Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 7, No. 2, September 2024, pp. 553 – 560
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Analysis of Student Dropout Potential Using the Multinomial Naive Bayes Algorithm

[*1]Dewi Afrianti, [2]Armansyah
[1,2]Computer Science Study Program, Faculty of Science and Technology,
Universitas Islam Negeri Sumatera Utara, Indonesia
E-mail:[1] dewiyapimafriyanti@gmail.com, [2]armansyah@uinsu.ac.id

Article Info
Article history:
Received Aug 2nd, 2024
Revised Sep 17th, 2024
Accepted Sep 30th, 2024

Keyword:
Accuracy
Classification Model
Multinomial Naive Bayes
Potential Dropout
Supporting Factors

ABSTRACT
The current situation of education in Indonesia is quite concerning, especially with the high dropout rate which is one of the main problems. The variation in dropout rates in various educational institutions, including at Muhammadiyah 9 Vocational High School in Medan, reflects the diversity of challenges faced. This study aims to analyze the supporting factors that influence the potential for student dropout using the Multinomial Naive Bayes method, especially at this school. The results of the study showed that the model could understand the data with a classification performance accuracy of 83.04% at the 20% dataset testing stage. Through this test, 76 active students, 11 students with the potential to drop out, and 25 students dropped out were obtained. Meanwhile, precision, recall, and f1-score in the class with the potential to drop out cannot be displayed because the class comparison is unbalanced.

Corresponding Author:
Dewi Afrianti,
Computer Science Study Program, Faculty of Science and Technology,
State Islamic University of North Sumatra,
Lap. Golf No.120, Kp. Tengah, Kec. Pancur Batu, Kabupaten Deli Serdang, Sumatera Utara 20353
Email: dewiyapimafriyanti@gmail.com

DOI: http://dx.doi.org/10.24014/ijaidm.v7i2.32316

1. INTRODUCTION

Dropping out of school refers to individuals who have left school before completing their education, or can also be interpreted as school-age children who do not attend school and do not have a diploma [1][2]. Factors that cause children to drop out of school include the inability to follow or remember lessons, lack of interest and motivation to go to school, family economic conditions, neglect from parents, and an unsupportive play environment [3]. According to Ki Hajar Dewantara, education is a basic need in a child's growth. Education functions to guide children to become good individuals and happy members of society. This process aims to develop human potential and respect the human rights of every individual. Students are not machines that can be controlled, but rather a very valuable generation [4]. Every individual in this country has the right to education as regulated in Article 31 Paragraph 1 of the 1945 Constitution. Basic education is a right that must be fulfilled by the state, and the cost of basic education is the government's obligation (Article 31 Paragraph 2). Article 31 Paragraph 3 stipulates that the government must organize a national education system to improve the quality of life and morals of the community. titled paper may never reach the audience for which it was intended, so be specific [5].

Looking at the current state of education in Indonesia, the situation is quite concerning due to the various problems faced. One of the main problems is the high number of children dropping out of school [6]. The number of dropouts varies in each school, including at the Muhammadiyah 9 Medan Vocational High School (SMK). The number of dropouts reached 232 students, dropping out 7% of the 2,924 active students, 93% taken from student data for the 2017/2018 academic year to the 2023/2024 academic year.

Several factors causing children to drop out of school include low economic conditions, limited parental educational background, lack of attention from parents, lack of interest and motivation to learn, an

unsupportive friendship environment, inadequate educational facilities and infrastructure, and an education system that does not meet students' needs [7][8][9].

The factors causing children to drop out of school vary in each school. Therefore, an in-depth analysis of these factors is needed to understand the main causes and find the right solutions. One method that can be used is the classification approach. Classification is the process of creating a model that aims to identify categories or classes from a new dataset that was previously unknown [10][11].

Classification can offer solutions to problems such as processing, organizing, and analyzing large amounts of unstructured data. To analyze the factors that influence children dropping out of school, one technique that can be used is the Naïve Bayes Multinomial classification. This technique allows to identify categories such as active students, students who have the potential to drop out, and students who have dropped out. Thus, this analysis can provide clearer insights into the status of students and the factors that influence them [12][13].

Naïve Bayes Multinomial is one of the algorithms in machine learning that is used for classification. In the context of education, this algorithm can be used to predict the potential for students to drop out of school. Naïve Bayes Multinomial is an effective tool for analyzing the risk of student dropout due to its simplicity and efficiency, as well as its ability to provide fairly accurate results[14] [15]. The advantages of the Multinomial Naive Bayes algorithm, including simplicity, efficiency, ability to handle categorical data, tolerance to imbalanced data, and ease of implementation, make it an ideal choice for this study. This algorithm is effective in identifying students at high risk of dropping out, even when faced with diverse and imbalanced data [16].

To support this research, it is important to refer to relevant previous studies. One of them is a study by Ayuni et al., 2023 This study shows that the Multinomial Naive Bayes algorithm is effective in classifying news topics about Indonesia with the highest accuracy reaching 88.3%. In the study, which collected 682 news data, it was found that the dominant topics included politics, socio-culture, and health [17]. The second study by Mulyani et al., 2021 entitled In this study, book data was obtained from the Indramayu State Polytechnic Library and the Indramayu Regency Regional Library. This study examines the effect of Unigrams, Bigrams, and Trigrams on book title classification using the Multinomial Naive Bayes algorithm. The results of the study showed that the Unigram method provided the highest classification accuracy of 74.4% [18].

The second study by Ige & Adewale, 2022 on examined cyberbullying data from 2019 and found that 95% of teenagers in the US were involved in cyberbullying cases. The results of validation and cross-validation showed that the model used achieved 92% accuracy, with low bias but high variance. In addition, the model also showed a very low mean square error (MSE) [19]. The third study conducted by Vol, 2022 From this study, it was found that an average of 256 students per year did not drop out, while 34 students per year dropped out. This model achieved an average accuracy of 98.745% per year, with an average classification error of 1.255% per year. A total of 1,178 student grade data from SMK Negeri 2 Kotabumi Lampung Utara were collected for this study, which confirmed that an average of 256 students did not drop out and 34 students dropped out each year [20]. The next study, as mentioned Herwanto et al., 2021 The test results showed that the Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM) models achieved the highest precision value of 93%. The SVM model showed the highest recall and f1-score values, each at 94%. In addition, the SVM model also achieved the highest accuracy value of 95%. The MNB model had the fastest testing time, which was 2.66 ms [21].

This study is expected to be able to produce a more accurate and efficient predictive classification model, which can support dropout prevention efforts. Thus, the results of this study can help schools in identifying students at high risk of dropping out of school and designing appropriate and suitable interventions.clear background, a clear statement of the problem, the relevant literature on the subject, the proposed approach or solution, and the new value of research which it is innovation. It should be understandable to colleagues from a broad range of scientific disciplines.

2. RESEARCH METHOD

In the research framework, research stages are needed that contain a research model. This is done so that the research stages are structured. The method in this research is quantitative because it is related to numbers and statistics. The following is the research process applied to figure 1.

2.1 Data collection

Data collection through interview techniques is the initial step in basic research conducted by researchers to identify potential problems that can be addressed with a research approach. In this study, interviews covered a variety of topics, including the number of students who dropped out, follow-up in handling dropout cases, direct observation in the field, and related literature reviews.

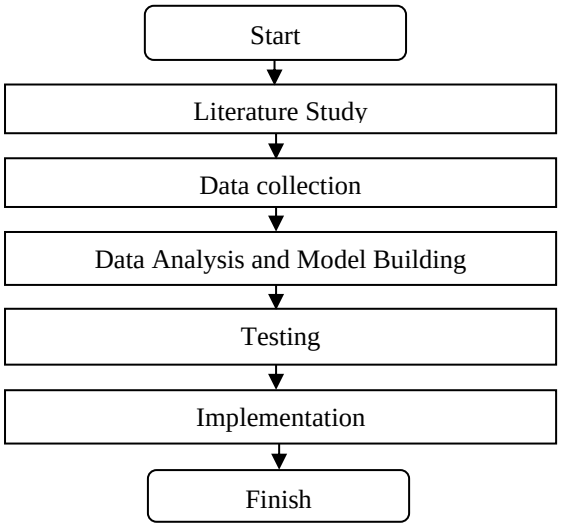

Figure 1. Research framework

2.2 Data analysis and model development

This study began with data collection from Muhammadiyah 9 Medan Vocational High School (SMK). The data that had been collected was then analyzed to group students into three categories: active students, students who had the potential to drop out, and students who had dropped out, based on the available attributes. After the data analysis stage was complete, the next step was model development. In this study, researchers adopted the Multinomial Naive Bayes approach as the main framework. The following is a flowchart of Multinomial Naive Bayes, can view figure 2.

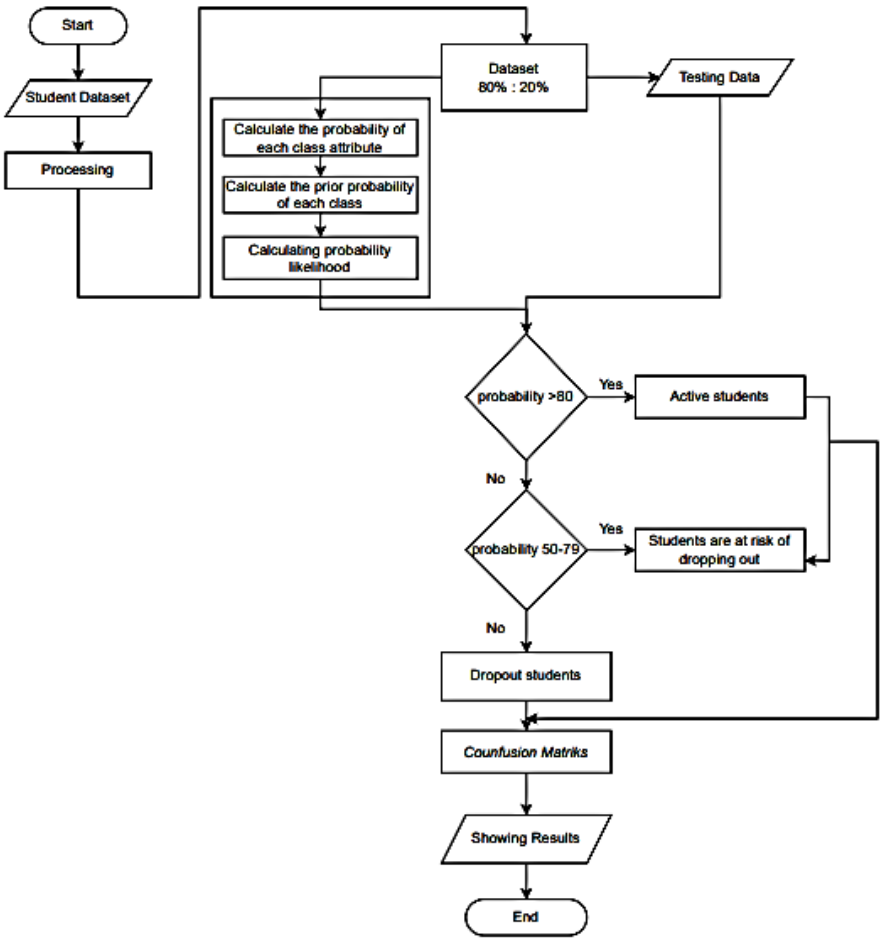

Figure 2. Naïve Bayes Multinomial Flowchart

This method was chosen because of its ability to handle categorical data as often encountered in the analysis of student dropout potential. The steps taken by the author include:

1. Data preparation
2. Data processing includes data normalization if necessary and changing categorical variables to numeric representations.
3. Feature selection includes Identifying relevant features for predicting dropout potential and Performing statistical analysis to select the most influential features.
4. Data division includes training data and testing data Calculate class and feature probabilities for the model.
5. Model training includes calculating class and feature probabilities for the model.
6. Model evaluation, namely evaluating the model using metrics such as accuracy, precision, recall.
7. Result interpretation, namely analyzing model results to gain an understanding of the factors that influence student dropout potential.

2.3 Testing

In the framework of this study, researchers will test the attributes and classes used in the predictive model to understand their impact on overall performance. The researcher will identify the most significant attributes in influencing the prediction results, as well as evaluate the class diversity that can affect the accuracy and reliability of the model. This approach will provide deeper insight into the factors that contribute to model performance, and allow for the development of more effective and reliable models.

2.4 Implementation

The results of the Multinomial Naive Bayes model were analyzed to identify the most influential attributes in predicting dropout potential. This analysis provides insight into the factors that contribute to students' decisions to drop out of school. The findings of this study can be used as a basis for developing effective intervention strategies to prevent dropout in educational settings.

3. RESULTS AND ANALYSIS

The results of the analysis and discussion related to the design of a model to analyze student dropout potential using Python and the Jupyter Notebook tool will be presented in stages, following the Multinomial Naive Bayes method, with the following results.

3.1. Data collection

In this study, data were collected through direct observation at the Muhammadiyah 9 Medan Vocational High School (SMK). The data used includes academic and biographical information of students. The Multinomial Naive Bayes method is applied to analyze the potential for student dropout. This method utilizes Bayes' theorem with the assumption of feature independence to predict the likelihood of students being in active, potential dropout, or dropout status. It is hoped that this classification model will provide more accurate and efficient insights in preventing dropout and assist school policies in designing appropriate interventions.

Table 1. Student dataset

No Sort	Gender	N_Average	Origin_SMP	Wil_Origin_Smp	Address	Access Distance	Label
1	MAN	77	SMP DAYA CIPTA	Medan	JL. JANGKA NO. 94B	Currently	active
2	MAN	87	MTS AL WASHILIYAH MEDAN KRIO	Sunggal	JLN PAYAGELI	Currently	active
3	MAN	77	SMP NURCAHAYA	Medan	JL. BUNGA CEMPAKA GG. GILINGAN PADI	Currently	active
4	MAN	85	SMP SWASTA NURCAHAYA	Medan	JL BARU PERTAMBANGAN PSR II TJ SARI	Currently	active
…	…	…	…	…	…	…	…
556	MAN	24	MTS ISLAMIYAH SUNGGAL	Sunggal	JL. SETIA BANGUN DUSUN IV SUNGGAL KANAN	Currently	DropOut

Table 2. Class probability

Status	Amount
Active	358
Potential Dropout	55
Dropout	143

The total number of students enrolled, there are 358 active students, 55 students who are at risk of dropping out, and 143 students who have dropped out. This shows that the majority of students are currently active, but there are a number of students who are at high risk of dropping out. The number of students who have dropped out is also quite significant, indicating the need for more attention in efforts to overcome dropout and support programs for students who are at risk of experiencing similar problems. Additional efforts are needed to reduce dropout rates and improve overall student retention.

3.2. Data analysis

The next stage is that the data will be sent to the Multinomial Naive Bayes classifier stage to analyze the data and predict which students are likely to drop out of school. The steps of the Multinomial Naive Bayes method analysis are explained as follows:

1. Gender probability
   Table 3 shows the distribution of gender against student status (Active, Potential Dropout, and Dropped Out). This table provides an overview of the number of students by gender in each category. This table also helps identify potential dropout issues based on gender, and shows that the number of male students who drop out is much higher than female students.

Table 3. Gender probability

Variabel		Active	Potential Dropout	Dropout
Gender	Man	320	53	128
	Woman	37	3	15

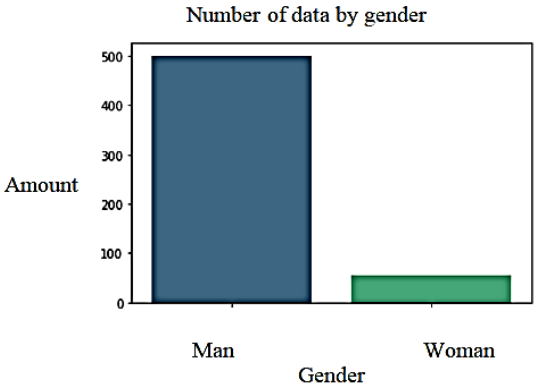

Figure 3. Gender graph

2. Probability of the average value
   The average value data is the result of calculations from all subjects for six semesters. This data representation shows that the average value between 81-95 is classified as "active", while the average value of 76-80 is in the "potential dropout" category, and the average value of 41-60 is included in the "dropout" category. In addition, there is a minimum value ranging from 20-40. Here is a graph of the average value, at figure 4.

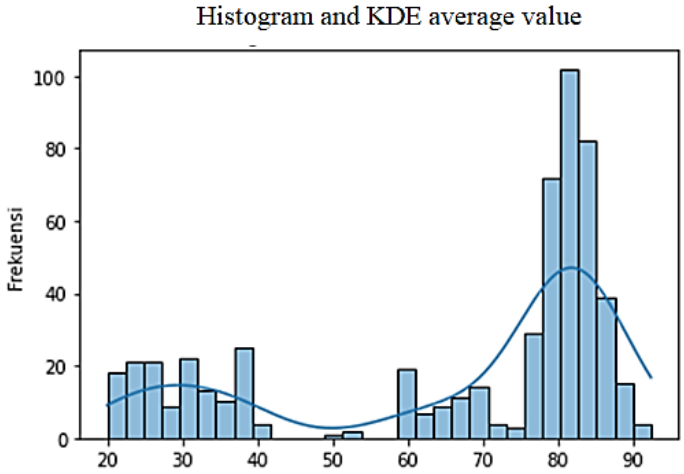

Figure 4. Average value

3. Probability of the area of origin of junior high schools based on sub-district
   Table 4 shows the probability distribution of junior high school origin areas against student status (Active, Potential to Drop Out, and Drop Out) based on sub-district. The sub-districts covered in this table include Sunggal, Medan, Pancur Batu, and outside Medan. This table serves to help understand the distribution of dropouts and potential dropouts based on junior high school origin areas.

Table 4. Probability of the area of origin of junior high schools

Variabel		Active	Potential Dropout	Dropout
Area of origin of junior high schools	Sunggal	144	32	52
	Medan	126	18	43
	Pancur Batu	62	0	52
	Luar medan	15	3	7

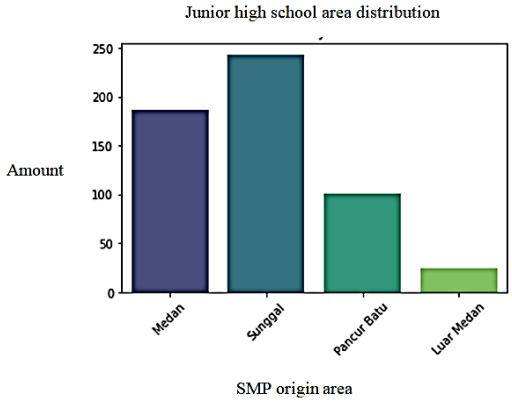

Figure 5. Graph of junior high school origin areas

4. Access distance
   Table 5 explains the possible distance of access related to student status, such as Active Potential, Potential to Drop Out, and Drop Out. This table illustrates how distance of access affects student status in terms of active potential, dropout risk, and dropout status.

Table 5. Access distance probability

Variabel		Active	Potential Dropout	Dropout
Access distance	Near	176	28	78
	Medium	165	19	59
	Far	17	8	6

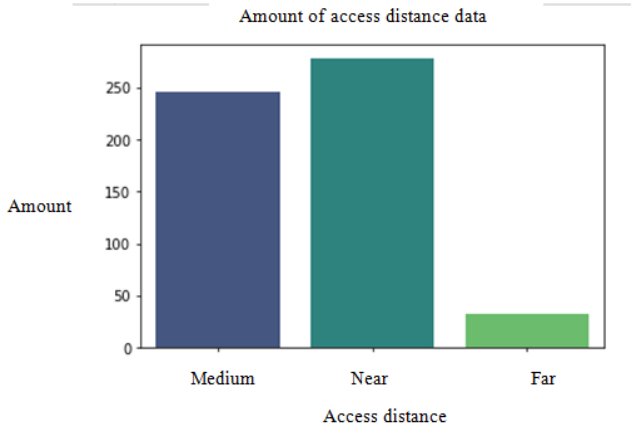

Figure 6. Graph of distance of access to school

In the study of student dropout potential analysis using the Multinomial Naive Bayes method in Jupyter Python, a systematic step will be described to identify factors that contribute to students' decisions to drop out of school before completing their degree. This model analyzes critical attributes such as academic and demographic data to understand and develop more effective intervention strategies. The accuracy of the model will be assessed using a confusion matrix. The following are the results of accuracy using a confusion matrix.

Tabel 6. Counfusion matrix

Actual / Predicted	Active	Potential Dropout Predicted	Dropout
Active	76	0	0
Potential Dropout	11	0	0
Dropout	8	0	25

Tabel 7. Evaluation

	Precision	Recall	F1-scroce	Support
Active	0.80	1.00	0.89	76
Potential dropout	0.00	0.00	0.00	11
Dropout	1.00	0.68	0.81	25
Accuracy			0.83	112
Macro avg	0.60	0.56	0.57	112
Weighted avg	0.77	0.83	0.78	112

The Multinomial Naive Bayes classification model applied to the dataset showed adequate performance in predicting student status with good accuracy. Evaluation through a confusion matrix, presented in the form of a heatmap, revealed the number of correct and incorrect predictions for the Active, Potential DropOut, and DropOut categories, and helped identify areas of misprediction. The classification report further provides metrics such as precision, recall, and F1-score, providing in-depth insights into the strengths and weaknesses of the model in classifying each student status. Overall, the model is effective in classifying categorical data, with results useful for understanding factors that influence predictions and for future model improvements.

4. CONCLUSION

Based on research using the Multinomial Naive Bayes algorithm, we successfully identified students who are potentially dropping out by taking into account four main factors, namely gender, average grades, distance between home and school, and junior high school of origin. This algorithm is able to analyze the relationship between these factors and student status, whether they are potentially dropping out or still active. The results of the study showed that the Multinomial Naive Bayes model has quite good performance in identifying active students, with an accuracy rate of 83.04% based on the confusion matrix in testing with 20% of the dataset. From this test, 76 students were identified as active, 11 students were potentially dropping out, and 25 students had dropped out. However, the precision, recall, and f1-score values for the potentially dropping out category could not be calculated due to the imbalance in the amount of data in the class. Therefore, the results of this study can provide insight to schools in identifying students who are at high risk of dropping out, as well as being the basis for designing appropriate interventions to prevent dropouts.

REFERENCES

[1] A. Hakim, "Faktor Penyebab Anak Putus Sekolah," *J. Pendidik.*, vol. 21, no. 2, pp. 122–132, 2020, doi: 10.33830/jp.v21i2.907.2020.

[2] B. Y. A. Lestari, F. Kurniawan, and B. R. Ardi, "Penyebab tingginya anak putus sekolah jenjang Sekolah Dasar (SD)," *J. Ilm. Sekol. Dasar*, vol. 4, no. 2, pp. 299–308, 2020.

[3] N. K. A. S. CAHYANI, N. L. P. SUCIPTAWATI, and K. G. SUKARSA, "Identifikasi Faktor Yang Memengaruhi Anak Putus Sekolah Di Kabupaten Badung," *E-Jurnal Mat.*, vol. 8, no. 4, p. 289, 2019, doi: 10.24843/mtk.2019.v08.i04.p267.

[4] S. Ujud, T. D. Nur, Y. Yusuf, N. Saibi, and M. R. Ramli, "Penerapan Model Pembelajaran Discovery Learning Untuk Meningkatkan Hasil Belajar Siswa Sma Negeri 10 Kota Ternate Kelas X Pada Materi Pencemaran Lingkungan," *J. Bioedukasi*, vol. 6, no. 2, pp. 337–347, 2023, doi: 10.33387/bioedu.v6i2.7305.

[5] J. J. Lanawaang and R. Mesra, "Faktor Penyebab Anak Putus Sekolah di Kelurahaan Tuutu Analisis Pasal 31 Ayat 1, 2, dan 3 UUD 1945," *J. Ilm. Mandala Educ.*, vol. 9, no. 2, pp. 1375–1381, 2023

[6] P. Astikaningtyas, "Peran Pendidikan Non Formal Untuk Membantu Siswa Drop Out Dalam Menyelesaikan Sekolahnya Berdasarkan Perspektif Islam (Studi Kasus Di Lembaga Ppap Seroja Jebres Surakarta)," *J. Pendidik. dan Keislam.*, vol. 157, no. 2, pp. 157–178, 2022, [Online]. Available: https://databoks.katadata.co.id/datapublish/2022/03/16/berapa-jumlah-anak-putus-sekolah-di-

[7] A. Yaneri, V. Suviani, and N. Vonika, "ANALISIS PENYEBAB ANAK PUTUS SEKOLAH BAGI KELUARGA MISKIN (Studi Kasus Anak Usia Sekolah Pada Keluarga Miskin di Kampung Lio Kota Depok)," *J. Ilm. Perlindungan dan Pemberdaya. Sos.*, vol. 4, no. 1, pp. 76–89, 2022, doi: 10.31595/lindayasos.v4i1.554.

[8] A. P. Tefa, "Analisis Faktor Penyebab Anak Putus Sekolah di Desa Oinlasi Kecamatan Mollo Selatan Kabupaten Timor Tengah Selatan," *PENSOS J. Penelit. dan Pengabdi. Pendidik. Sosiol.*, vol. 1, no. 1, pp. 47–56, 2023.

[9] Assa Riswan, "Jurnal Ilmiah Society," *Fakt. Anak Putus Sekol. Di Desa Sonuo Kec. Bolangitang Barat Kabupaten BolaangMongondow Utara*, vol. 2, no. 1, pp. 1–12, 2022.

[10] 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.

[11] H. F. Putro, R. T. Vulandari, and W. L. Y. Saptomo, "Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan," *J. Teknol. Inf. dan Komun.*, vol. 8, no. 2, 2020, doi: 10.30646/tikomsin.v8i2.500.

[12] Yuyun, Nurul Hidayah, and Supriadi Sahibu, "Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 4, pp. 820–826, 2021, doi: 10.29207/resti.v5i4.3146.

[13] M. A. Febriza, Q. J. Adrian, and A. Sucipto, "Penerapan Ar Dalam Media Pembelajaran Klasifikasi Bakteri," *J. BIOEDUIN Progr. Stud. Pendidik. Biol.*, vol. 11, no. 1, pp. 10–18, 2021, doi: 10.15575/bioeduin.v11i1.12076.

[14] A. Sabrani, I. G. W. Wedashwara W., and F. Bimantoro, "Multinomial Naïve Bayes untuk Klasifikasi Artikel Online tentang Gempa di Indonesia," *J. Teknol. Informasi, Komputer, dan Apl. (JTIKA)*, vol. 2, no. 1, pp. 89–100, 2020, doi: 10.29303/jtika.v2i1.87.

[15] M. Afriansyah, Joni Saputra, V. Y. P. Ardhana, and Yuan Sa'adati, "Algoritma Naive Bayes Yang Efisien Untuk Klasifikasi Buah Pisang Raja Berdasarkan Fitur Warna," *J. Inf. Syst. Manag. Digit. Bus.*, vol. 1, no. 2, pp. 236–248, 2024, doi: 10.59407/jismdb.v1i2.438.

[16] J. McCaffrey, "Multinomial Naive Bayes Classification Using the scikit Library." Accessed: Apr. 17, 2023. [Online]. Available: https://visualstudiomagazine.com/articles/2023/04/17/multinomial-naive-bayes.aspx

[17] A. Q. Ayuni *et al.*, "Klasifikasi Topik Berita Deutsche Welle Indonesia dengan Kata Kunci Indonesia Menggunakan Metode Multinomial Naive Bayes," *Jlk*, vol. 6, no. 1, pp. 11–16, 2023.

[18] E. Mulyani, F. P. B. Muhamad, and K. A. Cahyanto, "Pengaruh N-Gram terhadap Klasifikasi Buku menggunakan Ekstraksi dan Seleksi Fitur pada Multinomial Naïve Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 1, p. 264, 2021, doi: 10.30865/mib.v5i1.2672.

[19] T. Ige and S. Adewale, "AI Powered Anti-Cyber Bullying System using Machine Learning Algorithm of Multinomial Naïve Bayes and Optimized Linear Support Vector Machine Interception of Cyberbully Contents in a Messaging System by Machine Learning Algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 5, pp. 5–9, 2022, doi: 10.14569/IJACSA.2022.0130502.

[20] H. Herwanto, N. L. Chusna, and M. S. Arif, "Klasifikasi SMS Spam Berbahasa Indonesia Menggunakan Algoritma Multinomial Naïve Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 4, p. 1316, 2021,

BIBLIOGRAPHY OF AUTHORS

Dewi Afrianti is student of Computer Science Department, Bachelor of Science and Technology Program at the State Islamic University of North Sumatra. I have a special interest in Machine Learning, and Data Science.

Armansyah currently works as a teaching staff at State Islamic University North Sumatra, where he serves as a lecturer in Computer Science. especially in the field of Object-Oriented Programming. He completed his Bachelor's degree in Information Systems at STMIK Sisingamangaraja XII Medan in 2009 and earned his Master's degree from STMIK Eresha Jakarta in 2014.