Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

APPROACHES TO CUSTOMER TYPES CLASSIFICATION METHOD IN THE SUPERMARKET

Nanang Ruhyana-1*), Tati Mardiana-2

Sains Data
Universitas Nusa Mandiri
nanang.ngy@nusamandiri.ac.id, tati.ttm@nusamandiri.ac.id
(*) Corresponding Author

Abstract

The development of the retail industry in the economy is very rapid, so it provides good economic growth. One of the retailers is supermarkets. In supermarkets, consumers can buy goods directly, so they must be served well. The problem is how supermarkets can continue to increase their sales results because there is a lot of competition from supermarket competitors, so the marketing team, when creating events or promotions, must be right on target so that loyalty for member or non-member customers can be measured, which will be used as the right marketing strategy and can increase customer satisfaction when the customer is satisfied with the services, products and promotional activities at the supermarket, the customer will continue to make purchases and will increase the results of achieving good sales. Based on this problem, how will this research apply the classification method so that when we can make predictions from supermarket sales data for member and non-member customers, there will be a lot of insight for the marketing team so that marketing activities are suitable on target for member or non-member customers? This research uses machine learning methods for data classification, using the Support Vector Machine (SVM) and Naïve Bayes algorithms. The results of this research are from the Support Vector Machine (SVM) algorithm. Accuracy is 0.493, while using the Naïve Bayes algorithm is 0.535. From the results of this research, the use of the Naïve Bayes algorithm is better than SVM to approach the prediction of member and non-member customer classification in supermarket data in this research.

Keywords: Supermarkets, Customer Types, Classification, SVM, Naïve Bayes

Abstrak

Perkembangan industri retail dalam perekonomian sangat pesat sehingga memberikan pertumbuhan ekonomi yang baik, salah satu retail adalah supermarket, pada supermarket konsumen dapat membeli barang secara langsung, maka konsumen harus dilayani dengan baik. Permasalahannya bagaimana supermaket terus bisa meningkatkan hasil penjualannya, karena persaingan kompetitor supermarket ini sangat banyak, sehingga tim pemasaran saat membuat event-event atau promosi yang harus tepat sasaran sehingga loyalty bagi para pelanggan member atau nonmember dapat diukur, yang akan dijadikan strategi pemasaran yang tepat dan dapat meningkatkan kepuasan pelanggan, saat pelanggan tersebut merasa puas dengan layanan, produk dan kegiatan promosi pada supermarket, pelanggan akan terus melakukan pembelian dan akan meningkatkan hasil pencapaian penjualan yang baik. Dari permasalah tersebut, bagaimana penelitian ini akan menerapkan metode klasifikasi, sehingga pada saat kita bisa membuat prediksi dari data penjualan supermarket untuk pelanggan member dan nonmember, akan banyak wawasan bagi tim pemasaran, sehingga kegiatan pemasaran menjadi tepat sasaran untuk pelanggan member atau nonmember. Penelitian ini menggunakan metode pembelajaran mesin untuk klasifikasi data, menggunakan algoritma Support Vector Machine (SVM) dan Naïve Bayes,Hasil penelitian ini dari algoritma Support Vector Machine (SVM) Akurasi , 0.493 sedangkan menggunakan algoritma Naïve Bayes 0.535. Dari hasil penelitian tersebut, penggunaan algoritma Naïve Bayes lebih baik dibandingkan SVM sehingga dapat mendekati prediksi klasifikasi pelanggan member dan nonmember pada data supermarket pada penelitian ini.

Kata kunci: Supermarket, Tipe Pelanggan, Klasifikasi, SVM, Naïve Bayes

INTRODUCTION

The rapid retail industry in Indonesia is one of the economic factors that significantly influences its role in improving people's welfare. There are very diverse forms of retail, both traditional and modern. The global retail industry is transforming. The future of the retail industry will be very different. Online and offline payment technologies will bring value innovation. As digitalization moves across channel boundaries, online and offline retail channels will expand(Liao & Yang, 2020). One of the retail industries is supermarkets. Supermarkets are offline retail stores that currently have a considerable number. So how can this research approach in classifying predictions for members and non-member customers to be able to carry out marketing strategies, promotions, and loyalty programs for customers so that many benefits and insights can be increased in carrying out supermarket business processes, and can also be seen from consumer behavior from the data, consumer behavior is an essential factor in winning the competition, so every company will try to continue to optimize all factors that can increase consumer buying interest(Kojongian et al., 2021), a supermarket needs to carry out an analysis of the types of products that customers like to buy at a supermarket(Nengah Widya Utami & I Wayan Budi Suryawan, 2021). Before conducting research, the author also looks for literature as a reference by looking for comparisons, problems, or things that need to be improved from previous research so that this research can continue to develop and avoid the same research or duplication if it has been done previously.

Several research studies have been conducted, including implementing decision trees to classify supermarket payment methods. By understanding the factors influencing the payment method, companies can adjust promotions or discounts for each payment method more appropriately(Indrayana et al., 2023). There is also research discussing the Implementation of Data Mining to Classify Sales Data in Supermarkets Using the Naïve Bayes Algorithm by classifying mining data in supermarkets using the Naive Bayes algorithm to find out the types of products that customers prefer in 3 branches in 3 months, product categories obtained from classifying sales products can produce accuracy reaching 98.5%(Indriyani Indriyani & Agus Bahtiar, 2023). There is classification prediction research in several retail industries, including factors affecting the choice of payment method in modern retail shops. The research method in this research is quantitative research using the PLS-SEM method. The research results show that transaction value, income, level of education, ease of use awareness, and risk awareness positively and significantly increase the likelihood of consumer behavior to choose non-cash payment methods (Maya Nur Annisaa et al., 2023). The research also discusses the Segmentation and Classification of Customer Payment Behavior in Multimedia Companies Using the K-Means and C4.5 Algorithms. This research uses the classification method with the K-mean and C.45 algorithms in multimedia companies for customer payment behavior. In his research, customer payment behavior will be taken as customer potential as one of the classification attributes. Feature extraction with k-means can increase the accuracy of the C4.5 algorithm. This was noted by improving accuracy from 59.02% to 77.31% and AUC from 0.537 to 0.836. Potential customer attributes can also be used as a reference in promoting, retaining, and managing bankrupt customers(Nurelasari, 2019).

In 2022, there will be a research with the theme Application of SVM for Sentiment Classification in Review Comments in Indonesian in Online Shops, in its research classifying user comments on a product in an online shop. Comments are taken from one of the online shops, namely Shopee. The comments taken are of two types: positive feedback and negative feedback. Data collection and validation were assisted by expert Ony Jeselin S.Psi so that the process of labelling positive and negative data could be verified. Sentiment classification is carried out using the SVM method to see how effective this method is in classifying positive and negative feedback(Refo et al., 2022).

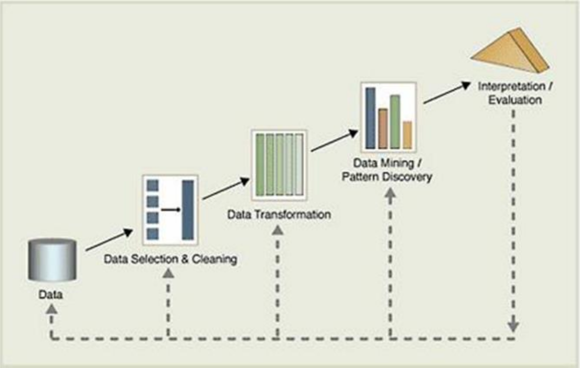
The author uses quantitative methods from various research and literature studies with transaction data from a retail industry, namely supermarkets, taken from https://www.kaggle.com/datasets/muhammadeh abmuhammad/supermarket-sales. The author conducts research using the SVM and Naïve algorithms. Bayes classifies label data for member and non-member customers so that the classification prediction results from each algorithm are better compared. The author uses SVM and Naïve Bayes from previous research journals, and no one has done this, so it differs from previous studies results. So, this algorithm will be used as a reference for future machine learning models. Even though there has been no prior research using SVM, SVM can solve classification and regression problems with linear or nonlinear kernels, which can become a learning algorithm that can be used for classification and regression. SVM also has high accuracy and a relatively small error rate. The ability to overcome overfitting does not require too large data and can be used to make predictions(Achyani, 2017).

RESEARCH METHODS

This research uses quantitative research. In quantitative research, the population and sample need to be considered carefully. Population can be defined as all members of a group of humans, animals, events, or objects who have specific characteristics or a set of characteristics in common(Rifka Agustianti, Pandriadi, Lissiana Nussifera, Wahyudi, L. Angelianawati, Igat Meliana, Effi Alfiani Sidik, Qomarotun Nurlaila, Nicholas Simarmata, Irfan Sophan Himawan, Elvis Pawan, Faisal Ikhram, Astri Dwi Andriani, Ratnadewi, 2022).

The author carries out data mining predictions using the SVM algorithm, Support Vector Machine (SVM), a technique for making predictions in classification and regression cases. SVM is in the same class as ANN regarding function and problem conditions that can be solved. Both are included in the supervised learning class, where implementation requires a training stage followed by a testing stage(Chrisdiyanti et al., 2023). Also, the author uses the Naïve Bayes algorithm. Bayes's theory is about finding the possibility of something based on previously existing data. This method can also classify opinions based on previously trained data. The essence of naive Bayes is to find the highest probability of the data(Amsury et al., 2022). to be able to compare between the two algorithms, which is better for predicting classification.

This author uses the Knowledge Discovery in Database framework to facilitate this research.


Interpretation / Evaluation
Data Mining / Pattern Discovery
Data Transformation
Data Selection & Cleaning
Data

Source: (Ruhyana et al., 2021)

Figure 1: Knowledge Discovery in Database.

In the data mining process, you can use a framework, one of which is KDD, as a reference in the process of producing research or development from data, which in KDD includes data, data selection, data cleansing, data transformation, data mining and evaluation, Knowledge discovery in Databases is an intersection of several disciplines like statistics, databases, AI, and visualization, and equally a link to the high performance of parallel computing. It affects an interdisciplinary knowledge domain and other general-purpose tools(Akanmu Semiu Ayobami, 2012).

In this research, data is taken from supermarket sales transactions: https://www.kaggle.com/datasets/muhammadehabmuhammad/supermarket-sales. There are 1000 datasets from this data, of which there are 16 sales attributes, and classification can be selected by data classification rules in the form of labels. It also uses 80% training data and 20% testing data. In this research stage, the following flow was obtained:

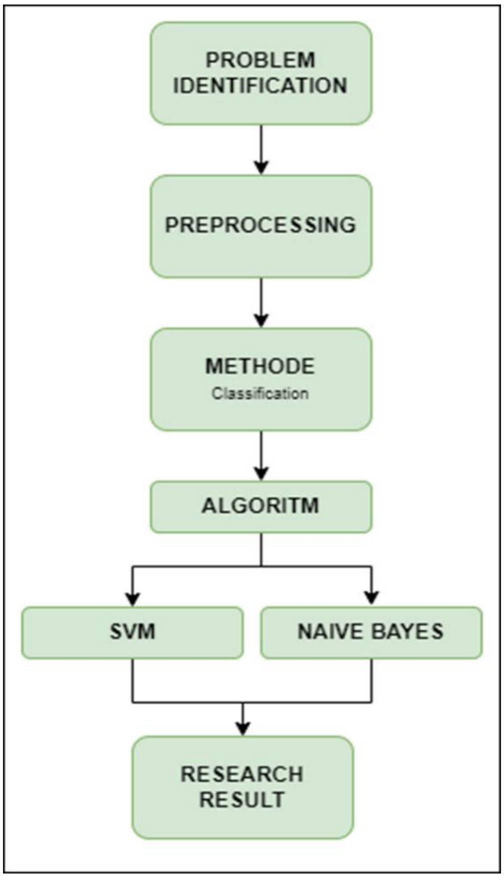


Figure 2: Research Stage

In the research stage figure above, the author describes the processes carried out in stages as follows:

1. Problem Identification

Identify broad problems in general sales sources to see general problem identification and look for more detailed solutions and contributions. Data mining is the analysis of data to find transparent relationships and draw conclusions that are not yet known or previously understood and are valuable for the owner of the data(Noviyanto, 2020).

2. Pre-processing

The next step is to recognize the existing data from the data that will predict the best solution. Sales transaction data must be processed before data mining techniques can be applied by removing noise from the data, selecting data, and transforming data so it can be processed using data mining techniques(Riyadi et al., 2022).

3. Method

The researcher will describe the classification method for predicting data with a class or label in this research. Classification is a process. Data processing is carried out to find a model or explain and differentiate concepts from data classes, aiming to estimate a class of particular objects whose class is unknown(Monika & Furqon, 2018)

4. Algorithm

In this research, we use two comparisons of the SVM and Naïve Bayes algorithms, which have been previously explained, so that we can apply and compare each of these algorithms.

5. Research Result

From the results of this research, we can create a pattern from the data to predict the classification of member and non-member customer type data, which can be used as a reference in sales strategies for the marketing team in a supermarket.

RESULTS AND DISCUSSION

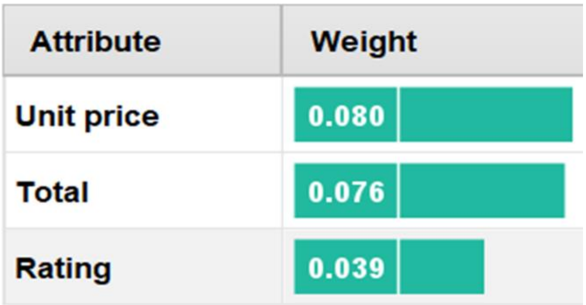
From the data that will be analyzed in this research, the data source has 1000 datasets and 17 data attributes in Table 1, and the description is as follows:

Table 1. Dataset Sales Supermarket

Attribute	Information
Invoice id	Computer-generated sales slip invoice identification number.
Branch	Branch of supercenter (3 branches are available identified by A/B and C).
City	Location of supercenters
Customer type	Type of customers by Members for customers using member cards and Normal for those without member cards.
Gender	Gender type of customer
Product line	General item categorization groups - Electronic accessories/ Fashion accessories/ Food and accessories/ Food and beverages/ Health and beauty/ Home and lifestyle and Sports and travel
Unit price	The price of each product is $
Quantity	Number of products purchased by customer
Tax	5% tax fee for customer buying
Total	Total price, including tax
Date	Date of purchase (Record available from January 2019 to March 2019)
Time	Purchase time (10 am to 9 pm)
Payment	Payment used by the customer for the purchase (3 methods are available – Cash/ Credit card and Ewallet)
COGS	Cost of goods sold
Gross margin percentage	Gross margin percentage
Gross income	Gross income
Rating	Customer stratification rating on their overall shopping experience (On a scale of 1 to 10)

From the data in Table 1 in this research, to utilize the analysis results of each attribute, which is interrelated with customer type data as class or label, we will first try to make feature selection with a correlation base to analyze the classification predictions. Feature selection is one data mining technique commonly used at the pre-processing stage. This technique reduces the complexity of attributes that will be managed in processing and analysis. (Adnyana, 2019).

After that, you will see each algorithm in feature selection, influencing how the SVM and Naïve Bayes algorithms work.

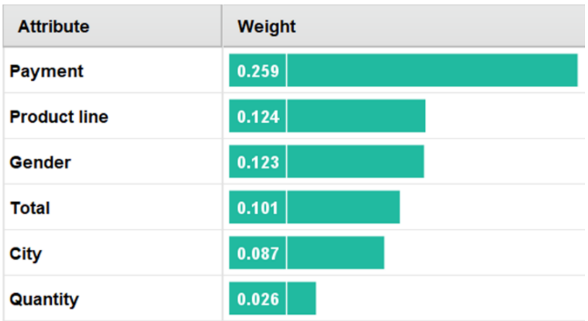


Attribute	Weight
Unit price	0.080
Total	0.076
Rating	0.039

Figure 3: Correlation Weight SVM

It can be seen from the correlation Weight Figure 3 for SVM that there are 3 attributes that I said are related to Unit Price, Total, and Rating, and the highest correlation is unit price.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v6i1.269
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023



Attribute	Weight
Payment	0.259
Product line	0.124
Gender	0.123
Total	0.101
City	0.087
Quantity	0.026

Figure 4: Correlation Weight Naïve Bayes

Meanwhile, for Naïve Bayes, there are 5 most significant attributes of payment, which can be seen in Figure 4 for the payment process, which may influence the customer type class. After that, in this research, we will test each algorithm by producing predictions in Table.

Table 2. Example Data Prediction

prediction (Customer type)	Customer type	confidence (Normal)	Confidence (Member)
Member	Member	0.487	0.513
Member	Normal	0.499	0.501
Member	Normal	0.499	0.501
Normal	Normal	0.503	0.497
Member	Member	0.493	0.507
Normal	Member	0.505	0.495
Member	Normal	0.499	0.501
Member	Member	0.499	0.501
Normal	Normal	0.504	0.496
Member	Member	0.499	0.501
Member	Normal	0.498	0.502
Normal	Member	0.518	0.482
Normal	Normal	0.501	0.499
Member	Normal	0.499	0.501
Member	Member	0.499	0.501
Member	Normal	0.494	0.506
Normal	Member	0.503	0.497
Member	Member	0.485	0.515
Member	Member	0.499	0.501
Member	Member	0.478	0.522

It can be seen from data table 2 that the overall predictions from the confusion matrix are as follows: One way is to apply the confusion matrix as a clarification model. The confusion matrix is used to obtain precision, recall, and accuracy values(Woro Isti Rahayu, Cahyo Prianto, 2021)

Table 3. Confusion Matrix SVM

	true Normal	true member	class precision
Pred. Normal	21	24	46.67%
pred. Member	121	120	49.79%
class recall	14.79%	83.33%	

Table 4. Confusion Matrix Naïve Bayes

	true Normal	true member	class precision
Pred. Normal	109	105	50.93%
pred. Member	28	44	61.11%
class recall	79.56%	29.53%	

Table 4 and Table 3 are the results of each confusion matrix from the SVM and Naïve Bayes algorithms using the same data from 1000 datasets to see the different results of the two algorithms for the confusion matrices. Confusion Matrix is a method usually used to calculate the level of accuracy in Data Mining for performance measurement and as a representation of the results of the classification process(Putri Ayu Mardhiyah, Riki Ruli A Siregar, 2020).

The results of this research using SVM and Naïve Bayes can be seen from the summary of the accuracy data and details of other evaluations in Table 5.

Table 5. Evaluation Prediction

Criterion	Value	
	SVM	Naïve Bayes
Accuracy	49.3%	53.5%
AUC	50.8%	51.4%
Precision	50.0%	60.4%
Recall	83.9%	29.3%
f_measure	62.3%	39.3%

The evaluation of two algorithms can be used to find results for classifying this data, which can influence the purchasing patterns of members and non-members so that they can provide appropriate recommendations for promotions, discounts, prices, and other influences. The growth of supermarkets is increasing, and there is high market competition. Supermarkets have a variety of products with different brands, various branches, and various types of customers(Wardhana et al., 2021).

CONCLUSIONS AND SUGGESTIONS

Conclusion

This research aims to look at predictions for data classification from customer type classes, which helps see member and non-member customer types that influence which attributes can be used as an important role so that in a strategy to increase sales better, knowledge can be used from member and class attribute data. Non-member, in terms of trying to use the results of supermarket sales where the attribute pattern is not much different, so it can be concluded that for this classification, supermarkets can be used which have data that is almost the same as the attributes from this research, in this research, we use classification with the SVM and Naïve algorithms Bayes so that you can compare the two algorithms so that with SVM the accuracy is 0.493 while for Naïve Bayes 0.535% is slightly better than Naïve Bayes. As explained in the introduction, no one has done previous research using this algorithm method, so the author tried to do this research. And done by selecting attributes.

Suggestion

For further research, perhaps you could try algorithms from other classifications so that they can vary, get better accuracy, and select the attribute data used.

REFERENCES

Achyani, Y. E. (2017). Prediksi Pemasaran Langsung Menggunakan Metode Support Vector Machine. *Jurnal Teknik Komputer*, *III*(2), 2–7. https://ejournal.bsi.ac.id/ejurnal/index.php/jtk/article/download/1719/1503

Adnyana, I. made B. (2019). Penerapan Feature Selection untuk Prediksi Lama Studi Mahasiswa. *Jurnal Sistem Dan Informatika*, *13*, 72–76. https://jsi.stikom-bali.ac.id/index.php/jsi/article/view/211/170

Akanmu Semiu Ayobami, S. R. (2012). Knowledge Discovery in Database : A knowledge management strategic Knowledge Discovery in Database : A knowledge management strategic approach. *Proceedings of 6th Knowledge Management International Conference (KMICe), July 2012*, 6–11. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2088965

Amsury, F., Ruhyana, N., & Mardiana, T. (2022). Comparison Of Classification Algorithms For Analysis Sentiment Of Formula E Implementation In Indonesia. *Jurnal Riset Informatika*, *4*(3), 291–298. https://doi.org/10.1109/siu.2012.6204469

Chrisdiyanti, I. N., Fa'rifah, R. Y., & ... (2023). Klasifikasi Review Customer Di E-Commerce Bukalapak Menggunakan Metode Support Vector Machine (SVM). *eProceedings ...*, *10*(3), 3200–3206. https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/20576%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/20576/19889

Indrayana, Y. K., Ramadhan, R. K., & ... (2023). Implementasi Decision Tree untuk Mengklasifikasikan Metode Pembayaran di Supermarket. *Prosiding Seminar ...*, *November*, 43–53. https://ojs.amikomsolo.ac.id/index.php/semnasa/article/view/86%0Ahttps://ojs.amikomsolo.ac.id/index.php/semnasa/article/download/86/5

Indriyani Indriyani, & Agus Bahtiar. (2023). Implementasi Data Mining Untuk Mengklasifikasikan Data Penjualan Pada Supermarket Menggunakan Algoritma Naïve Bayes. *Jurnal Manajemen Dan Bisnis Ekonomi*, *1*(1), 207–220. https://doi.org/10.54066/jmbe-itb.v1i1.70

Kojongian, V., Lapian, J., & Lumanauw, B. (2021). Pengaruh Bauran Pemasaran Terhadap Minat Beli Konsumen Di Cool Supermarket Tomohon. *Jurnal EMBA*, *9*(4), 811–820. https://ejournal.unsrat.ac.id/index.php/emba/article/view/36618

Liao, S. H., & Yang, L. L. (2020). Mobile payment and online to offline retail business models. *Journal of Retailing and Consumer Services*, *57*(151), 102230. https://doi.org/10.1016/j.jretconser.2020.102230

Maya Nur Annisaa, Nuryasman MN, & Ira Geraldina. (2023). Factors Affecting The Choice Of Payment Method In Modern Retail Shops. *Jurnal Ekonomi*, *28*(3), 327–348. https://doi.org/10.24912/je.v28i3.1780

Monika, I. P., & Furqon, M. T. (2018). Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *2*(10), 3165–3166. http://j-ptiik.ub.ac.id

Nengah Widya Utami, & I Wayan Budi Suryawan. (2021). Implementasi Data Mining Untuk Mengklasifikasikan Produk Pada Sebuah

Supermarket Mengunakan Algoritma Id3 Pada Orange. *Smart Techno (Smart Technology, Informatics and Technopreneurship)*, *3*(1), 33–36. https://doi.org/10.59356/smart-techno.v3i1.33

Noviyanto, N. (2020). Penerapan Data Mining dalam Mengelompokkan Jumlah Kematian Penderita COVID-19 Berdasarkan Negara di Benua Asia. *Paradigma - Jurnal Komputer dan Informatika*, *22*(2), 183–188. https://doi.org/10.31294/p.v22i2.8808

Nurelasari, E. (2019). *Segmentasi Dan Klasifikasi Perilaku Pembayaran Pelanggan Pada Perusahaan Multimedia Dengan Algoritma K-Means Dan C4.5*. *XXI*(1), 69–76. https://doi.org/10.31294/p.v20i2

Putri Ayu Mardhiyah, Riki Ruli A Siregar, P. P. (2020). Klasifikasi Untuk Memprediksi Pembayaran Kartu Kredit Macet Menggunakan Algoritma C4.5 Putri. *Jurnal Teknologia*, *3*(1), 91–101. https://aperti.e-journal.id/teknologia/article/view/66/44

Refo, Y., Rostianingsih, S., & Liliana, L. (2022). Penerapan SVM untuk Klasifikasi Sentimen pada Review Comment Berbahasa Indonesia di Online Shop. *Jurnal Infra*, *Vol 10*, *No*, 1–6. https://publication.petra.ac.id/index.php/teknik-informatika/article/view/12813%0Ahttps://publication.petra.ac.id/index.php/teknik-informatika/article/download/12813/11113

Rifka Agustianti, Pandriadi, Lissiana Nussifera, Wahyudi, L. Angelianawati, Igat Meliana, Effi Alfiani Sidik, Qomarotun Nurlaila, Nicholas Simarmata, Irfan Sophan Himawan, Elvis Pawan, Faisal Ikhram, Astri Dwi Andriani, Ratnadewi, I. R. H. (2022). Metode penelitian kuantitatif & kualitatif. In N. M. Ni Putu Gatriyani (Ed.), *Tohar Media* (Cetakan Pe, Nomor Mi). TOHAR MEDIA.

Riyadi, A. A., Amsury, F., Ruhyana, N., & Rahman, I. A. (2022). Implementation of the Association Method in the Analysis of Sales Data from Manufacturing Companies. *Jurnal Riset Informatika*, *5*(1), 593–598. https://doi.org/10.34288/jri.v5i1.491

Ruhyana, N., Mardiana, T., Amsury, F., & Sulistyowati, D. N. (2021). Classification of Student Satisfaction With Online Lecture. *Jurnal Riset Informatika*, *4*(1), 105–110. https://doi.org/10.34288/jri.v4i1.299

Wardhana, A. W., Patimah, E., Shafarindu, A. I., Siahaan, Y. M., Haekal, B. V., & Prasvita, D. S. (2021). Klasifikasi Data Penjualan pada Supermarket dengan Metode Decision Tree. *Senamika*, *2*(1), 660–667. https://conference.upnvj.ac.id/index.php/senamika/article/view/1389

Woro Isti Rahayu, Cahyo Prianto, E. A. N. (2021). Perbandingan Algoritma K-Means dan Naïve Bayes Untuk Memprediksi Prioritas Pembayaran Tagihan Rumah Sakit Berdasarkan Tingkat Kepentingan Pada PT. PERTAMINA (PERSERO). *Jurnal Teknik Informatika*, *13*(2), 1–8. https://ejurnal.poltekpos.ac.id/index.php/informatika/article/view/1383/809