Electronic Integrated Computer Algorithm Journal

Vol. 1, No.1, October 2023, pp 9~15

ISSN: 3031-0350 , DOI: 10.62123/enigma.v1i1.9

Timeliness Of Spp Payments At Smk Tritech Infomatika Using Naive Bayes Algorithm

Muhammad Fakhri Al-Zikri[1*] , Rial Beimar Volado Sibuea[1] , Tri Krisandi Silalahi[1]

[1]Department of Computer Engineering and Informatics, Politeknik Negeri Medan, 20155, Indonesia

DOI: 10.62123/enigma.v1i1.9

Received : September 13, 2023
Revised : September 24, 2023
Accepted : October 01, 2023

Keywords:
Rapid Miner, SMK Tritech Informatika, Naive Bayes Method

ABSTRACT

SMK Tritech Informatika is one of the vocational schools located on Jalan Bhayangkara Medan. Smk Tritech plays an important role in education. However, many schools are late in financing their operations. So there are many problems related to school operational payments such as tuition payments. SPP payment is an important problem at Smk Tritech Infomatika because many of the students are late in paying SPP even though the payment deadline has been set. Therefore, it is necessary to evaluate the SPP payment. To overcome this problem, it is necessary to detect the factors that cause late tuition payments using data mining. The data mining technique used is classification with the Naive Bayes algorithm method.

1. INTRODUCTION

SMK Tritech Informatika is one of the private educational institutions that focuses on vocational education. The cost of education is one of the supporting inputs for the implementation of education. This cost plays a very important role in achieving good education. In this case, one of the education costs found at Smk Tritech Informatika is the student fee that must be paid every month or better known as the Education Development Contribution (SPP). This SPP fee is generally applied by private institutions / schools which are charged to each student. Because private schools in the management of education are charged to the community or local policy. In contrast to public schools whose costs are generally borne by the government. So in this case the school charges tuition fees to student guardians. for each month. Which aims for the continuity of education atSmk Tritech Informatika[1][2].

In this case, the problem that arises regarding tuition payments is if students are late in paying tuition fees from the specified time[3][4]. This is a problem because tuition payments are one of the important factors in providing good service quality for the school. Quoting data from interviews with school caregivers, that students who are late in paying tuition fees[5][6]. Generally, the factors that cause the above problems are the economic factors of income of parents or guardians of students, and also the habit of students who slow down payments even though they have been given money by their parents, and often students use money that should be used to pay for the pesantren, used for their personal expenses[7][8]. This is a difficult problem because many students are late in paying their tuition fees, causing the school to have less income. Meanwhile, these costs are needed for the continuity of school education, such as paying teacher salaries and paying for school needs[9]. So there needs to be a solution to this problem by classifying payments based on the level of timeliness of payment. So that it can be an evaluation material for the school to increase tuition payments at the specified time[10][11]. The purpose of this research is to classify tuition payments based on the timeliness of payments using the Naive Bayes algorithm. The Naive Bayes algorithm is one of the Top 10 Algorithms in Data Mining which was published in December 2006 by the IEEE International Conference on Data Mining[12][13]. Naïve bayes merupakan algoritma dalam melakukan yang berakar dari teorema probabilitas Bayes. This algorithm is often used in machine learning for classification, namely predicting the category or class of a sample based on its features. The Naive Bayes algorithm assumes that each feature is conditionally independent of its class[14][15]. Although these assumptions are often not completely true in real-world contexts, the Naive Bayes algorithm remains quite simple, fast, and often quite effective. Naïve Bayes is included in the subfield of data mining so that many cases in data mining use the Naïve Bayes algorithm. Data mining is the process of collecting information and patterns that can be used as needed on large data sets. So this research will utilize the data mining process and the Naïve Bayes algorithm[16][17].

2. LITERATURE REVIEW

This study describes that the purpose of this research is to apply the Naive Bayes method with the Information Gain method for late school payment time. This research is conducted to predict late school payments based on data information on parents' income, parents' education, age, and also parents' dependents. This prediction is expected to be a solution to optimise late school payments. This research analyses late payment data by predicting the level of late payment using the Naive Bayes algorithm and also Information Gain with the results of an accuracy rate of 80% using only the Naive Bayes algorithm and 90% with the Information Gain method.

*Corresponding Author Email: fakhrimpolmed10@gmail.com

2.1. Data Mining

Data mining is a computational process for discovering patterns in large data sets. An overview of the data mining process as knowledge discovery in databases includes: the use of algorithms, statistical tools, and machine learning to extract previously unknown patterns[18]. Data mining supports the data analysis process by identifying clusters, detecting anomalies, discovering dependencies, and finding correlations[19].

2.2. Classification

Classification is also one of the data mining methods or techniques[20][21]. The definition of classification is a job of assessing data objects that aim to be included in certain classes from a number of available classes. In classification there are two main jobs, namely the construction of a model as a prototype to be stored as memory and the use of the model is used for recognition/classification/prediction on another data object so that it is known which class the data object is stored in the model that has been built[22]. In the classification method there are several phases of completion, starting from training data and ending with the data testing process so that an accurate decision is produced. The following is a picture of the solution flow of the Classification method[23].

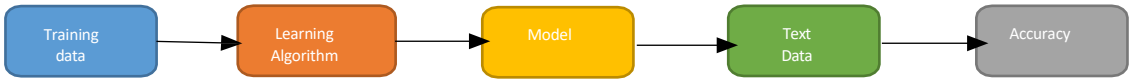

Figure 1. Classification Method Solving Flow

2.3. Payment of tuition fees

SPP payment is a monthly payment or can also be interpreted as payment of the operational costs of the institution / education that must be paid monthly. SPP can be interpreted as an educational development contribution paid by students in schools. The purpose of the SPP is so that the institution/education can finance educational operations and also finance educational facilities so that the institution/school can carry out better learning activities.

2.4. Naive Bayaes Algorithm

Algorithms are techniques for organising stages to solve problems in the form of sentences with a limited number of words, arranged logically and systematically. Algorithms are also often defined as a procedure for solving problems using certain steps and limited in number. The accuracy test used in the Naive Bayes Algorithm generally uses the Confusion Matrix method with the following formula[24].

Table 1. Confusion Matrix Formula

Correct Classification	Classifields as (+)	Classifields as (-)
+	True Positive (TP)	False Negative (FN)
-	False Postive (FP)	True Negative (TN)

The following is an explanation of the confusion matrix formula table:

- *Precision* is used to measure how large the proportion of positive classes that are successfully predicted correctly from the entire postive class, which is calculated using the formula :

  Precession $\frac{TP}{TP+FP}$ (1)

- *Recall* is used to show the percentage of positive data classes that are successfully predicted correctly from all positive class data, which is calculated using the formula:

  Recall $\frac{TP}{TP+FN}$ (2)

- *Accuracy* is the sum of the ratio of correct data to the total amount of data.

2.5. RapidMiner

RapidMiner is software used for knowledge discovery that has approximately 400 data mining operators, including input operators, output operators, data preprocessing and visualisation. Another definition of RapidMiner is data processing software using data mining principles. RapidMiner extracts patterns from large data sets by combining statistical methods, artificial intelligence and databases. RapidMiner can also be defined as tools used in techniques that are in the environment of machine learning, data mining, text mining, and predictive analytics[25]. The advantage of RapidMiner in data mining processing is that it is very easy to use to calculate a lot of data by using operators. This operator functions to modify data, where the data is connected to the nodes on the operator, then the user only needs to connect to the node to see the results. The results that RapidMiner displays visually with a graph. With these advantages, making RapidMiner the software of choice for extracting data with data mining methods[26].

3. RESEARCH METHODS

3.1 Research Flow

The research flow used in this study will be explained in the following figure:

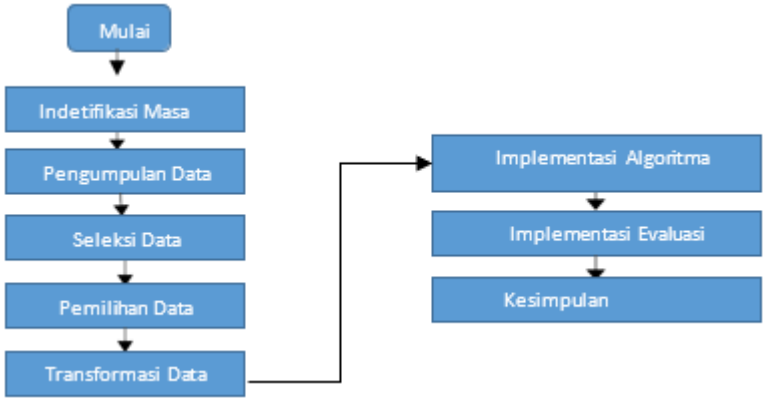

Figure 2. Research Flow

Based on the picture above, the research flow can be explained as follows:

1. Identify the problem, determine the background of the problem, parameters and solution of the problem
2. Data collection, collecting by using observation, interviews and literature studies. So as to produce SPP payment data provided by the cottage both primary data and secondary data and also literature related to SPP.
3. discussion in this research.

Table 2. Primary data of tuition payment

No.	Full Name	Gender	Address	Ket. Payment
1	Aldi Wiguna	L	Jln.Bromo	EXACTLY
2	Ardika Kuswahyudi	L	Jln.Tembung	EXACTLY
3	Dandy Prawira	L	Jln.Denai	EXACTLY
4	Dicky Syahbana	L	Jln.Pancing	LATE
5	Farah Nabila	P	JLn.Cemara	LATE
6	Hamdoko	L	Jln.Lau Dendang	EXACTLY
7	Hery Sugianto	L	Jln.Sutrisno	LATE
8.	Kevin Mahmuhadi	L	Jln.Sutomo	EXACTLY
9.	Muhammad Daffa	L	Jln.Pancasila	EXACTLY
--	--	--	--	--
--	--	--	--	--
49	Daffa Ahmad	L	Legendary street	EXACTLY
50.	Primus Raihan	L	Jln Rrq	LATE

Table 3. Secondary Data of Tuition Payment

No.	Full Name	Father's Name	Mum's name	Jobs	Revenue
1	Aldi Wiguna	Ulul Absor	Neli Rahmawati	Entrepreneurship	5 million
2	Ardika Kuswahyudi	Ade Juluri	Carsih	Entrepreneurship	2Million
3	Dandy Prawira	Yudianto	Dawn	PNS	2Million
4	Dicky Syahbana	Rustayim	Farida Hanum	Entrepreneurship	3 million
5	Farah Nabila	Gustomi	Artati	Entrepreneurship	3Jua
6	Hamdoko	Carlam	Yati	PNS	2Million
7	Hery Sugianto	Ahmad Yani	Diani	Entrepreneurship	3 million
8.	Kevin Mahmuhadi	Jumena	Sumiyati	Entrepreneurship	5 million
9.	Muhammad Daffa	Januardi	Eka	Police	5 million
10	Muhammad Ridwan	Supriadi	Yuli	PNS	4Million
--	--	--	--	--	--
--	--	--	--	--	--
49	Daffa Ahmad	Rengoku	Hima	Police	5 million
50	Primus Raihan	horas	Rista	Entrepreneurship	4Million

4. Data *Selection*, selecting data that has been collected to be used as a data mining process separated from operational data or datasets.
5. *Preprocessing/Cleaning*, discarding unused data and also at this stage includes checking the data and correcting the data if there are errors, such as *typographical* errors.

Table 4. Data Processing Results

No.	Full Name	Jobs	Revenue	Ket. Payment
1	Aldi Wiguna	Entrepreneurship	5 million	Exactly
2	Ardika Kuswahyudi	Entrepreneurship	2Million	Exactly
3	Dandy Prawira	PNS	2Million	Exactly
4	Dicky Syahbana	Entrepreneurship	3 million	Late
5	Farah Nabila	Entrepreneurship	3Jua	Late
6	Hamdoko	PNS	2Million	Exactly
7	Hery Sugianto	Entrepreneurship	3 million	Late
8.	Kevin Mahmuhadi	Entrepreneurship	5 million	Exactly
9.	Muhammad Daffa	Police	5 million	Exactly
10	Muhammad Ridwan	PNS	4Million	Exactly
49	Daffa Ahmad	Police	5 million	Exactly
50	Primus Raihan	Entrepreneurship	4Million	Late

6. In this phase, the process of transforming data forms that do not have clear entities into valid or ready data forms for the Data Mining process is carried out.
7.

Table 5. Training Data

Jobs	Revenue	Payment terms
Entrepreneurship	5 million	Exactly
Entrepreneurship	2Million	Exactly
PNS	2Million	Exactly
Entrepreneurship	3 million	Late
Entrepreneurship	3Jua	Late
PNS	2Million	Exactly
Entrepreneurship	3 million	Late
Entrepreneurship	5 million	Exactly
Police	5 million	Exactly
PNS	4Million	Exactly
Police	5 million	Exactly
Entrepreneurship	4Million	Late

8. Data mining implementation, in this phase data analysis is carried out using the *Naive Bayes Classifier* Algorithm. So as to produce a model and level of accuracy in the application of the algorithm.
9. Interpretation/Evaluation, in the last phase carried out is the process of forming conclusions from the results that have been obtained.

4 RESULTS AND DISCUSSIONS

4.1 Testing Results

In the results of data testing there are several *outputs* generated by *RapidMiner software*, namely:
Figure 3. It can be explained that the Naïve Bayes calculation process using Rapidminer tools is first carried out by importing training data and testing data, then the data is connected to the *tools* operators in RapidMiner as shown above. Next, the run process is carried out.

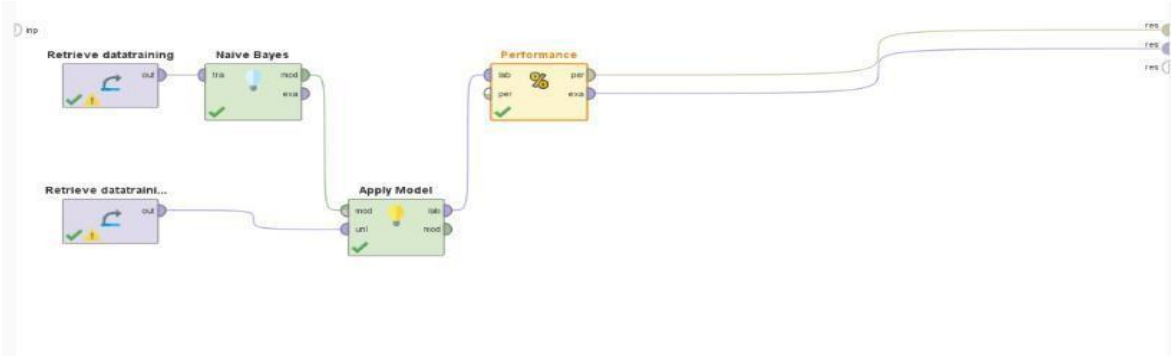

Figure 3. Rapidminer Naive Bayes Operator Configuration Process

After running, the classification results will appear on the example set data. The following are the results of the naive bayes calculation using RapidMiner *tools*.

Table 6. Classification results

Jobs	Revenue	Ket. Payment	Confidence	Confidence	Prediction
Entrepreneurship	5 Million	Exactly	1.0	0.0	Exactly
Entrepreneurship	2Million	Exactly	1.0	0.0	Exactly
PNS	2Million	Exactly	1.0	0.0	Exactly
Entrepreneurship	3 Million	Late	0.4	0.6	Late
Entrepreneurship	3Jua	Late	0.0	1.0	Late
PNS	2Million	Exactly	1.0	0.0	Exactly
Entrepreneurship	6 Million	Exactly	0.5	0.5	Exactly
PNS	3 Million	Exactly	0.6	0.4	Exactly
Police	5 Million	Exactly	1.0	0.0	Exactly
Farmers	4Million	Exactly	1.0	0.0	Exactly
Police	5 Million	Exactly	1.0	0.0	Exactly
Entrepreneurship	4Million	Late	0.6	0.4	Exactly

The results of the SPP payment data classification process with Ms. Excel *tools* and RapidMiner *tools are the* SAME.


accuracy: 76.00%

	true TEPAT	true TERLAMBAT	class precision
pred. TEPAT	31	9	77.50%
pred. TERLAMBAT	3	7	70.00%
class recall	91.18%	43.75%	

Figure 4. Results of Rapidminer Tools

In Figure 4. is the result of testing with RapidMiner *tools* with the number of *True Positive* is 11 data classified as pred.TEPAT and *class* TEPAT, *False Positive is* 6 data classified as pred.TEPAT but *class* TERLAMBAT, *True Negative is* 3 data classified as pred. SLOWEST and *class* SLOWEST, and *False Negative as* much as 2 data is classified as pred. SLOWEST but *class* FAST. Therefore, it can be concluded that the classification results of tuition payment data using RapidMiner and Ms. Excel *tools* are the SAME.

4.2 Accuracy Level of Algorithm Model Implementation

The method of testing the accuracy level used is the *confusion matrix* method which consists of *precission, recall,* and *accuracy*. *Confusion matrix* testing for testing data processed using Ms. Excel *tools* for accuracy values can be seen in the following table:

Table 7. Testing Results Confusion Matrix tools Ms. Excel

Class	Classification		Precision
Exactly		Late	
Pred	31	9	77,50%
Pred Late	3	7	70,00%
Recall	91,18%	43,75%	

Acuracy	73,35%

In applying the Performance operator, not only the confusion matrix is generated. However, it also produces an *Area Under Curve* (AUC) which will be displayed in the following *Receiver Operating Characteristic* (ROC) curve image:

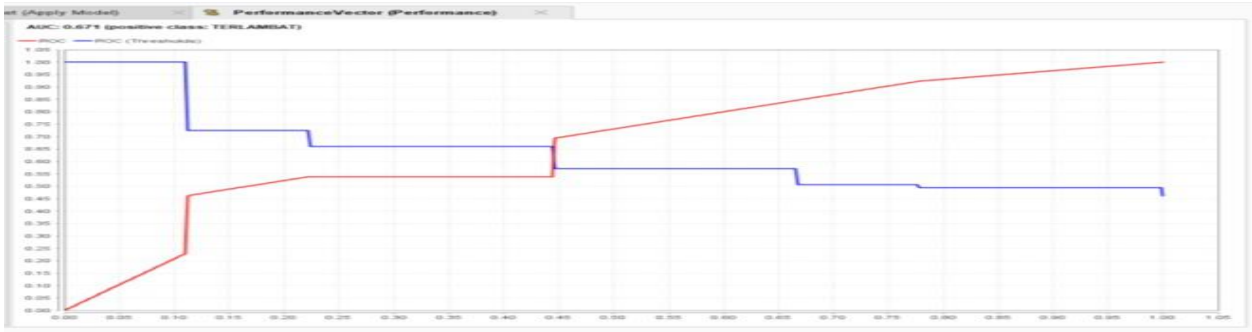

Figure 5. Receiver Operating Characteristic (ROC) curve

Receiver Operating Characteristic (ROC) curves are used to express the *confusion matrix* data. The horizontal line represents the *False Positive* (FP) value and the vertical line represents the *True Positive* (TP) value. From Figure 5. it can be seen that the *Area Under Curve* (AUC) value of the *Naive Bayes* algorithm model is 0.671 with the SLOWEST *positive class*, AUC (*optimistic*) is 0.718 with the SLOWEST *positive class*, and AUC (*pessimistic*) is 0.624 with the SLOWEST *positive class*.

4.3 DISCUSSION

At this stage is a discussion of the results of the Naive Bayes Algorithm testing that has been carried out on Ms. Excel *tools* and RapidMiner *tools*. The following is an explanation of this research

1. In this study the authors used 50 datasets consisting of primary data and secondary data. Of the 50 data, 50 data are used as training data and from the training data there are 50 data used as testing data which are then used as tests in this study both using manual *tools* Ms. Excel and with RapidMiner *tools*.
2. For the *Class Prediction* section using 50 data on *testing* data. There is matching data between the Right class and pred.Tepat which is 31 Data (*True Positive*) and for Slow class data that matches pred.Terlambat as much as 7 data (*True Negative*) and data for the Right class but pred.Terlambat as much as 9 data (*False Negative*) and data for the Late class but pred.Tepat as much as 3 data (*False Postive)*.
3. The level of *accuracy* in manual calculations in Ms. Excel tools and calculations using RapidMiner *tools is the* same, which is 73.35%. *Precision* for the FAST class is 77.50%, while for the SLOWEST class is 70.00%. And *Recall* for the FIT class is 91.18%, while for the SLOWEST class is 43.75%.
4. Factors that are problematic for the late payment of tuition fees are related to the economic status of parents, as well as parents' occupation and gender.

5 CONCLUSSION

Based on the test results and discussion in this study, it can be concluded as follows:

1. *The Naive Bayes* algorithm can be used to classify tuition fee payment data based on the timeliness of payment.
2. Of the 50 testing data used, 31 data can be classified as *True Positive*, 7 data as *True Negative*, 9 data as *False Negative*, and 3 data as *False Positive*.
3. The Naive Bayes algorithm in predicting whether students pay tuition fees correctly or late has an *accuracy* rate of 73.35%, *precision* in the *ACCURATE* class is 77.50% while in the TARDY class it is 70.00%, and *recall* in the ACCURATE class is 91.18% while in the TARDY class it is 43.75%.
4. Factors that influence the level of punctuality of tuition payments at Smk Tritech Informatika are not only from the factors of parental employment and parents' monthly income, but external factors also influence.

REFERENCES

[1] I. M. Talha, I. Salehin, S. C. Debnath, M. Saifuzzaman, N. N. Moon, and F. N. Nur, "Human behaviour impact to use of smartphones with the python implementation using naive Bayesian," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, 2020, pp. 1–6.

[2] M. Tabash, M. Abd Allah, and B. Tawfik, "Intrusion detection model using naive bayes and deep learning technique.," *Int. Arab J. Inf. Technol.*, vol. 17, no. 2, pp. 215–224, 2020.

[3] S. Widaningsih, "Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4,5, Naïve Bayes, Knn Dan Svm," *J. Tekno Insentif*, vol. 13, no. 1, pp. 16–25, 2019, doi: 10.36787/jti.v13i1.78.

[4] A. R. Lubis, M. K. M. Nasution, O. S. Sitompul, and E. M. Zamzami, "The effect of the TF-IDF algorithm in times series in forecasting word on social media," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 22, no. 2, p. 976, 2021, doi: 10.11591/ijeecs.v22.i2.pp976-984.

[5] 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.

[6] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowledge-Based Syst.*, vol. 192, p. 105361, 2020.

[7] Sarwo and S. Aisyah, "Penerapan Data Mining Menggunakan Algoritma Naive Bayes Classifier Untuk Memberikan Rekomendasi Bermain Golf Pada PT. Asiamadya Selaras," *J. Teknol. Pelita Bangsa*, vol. 6, no. 2, pp. 99–104, 2017.

[8] P. Koukaras, C. Tjortjis, and D. Rousidis, "Mining association rules from COVID-19 related twitter data to discover word patterns, topics and inferences," *Inf. Syst.*, vol. 109, p. 102054, 2022.

[9] 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)*, IEEE, 2021, pp. 41–44.

[10] G. Nguyen *et al.*, "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 77–124, 2019, doi: 10.1007/s10462-018-09679-z.

[11] S. Theodoridis, "Chapter 12 - Bayesian Learning: Inference and the EM Algorithm," S. B. T.-M. L. (Second E. Theodoridis, Ed., Academic Press, 2020, pp. 595–646. doi: https://doi.org/10.1016/B978-0-12-818803-3.00023-4.

[12] 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)*, IEEE, 2020, pp. 324–326.

[13] 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)*, IEEE, 2020, pp. 217–220.

[14] S. Ramadani, I. Ambarita, and A. M. H. Pardede, "Metode K-Means Untuk Pengelompokan Masyarakat Miskin Dengan Menggunakan Jarak Kedekatan Manhattan City Dan Euclidean (Studi Kasus Kota Binjai)," vol. 04, no. 2, pp. 15–29, 2019.

[15] U. Yabas, H. C. Cankaya, and T. Ince, "Customer churn prediction for telecom services," *Proc. - Int. Comput. Softw. Appl. Conf.*, pp. 358–359, 2012, doi: 10.1109/COMPSAC.2012.54.

[16] B. Zhu, X. Wu, L. Yang, Y. Shen, and L. Wu, "Automatic detection of books based on Faster R-CNN," in *2016 third international conference on digital information processing, data mining, and wireless communications (DIPDMWC)*, IEEE, 2016, pp. 8–12.

[17] Z. Y. Shu, "The study and application of the technology of data warehouse and data mining in the library," *2011 Int. Conf. Electr. Inf. Control Eng. ICEICE 2011 - Proc.*, pp. 4671–4673, 2011, doi: 10.1109/ICEICE.2011.5778341.

[18] A. F. Firdaus, R. Saedudin, and R. Andeswari, "Implementasi Metode Klasifikasi Naive Bayes Dalam Memprediksi Kelulusan Mahasiswa," *e-Proceeding Eng.*, vol. 8, no. 5, pp. 9274–9279, 2021.

[19] L. M. Jose and K. Rahamathulla, "A semantic graph based approach on interest extraction from user generated texts in social media," *Proc. 2016 Int. Conf. Data Min. Adv. Comput. SAPIENCE 2016*, pp. 101–104, 2016, doi: 10.1109/SAPIENCE.2016.7684118.

[20] M. Abdel-Basset, M. Mohamed, F. Smarandache, and V. Chang, "Neutrosophic association rule mining algorithm for big data analysis," *Symmetry (Basel).*, vol. 10, no. 4, p. 106, 2018.

[21] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text mining in big data analytics," *Big Data Cogn. Comput.*, vol. 4, no. 1, p. 1, 2020.

[22] M. Bouazizi and T. Ohtsuki, "Opinion mining in Twitter: How to make use of sarcasm to enhance sentiment analysis," *Proc. 2015 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2015*, pp. 1594–1597, 2015, doi: 10.1145/2808797.2809350.

[23] S. Sagadevan, N. H. A. H. Malim, and M. H. Husin, "A Seed-Guided Latent Dirichlet Allocation Approach to Predict the Personality of Online Users Using the PEN Model," *Algorithms*, vol. 15, no. 3, 2022, doi: 10.3390/a15030087.

[24] I. I. Kholod, "Conditions for parallel execution of functions in data mining algorithm," *Proc. 2018 IEEE Conf. Russ. Young Res. Electr. Electron. Eng. ElConRus 2018*, vol. 2018-Janua, pp. 308–312, 2018, doi: 10.1109/EIConRus.2018.8317094.

[25] M. Watts and N. Kasabov, "Evolutionary optimisation of evolving connectionist systems," *Proc. 2002 Congr. Evol. Comput. CEC 2002*, vol. 1, no. June 2014, pp. 606–610, 2002, doi: 10.1109/CEC.2002.1006995.

[26] A. Karimi, L. Rossi, and A. Prati, "Aeda: An easier data augmentation technique for text classification," *arXiv Prepr. arXiv2108.13230*, 2021.