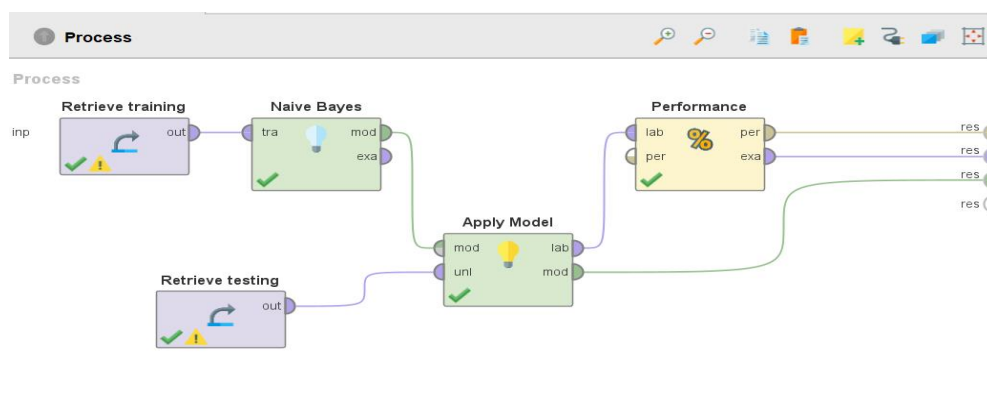

Figure 2. Connecting excel testing and Naïve Bayes Operators with Apply Mode and Performance (Clacification)

Table 3. Accuracy Performance Value

accuracy :95%			
	sold	undersold	class precision
predicted sold	7	0	100%
predicted undersold	1	12	92.31%
class recall	87.50%	100%	

Based on Table 3 above, it is known that the accuracy value is 95.00% with the number of True Sold is 7 and True Undersold is 0 where the results using software with the process carried out by manual calculation are the same, thus it can be said that the calculation process has been successfully carried out correctly.

The following is the calculation process of Naive Bayes which is done manual

$$P(A) = \frac{n(A)}{xn(S)} \qquad (1)$$

Description:
P (A) = Probability of occurrence of event A
n (A) = The number of events in question
n (S) = Total Data

1. Count the number of classes/labels.
   P(Applicable) = 38/100 : 0.38
   P(Less Applicable) = 62/100 : 0.62
   P(Fast) = 38/100
   P(Slow) = 62/100

2. Count the number of cases with the same class.
   P(High | Applicable) = 21/28
   P(High | Less behavioural) = 20/62
   P(Low | Applicable) = 19/38
   P(Low | Less Applicable) = 40/62
   The results of the probability data can be seen in the following table:

Table 4. Behavioural/Inappropriate Probability Results

	Apply	Less Applicable	Sold (Result)	Less Applicable (Result)
P(Applicable/Lack of Applicability)	38	62	0,38	0,62

Table 5. Product Stock Probability

P(S|...	Apply	Less Applicable	Sold (Result)	Less Applicable (Result)
High	21	20	0,55263	0,32258
Low	19	40	0,5	0,64516
Total	40	60	1,05263	0,96774

In the table above, it is known that for the Stock variable, the value of (High | Sold) is 0.55263, (High | Less Applicable) is 0.32258. (Low | Applicable) is 0.5, (Low | Less Applicable) is 0.64516. The sum of the probabilities of the Applicable category is 1.05263 and the sum of the probabilities of the Less Applicable category is 0.96774.

Table 6. Product Sales Probability

	Apply	Less Applicable	Sold (Result)	Less Applicable (Result)
P (Fast/Slow)	38	62	0,38	0,62

Table 7. Product Type Probability

P(J|...	Apply	Less Applicable	Sold (Result)	Less Applicable (Result)
Snacks	4	9	0,10526	0,23684
Instant noodles	3	1	0,07894	0,02631
Parfume	0	3	0	0,07894
Deodorant	2	0	0,05263	0
Hair vitamins	1	0	0,02631	0
Lip balm	0	1	0	0,02631
Lotion	2	0	0,05263	0
Serum	0	1	0	0,02631
Supplement	0	2	0	0,05263
Shampoo	1	3	0,02631	0,07894
Conditioner	1	0	0,02631	0
Bread	3	1	0,07894	0,02631

Ice cream	0	2	0	0,05263
Full cream milk	1	1	0,02631	0,02631
Milk powder	1	2	0,02631	0,05263
Cookies	0	1	0	0,02631
Frozen food	2	2	0,05263	0,05263
Chilli sauce	1	2	0,02631	0,05263
Sweet soy sauce	1	0	0,02631	0
Bottled drinks	6	3	0,15789	0,07894
Canned drinks	1	4	0,02631	0,10526
Boxed drinks	2	3	0,05263	0,02631
Dry tissue	0	2	0	0,05263
Wet tissue	0	1	0	0,02631
Telon oil	0	2	0	0,05263
Toothpaste	1	1	0,02631	0,02631
Detergent powder	2	0	0,05263	0
Liquid detergent	1	0	0,02631	0
Fragrance	2	0	0,05263	0
Anti-mosquito	0	2	0	0,05263
Air freshener	0	3	0	0,05263
Candy	1	3	0,02631	0,05263
Honey	0	1	0	0,02631
Jam	0	2	0	0,05263
Meses	0	1	0	0,02631
Total	38	62	2, 44705	1,44726

5 CONCLUSIONS

Based on the calculations carried out to predict product sales at Alfamart Simpang Pos using the naive bayes algorithm, it can be concluded that the naive bayes classification data mining algorithm is successfully used to predict product sales at Alfamart Simpang Pos. When predicting using the Naive Bayes algorithm, it can create a probability table model that is used to make predictions with a certain accuracy rate of 95.00%. Based on the calculations carried out, it is known that the chances of success are high affecting product sales, namely for the stock variable to be "High", for the sales variable to be "Fast", and for the output is "Sold".

REFERENCES

[1]   K. A. Pranajaya and S. I. Prasetyo, "D Digitizing BUMDes: Application, Opportunities, and Challenges in the Special Region of Yogyakarta," *J. Din.*, vol. 2, no. 2, pp. 1–8, 2022.

[2]   A. R. Lubis, S. Prayudani, M. Lubis, and O. Nugroho, "Sentiment Analysis on Online Learning during the Covid-19 Pandemic Based on Opinions on Twitter using KNN Method," in *2022 1st International Conference on Information System and Information Technology, ICISIT 2022*, 2022, pp. 106–111, doi: 10.1109/ICISIT54091.2022.9872926.

[3]   C. Villavicencio, J. J. Macrohon, X. A. Inbaraj, J.-H. Jeng, and J.-G. Hsieh, "Twitter sentiment analysis towards covid-19 vaccines in the Philippines using naïve bayes," *Information*, vol. 12, no. 5, p. 204, 2021.

[4]   I. Safira and M. Muhathir, "Analysis of Different Naïve Bayes Methods for Categorizing Spices Through Photo using the Speeded-up Robust Feature," in *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, 2023, pp. 29–34.

[5]   G. L. Y. Londo, D. H. Kartawijaya, H. T. Ivariyani, Y. S. P. WP, A. P. M. Rafi, and D. Ariyandi, "A Study of Text Classification for Indonesian News Article," in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, 2019, pp. 205–208.

[6]   F. F. Shahare, "Sentiment analysis for the news data based on the social media," *Proc. 2017 Int. Conf. Intell. Comput. Control Syst. ICICCS 2017*, vol. 2018-Janua, pp. 1365–1370, 2017, doi: 10.1109/ICCONS.2017.8250692.

[7]   N. A. Febriyati, A. D. GS, and A. Wanto, "GRDP Growth Rate Clustering in Surabaya City uses the K-Means Algorithm," *IJISTECH (International J. Inf. Syst. Technol.*, vol. 3, no. 2, pp. 276–283, 2020.

[8]   S. Pradha, M. N. Halgamuge, and N. Tran Quoc Vinh, "Effective text data preprocessing technique for sentiment analysis in social media data," *Proc. 2019 11th Int. Conf. Knowl. Syst. Eng. KSE 2019*, pp. 1–8, 2019, doi: 10.1109/KSE.2019.8919368.

[9]   N. Ginantra *et al.*, "Performance One-step secant Training Method for Forecasting Cases," in *Journal of Physics: Conference Series*, 2021, vol. 1933, no. 1, p. 12032.

[10] A. R. Lubis, S. Prayudani, Y. Fatmi, and O. Nugroho, "Latent Semantic Indexing (LSI) and Hierarchical Dirichlet Process (HDP) Models on News Data," in *2022 5th International Conference on Computer and Informatics Engineering, IC2IE 2022*, 2022, pp. 314–319, doi: 10.1109/IC2IE56416.2022.9970067.

[11] H. Mehta and K. Passi, "Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI)," *Algorithms*, vol. 15, no. 8, 2022, doi: 10.3390/a15080291.

[12] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between multinomial and Bernoulli naïve Bayes for text classification," in

2019 International Conference on Automation, Computational and Technology Management (ICACTM), 2019, pp. 593–596.

[13] M. R. Romadhon and F. Kurniawan, "A comparison of naive Bayes methods, logistic regression and KNN for predicting healing of Covid-19 patients in Indonesia," in *2021 3rd east Indonesia conference on computer and information technology (eiconcit)*, 2021, pp. 41–44.

[14] T. M. Ma, K. Yamamori, and A. Thida, "A comparative approach to Naïve Bayes classifier and support vector machine for email spam classification," in *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, 2020, pp. 324–326.

[15] S. Ghazal, W. S. Qureshi, U. S. Khan, J. Iqbal, N. Rashid, and M. I. Tiwana, "Analysis of visual features and classifiers for Fruit classification problem," *Comput. Electron. Agric.*, vol. 187, p. 106267, 2021.

[16] V. R. Arunachalaeshwaran, H. F. Mahdi, T. Choudhury, T. Sarkar, and B. P. Bhuyan, "Freshness classification of hog plum fruit using deep learning," in *2022 International congress on human-computer interaction, optimization and robotic applications (HORA)*, 2022, pp. 1–6.

[17] A. R. Lubis *et al.*, "Enhancing Text Summarization with a T5 Model and Bayesian Optimization," *Rev. d'Intelligence Artif.*, vol. 37, no. 5, pp. 1213–1219, 2023, doi: 10.18280/ria.370513.

[18] Y. Y. Lase, A. R. Lubis, F. Elyza, and S. A. Syafli, "Mental Health Sentiment Analysis on Social Media TikTok with the Naïve Bayes Algorithm," in *2023 6th International Conference of Computer and Informatics Engineering (IC2IE)*, 2023, pp. 186–191, doi: 10.1109/IC2IE60547.2023.10331126.

[19] A. R. Lubis, Y. Y. Lase, D. A. R, and D. Witarsyah, "Optimization of SVM Classification Accuracy with Bayesian Optimization Utilizing Data Augmentation," in *2023 6th International Conference of Computer and Informatics Engineering (IC2IE)*, 2023, pp. 169–174, doi: 10.1109/IC2IE60547.2023.10331580.

[20] A. R. Lubis, M. K. M. Nasution, O. S. Sitompul, and E. M. Zamzami, "The feature extraction for classifying words on social media with the Naïve Bayes algorithm," *IAES Int. J. Artif. Intell.*, vol. 11, no. 3, pp. 1041–1048, 2022, doi: 10.11591/ijai.v11.i3.pp1041-1048.

[21] M. Wongkar and A. Angdresey, "Sentiment analysis using Naive Bayes Algorithm of the data crawler: Twitter," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*, 2019, pp. 1–5.

[22] S. Kurniawan and I. Budi, "Indonesian tweets hate speech target classification using machine learning," *2020 5th Int. Conf. Informatics Comput. ICIC 2020*, pp. 1–5, 2020, doi: 10.1109/ICIC50835.2020.9288515.

[23] E. Aker, M. L. Othman, V. Veerasamy, I. bin Aris, N. I. A. Wahab, and H. Hizam, "Fault detection and classification of shunt compensated transmission line using discrete wavelet transform and naive bayes classifier," *Energies*, vol. 13, no. 1, p. 243, 2020.

[24] S. Dey, S. Wasif, D. S. Tonmoy, S. Sultana, J. Sarkar, and M. Dey, "A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews," in *2020 International Conference on Contemporary Computing and Applications (IC3A)*, 2020, pp. 217–220.

[25] E. Rosmaini, T. F. Kusumasari, M. Lubis, and A. R. Lubis, "Study to the current protection of personal data in the educational sector in Indonesia," in *Journal of Physics: Conference Series*, 2018, vol. 978, no. 1, doi: 10.1088/1742-6596/978/1/012037