Jurnal Mantik, 6 (1) (2022) 293-302

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

Sentiment Analysis of Detergen Products at Suzuya Mall Rantauprapat Navie Bayes Method

Nurhikmah Nasution[1], Ibnu Rasyid Munthe[2], Gomal Juni Yanris[3]

[1,2,3] Faculty of Science and Technology, Labuhanbatu University, Rantauprapat, 21418, Indonesia

E-mail: hikmahnasution31@gmail.com[1], ibnurasyidmunthe@gmail.com[2], gomaljuniyanris@gmail.com[3]

ARTICLE INFO

Article history:
Received: Mar 12, 2022
Revised: Aprl 9, 2022
Accepted: Aprl 29, 2022

Keywords:
Suzuya Mall, Naïve Bayes, Confusion Matrix, Accuracy, Model Classification

Suzuya Mall Rantauprapat is an Indonesian convenience store chain with numerous locations. Based on data from top brands from 2014 to 2018, we discovered that the market share of powder detergent is uncertain each year. Powder detergent companies must understand consumer desires and monitor the position of their products on a regular basis in order to anticipate market share changes.Data is cleaned by removing capital letters and punctuation marks, followed by feature extraction. Feature Extraction aims to perform calculations and comparisons that can be used for the classification of an image. Nave Bayes is a method or stage of data processing with the nave bayes method. The Naive Bayes classifier is a method of classifying based on Bayes" tyrannology. It uses probability and statistical methods first put forward by Thomas Bayes. The classification model is then assessed to determine its accuracy and performance in computer science.Data set is data that has been changed in the form of tabulation from research data into excel form. After the data is processed by the naïve bayes method, the confusion matrix is obtained as follows. It can be known the factors that consumers use in choosing detergent products are the fragrance factor, price, foam produced, and the effect on the hands.Machine learning, specifically Nave Bayes, will be used in this research methodology. The Naive Bayes classifier is a method of classification rooted in Bayes" theorem. A confusion matrix is a table that contains many rows of test data that the classification model predicts to be true or false. Process data mining can be made up of any number of nested operators that are described in XML files and built using RapidMiner.

1. Introduction

According to a survey conducted in 2016, 95 percent of Indonesians use detergent, with 73 percent using powdered detergents. Rinso, Daia, So Klin, Attack, and Molto are some of the most well-known brands in Indonesia. We discovered that the market share of powder detergent each year is uncertain based on data from top brands from 2014 to 2018. The increase in the number of brands on the market has caused this shift.[1]

Many factors contribute to brand switching, including discounted prices and lower product prices than competitors. Powder detergent companies must understand consumer desires and monitor the position of their products on a regular basis in order to anticipate market share changes. To avoid being usurped by competitors, the company must improve the quality of its products and continue to innovate. Suzuya is an Indonesian convenience store chain with numerous locations. At affordable prices, more than 200 food items and other necessities are available. Due to the large number of products available at Suzuya mall Rantauprapat, public opinion on detergent products varies. This can be caused by a variety of factors, including the consumer's previous product experience and product knowledge.[2]

Sentiment analysis is the process of gaining information by comprehending, evaluating, or assessing people's feelings. The classification of polarity contained in the text, which is put forward as positive, negative, and neutral classes, is a basic principle of sentiment analysis. The use of machine learning is used to conduct sentiment analysis.[3] Machine learning (ML) is a technology that allows machines to learn on their own without the need for human intervention. ML can also analyze existing data as well as data it collects in order to perform specific tasks. Several mathematical scientists, including Adrien Marie Legendre,

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Thomas Bayes, and Andrey Markov, coined the term "machine learning" in the 1920s. IBM's "deep Blue," released in 1996, is an example of machine learning in action. [4] Based on research conducted by Lestari et al. (2019), the market share positions for detergent brands in the 31st period (July 2021) are Rinso 40%, Daia 20.8%, So Klin 18.8%, Attack 12.9%, and Molto 7.5%. The brand that has the highest market share is Molto. [5] As a result, this study will use the nave Bayes method in data recognition to analyze sentiment toward detergent products in Suzuya Mall Rantauprapat. The purpose and goal of this study is to identify the variables that have the greatest impact on the community's detergent product selection at Suzuya Mall Rantauprapat, as well as to determine which detergent products are the most popular among Suzuya Mall Rantauprapat residents.

2. Method

The first step in this research methodology is to collect data from various sources. The next step is to manually label each dataset with positive, negative, and neutral labels. After the labeling procedure is completed, the data cleaning procedure follows. Data is cleaned by removing capital letters and punctuation marks, followed by feature extraction. Machine learning, specifically Nave Bayes, will be used to process these features. Here's how the research progressed.

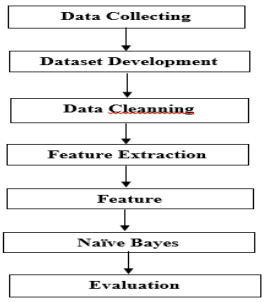

Figure 1. Research Flow [10]

2.1 Data Collecting

Data collection techniques, also known as "data collecting," is a research method in which researchers collect data systematically for analysis using the scientific method. [6]

Table 1. Data Collecting

No	Name	Age	Gender	Last Education
1	Kinanti Sekar	18	WOMAN	HIGH SCHOOL
2	Sukma Asrofi	21	MAN	HIGH SCHOOL
3	Mustika	25	WOMAN	HIGH SCHOOL
4	Haifa Zahra Dzkiani	20	WOMAN	HIGH SCHOOL
5	Intan Widyarti	21	WOMAN	HIGH SCHOOL
…	……...	…	……	……
80	Eva Nur Syafitri	23	WOMAN	S1
81	Fitri Nadia	24	WOMAN	HIGH SCHOOL
82	Erika Putri Sinaga	26	WOMAN	HIGH SCHOOL
83	Aida	23	WOMAN	S1
84	Meysi Rusliana	24	WOMAN	HIGH SCHOOL
85	Fuji Astriani Sirait	21	WOMAN	HIGH SCHOOL

The most common form of diagram that students can easily find is a diagram of drawings, bars, lines and circles. Students can start the activity by constructing and interpreting bar diagrams by using one-on-one correspondence ideas and concepts on drawing diagrams. [16]

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

2.2 Dataset Development

A database is a collection of data that is logically related and designed to meet the information needs of an organization. [9]. Dataset development can be summed up as a data set consisting of some data of important people or people who have high positions in an organization or company.

Table 2. Dataset Development

No	Product Type	Positive Response	Negative Responses
1	Daia	Better than liquid detergent	
2	Rinso Liquid	Make clothes softer and last longer	
3	Rinso Liquid	Long lasting fragrance	
4	Rinso Bubuk	Eradicates stubborn stains and lots of foam	
5	So Klin Liquid	Fragrant and cheap and softer in the hands	
6	Daia	Easy to find	
7	Attack Gel	Use is more hygienic and effective	
80	Daia		Hot in hand
81	Daia		The content is too little
82	So Klin Liquid		Give off an unpleasant odor
83	Daia		Hot in hand
84	Rinso		Expensive
85	Attack Plus Softener		Rarely Promo

2.3 Data Cleanning

Data cleaning is the process of removing noise, inconsistencies, and irrelevant data from a database. Data cleaning is the process of removing noise, inconsistencies, and irrelevant data from a database.

Table 3. Data Cleanning

Responses	Point
More expensive and draining pockets	negative
Hands get rough	negative
The size is nothing small.	negative

2.4 Feature Extraction

Feature Extraction is one way that can be used to recognize an object based on the special histogram that the object has. Feature Extraction aims to perform calculations and comparisons that can be used for the classification of an image based on the characteristics of the histogram owned. [7]

Table 4. Feature Extraction

Row No	Word	In Document	Total
1	Fragrant	31	31
2	Hands	25	26
3	Endure	26	26
4	Hot	24	24
5	The foam	22	22
6	Price	19	19
7	The fragrance	19	19
8	Detergent	12	17
9	Stain	16	17
10	Hand	17	17

2.5 Feature

Feature is data that has gone through future extraction.

Table 5. Feature

Row No	Word	In Document	Total
189	Supermarket	1	1
190	Hands	1	1
191	Texture	1	1
192	The texture	1	1
193	Hang	1	1
194	Sometimes	1	1
195	Famous	1	1
196	Injured	1	1
197	Not	1	1
198	Ink	1	1

2.6 Naïve Bayes

Naïve bayes is a method or stage of data processing with the naïve bayes method

Table 6. Data processing by Naïve Bayes Method

No	text	label
1	Much thicker but quickly soluble in water. Its use is easier to measure so that its use is more hygienic and more able to lift stains	positive
2	More fragrant and sidey	positive
3	Because the price is cheaper and easier to find	positive
4	Cheap price is affordable and can clean stains	positive
…….	…….	…
167	More expensive and draining pockets	negative
168	Hands get rough	negative
169	The size is nothing small.	negative
170	Less foam	negative

2.7 Evaluation

Evaluation is the final stage after processing data by method, where the data will be analyzed.

a. Product

If a product is a tool or something that is the answer or solution to a problem of consumer needs, then we must consider problems or consumer needs when developing products. Meanwhile, according to William J. Stanton, products are a collection of tangible and intangible characteristics such as color, price, the product's good name, the store's good name (retailer), and factory and retailer services received by buyers to meet the store's needs and desires.[1]

b. Sentiment analysis

Sentiment analysis is a method for determining whether the content of a text dataset (documents, sentences, paragraphs, and so on) is positive, negative, or neutral.[3] The computational analysis of opinion, feelings, and subjectivity in text is known as sentiment analysis.[4]

c. Naive Bayes Classifier

The Naive Bayes Classifier is a method of classifying based on Bayes' theorem. It uses probability and statistical methods first put forward by Thomas Bayes. The method predicts future opportunities based on previous experience, so this is known as the Bayes Reality.[6]. The Naive Bayes classifier is a method of classification rooted in Bayes' theorem. The method of classifying using probability and statistical methods, namely predicting odds based on previous experience (Bayes' with its characteristics), is a very strong

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

assumption (nave).[7] The Naive Bayes algorithm is based on the application of Bayes' theorem (bayes rule) assuming strong (naïve) independence for a computer programme to predict its own behaviour in real-world situations, i.e., human behaviour and other human intelligence.[8]. The Bayes formula is the foundation of Nave Bayes, which is used in programming. The model used in Naive Bayes is a non-equivalence model, i.e. it does not have to be identical to the one used in Bayes theory. [12]

Equation 1

$$P(A|B) = (P(B|A) * P(A))/P(B) \quad \text{(i)}$$

Information:

The chance of event A as B is determined by chance B when A, chance A, and chance B.

In its application the formula in equation 1 turns into

Equation 2.

$$P(Ci|D) = (P(D|Ci) * P(Ci)) / P(D) \quad \text{(ii)}$$

Naïve Bayes or can be referred to as Multinomial Naïve Bayes is a simplification model of the Bayes Method that is suitable in classifying text or documents where

Equation 3

$$VMAP = \arg\max P(Vj \mid a1 , a2 ,.....an) \quad \text{(iii)}$$

Based on equation 3, equation 1 can be written as follows:

Equation 4

$$VMAP = \text{agr max (Vj eV)}\ P(a1,a2\ldots\ldots an \mid P\ (Vj)|P\ (a1,a2\ldots.an) \quad \text{(iv)}$$

Here is the completion flow of the Naïve Bayes method:

1. Read training data
2. Calculate the Number and probability
   If there is numerical data, then find the mean value and standard deviation of each parameter that describes the number data. Here is the formula used to calculate the average (mean) can be seen as follows:

$$\mu = \sum_{i=1}^{n} x_i \text{ atau } \mu = \frac{x_1+x_2+x_3+\cdots+x_n}{n} \quad \text{(v)}$$

   Information:
   μ : average count (mean)
   xi : sample value to –i
   n : number of samples

   And the following formula used to calculate the standard deviation value can be seen below:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}} \quad \text{(vi)}$$

   Information:
   σ: standard deviation
   xi : x to –i value
   μ: counting average
   n : number of samples

   If the data is not numeric, calculate the probability value of each same category, with the amount of data from the same category and then divided by the data in that category..
3. Probability Value of Each Class Feature
   In the table below, the probability value of each feature in a class is calculated by multiplying it by the weighting data from that category by the total amount of data in that category.
4. Gaussian Distribution Value
   The above is an example of how to calculate the probability value for a data testing feature that has numerical data or numbers. Here is the equation for finding gaussian distribution values for such a feature, as well as a generalised version of the Equation for Finding Gaussian Distributions.
5. The Final Probability of Each Class
   Calculating the final probability for each class means entering all existing gaussian distribution value data into the same class.
6. Final Probability

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

The final probability is obtained by sharing the probability value of one category with the sum of the values of all other categories. It is calculated by combining the probabilities of each class into the Nave Bayes Classifier formula. After obtaining the final probability, the last step is normalized to normalise the number of classes into a single category.

d. Confusion Matrix

A confusion matrix is a table that contains many rows of test data that the classification model predicts to be true or false. An assessment is carried out to determine the performance and accuracy of the classification methods used in the computer science field. The classification model is then used to determine its accuracy and performance as indicated.[14]

Table 7. Confusion Matrix

Kategori x	*Predicted*	
Actual	*True Positive*	*False Positive*
	True Negative	*False Negative*

Information:

TP (True Positive) indicates the amount of test data that the system has classified into category x, and all such data is indeed included in category x. FN (False Negative) indicates such data should not belong to category x but rather a range of other categories. TN (True Negative) means the system does not have any data that it does not classify into x category.

e. Precision, Recall dan Accuracy

The success rate of rediscovering information is called recall, and the degree of clinability is called accuracy.. [11]

3. Result and Discussion

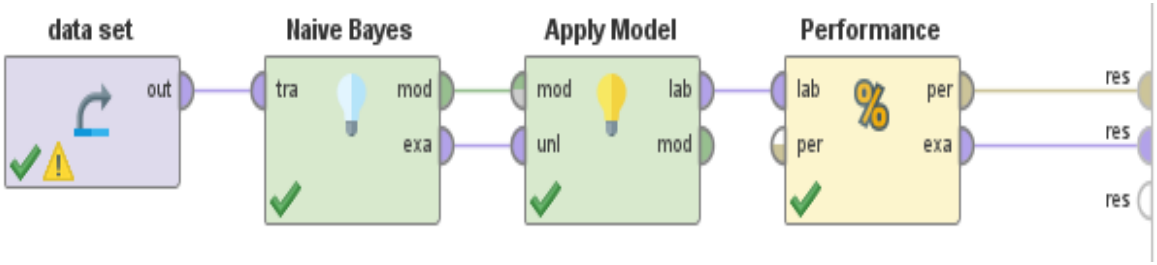

Figure 2. Process Metode Naïve Bayes Pada Rapid Miner

Rapid Miner is a data mining and machine learning environment that includes data loading and transformation (ETL), preprocessing and data visualization, modeling, evaluation, and deployment procedures. Process data mining can be made up of any number of nested operators that are described in XML files and built using RapidMiner. [13] Here are the stages performed in the analysis of the naïve bayes:

1. Data set is data that has been changed in the form of tabulation from research data into ecel form. Here's the data set

Table 8. Dataset

No	text	label
1	Much thicker but quickly soluble in water. Its use is easier to measure so that its use is more hygienic and more able to lift stains	positive
2	More fragrant and sidey	positive
3	Because the price is cheaper and easier to find	positive
4	Cheap price is affordable and can clean stains	positive
…….	…….	…
167	More expensive and draining pockets	negative
168	Hands get rough	negative

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

169	The size is nothing small.	negative
170	Less foam	negative

2. Part attribute naïve bayes is an algorithm naïve bayes that serves as a method used in processing data. [15]

Table 9. Data Sharing Training and Data Testing

No	*Class*	*Data Training*	*Data Testing*
1	positive	51	48
2	negative	35	48
	Total	86	86

Table 10. Example of Frequency Term Occurrence

Word	Frequency of Word occurrence (W_k)	
	positive	negative
Endure	28	10
Soft	3	1
More	2	2
Fragrant	1	3
Hygienic	1	4
Stubborn	1	6
Affordable	-	-
Stain	10	1
Thrifty	1	1
Hand	1	1

Next Look for the probability of the word resistant, soft, more, fragrant, hygienic:
Known:

n*Tweet* Positive :48
n*Tweet* Negative :48

a. The Word Hold

P(Hold|Positive) = $\frac{28+1}{48+86}$ = 0,21

P(Hold|Negative) = $\frac{10+1}{48+86}$ = 0,08

b. The Word Soft

P(Soft|Positive) = $\frac{3+1}{48+86}$ = 0,02

P(Softt|Negative) = $\frac{1+1}{48+86}$ = 0,01

c. The Word More

P(More|Positive) = $\frac{2+1}{48+86}$ = 0,02

P(More|Negative) = $\frac{2+1}{48+86}$ = 0,02

d. The Word Fragrant

P(Fragrant|Positive) = $\frac{1+1}{48+86}$ = 0,01

P(Fragrant|Negative) = $\frac{3+1}{48+86}$ = 0,02

e. The Word Hygenic

P(Hygenic|Positive) = $\frac{1+1}{48+86}$ = 0,01

P(Hygenic|Negative) = $\frac{4+1}{}$ = 0,03

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

48+86

Here is the result of the probability of the word

3. The apply model attribute is an attribute used to apply or run the naïve bayes method earlier.
4. Finally, the performance attribute is an attribute that serves to determine the level of accuracy of the method being passed.

After the data is processed by the naïve bayes method, the confusion matrix is obtained as follows.

Table 11. Confusion Matrix Results

	true positif	true negatif	class precision
pred. positif	85	0	100.00%
pred. negatif	0	58	100.00%
class recall	100.00%	100.00%	

From the Table it is known that the value for true positive is 85 and the negative tue is 58 with an accuracy of 100%. Here is a chart for the negative and positive labels on the study.

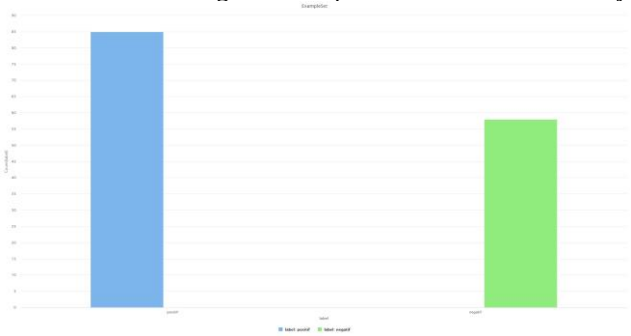

Figure 3. Positive and negative label graph

Where:
Blue = Positive
Green = Negative.

Based on the results of data processing, several words are often used as reference detergent users in providing opinions about products can be seen in the following figures:

Figure 4. Words that represent the product

Based on the figure , it can be known the factors that consumers use in choosing detergent products are the fragrance factor, price, foam produced, and the effect on the hands.

4. Conclusion

Machine learning, specifically Nave Bayes, will be used in this research methodology. The Naive Bayes classifier is a method of classification rooted in Bayes' theorem. A confusion matrix is a table that contains many rows of test data that the classification model predicts to be true or false. Process data mining can be made up of any number of nested operators that are described in XML files and built using RapidMiner. Confusion Matrix Results. known that the value for true positive is 85 and the negative tue is 58 with an accuracy of 100%. Here is a chart for the negative and positive labels on the study it can be known the

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

factors that consumers use in choosing detergent products are the fragrance factor, price, foam produced, and the effect on the hands.

References

[1] Oscar, B., & Megantara, H. C. (2020). Pengaruh atribut produk terhadap keputusan pembelian produk muslim army. *Jurnal Bisnis Dan Pemasaran*, *10*, 1–12.

[2] Juniarti, A. D. (2019). Produk dan Harga Koran Radar Banten. *Jurnal InTent*, *2*(2), 113–121.

[3] Chandani, V., & Wahono, R. S. (2015). Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentimen Review Film. *Journal of Intelligent Systems*, *1*(1), 55–59.

[4] Indrayuni, E. (2019). Klasifikasi Text Mining Review Produk Kosmetik Untuk Teks Bahasa Indonesia Menggunakan Algoritma Naive Bayes. *Jurnal Khatulistiwa Informatika*, *7*(1), 29–36. https://doi.org/10.31294/jki.v7i1.1

[5] Lestari, W. A., Samanhudi, D., & Wati, E. P. (2019). Analisis Pangsa Pasar Detergen Bubuk Dan Penentuan Strategi Pemasaran Pada Merek Yang Memiliki Pangsa Pasar Terkecil Dengan Metode Markov Chain Dan Swot Di Wilayah Surabaya Timur. *Tekmapro : Journal of Industrial Engineering and Management*, *14*(2), 1–12. https://doi.org/10.33005/tekmapro.v14i2.52

[6] Syukri Mustafa, M., Rizky Ramadhan, M., & Thenata, A. P. (2017). Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. *Citec Journal*, *4*(2), 151–162.

[7] Ratnawati, F. (2018). Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter. *INOVTEK Polbeng - Seri Informatika*, *3*(1), 50. https://doi.org/10.35314/isi.v3i1.335

[8] Wijaya, H. D., & Dwiasnati, S. (2020). Implementasi Data Mining dengan Algoritma Naïve Bayes pada Penjualan Obat. *Jurnal Informatika*, *7*(1), 1–7. https://doi.org/10.31311/ji.v7i1.6203

[9] Oktavia, T. (2012). *Pemodelan Sistem Basis Data Relasional Pada Unit Operasional Pelayanan Kesehatan*. *2012*(semnasIF), 229–236.

[10] Wahyu Sholeha, E., Yunita, S., Hammad, R., Cahya Hardita, V., Rekayasa Komputer Jaringan, T., & Tanah Laut, P. (2022). *Analisis Sentimen Pada Agen Perjalanan Online Menggunakan Naïve Bayes dan K-Nearest Neighbor (Sentiment Analysis of Online Travel Agent Using Naïve Bayes and K-Nearest Neighbor)*. *3*(4), 203–208.

[11] Sari, Z. T., Indarto, W., Puspitasari, E., Teacher, S., Program, E., & Childhood, E. (n.d.). *Description of Knowledge About the Importance of Peer To Play With Friends for Children Ages 4-6 Years in District Sukajadi Tk Ridha Pekanbaru*. 1–9.

[12] Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional. *Jurnal Tekno Kompak*, *15*(1), 131. https://doi.org/10.33365/jtk.v15i1.744

[13] Ramamohan, Y., Vasantharao, K., Chakravarti, C. K., & Ratnam, a S. K. (2012). A Study of Data Mining Tools in Knowledge Discovery Process. *International Journal of Soft Computing and Engineering*, *2*(3), 191–194.

[14] Dwiyanti, Y., Herdiani, A., & Puspitasari, S. Y. (2017). *Memprediksi Status Berlangganan Klien Bank Pada Kampanye Pemasaran Langsung Dengan Menggunakan Metode Klasifikasi Dengan Algoritma C5.0*. *4*(2), 3138–3147.

[15] Imandasari, T., Irawan, E., Windarto, A. P., & Wanto, A. (2019). Algoritma Naive Bayes Dalam Klasifikasi Lokasi Pembangunan Sumber Air. *Prosiding Seminar Nasional Riset Information Science (SENARIS)*, *1*(September), 750. https://doi.org/10.30645/senaris.v1i0.81

[16] Kurnia, A. B., & Suryowati, E. (2016). Penerapan Realistic Mathematics Education Dalam Pembelajaran Membaca Diagram Batang Dan Garis Siswa Smp Kelas Vii. *AdMathEdu : Jurnal Ilmiah Pendidikan Matematika, Ilmu Matematika Dan Matematika Terapan*, *4*(2). https://doi.org/10.12928/admathedu.v4i2.4793

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.