Received November 11th, 2024; accepted Deember 4th, 2024. Date of publication December 30th, 2024
Digital Object Identifier: https://doi.org/10/25047/jtit.v11i2.434

Educational Data Mining for Student Academic Performance Analysis

KHOIRUNNISA' AFANDI[1], M. HABIBULLAH ARIEF[2], MARTIANA KHOLILA FADHIL[3]

[1,2,3]Information System, Universitas Jember, Indonesia

CORESPONDING AUTHOR: KHOIRUNNISA' AFANDI (email:oni.pssi@unej.ac.id)

ABSTRACT Good student academic performance is the key to success in the quality of education at university. One of the factors that influence academic success by utilising information technology and data analytics. This research incorporates GPA scores and other external factors that can affect students' academic performance such as parents' job and latest education, address, gender, extracurricular, etc. This research uses Machine Learning; Decision Tree, Random Forest, K-Nearest Neighbour, Support Vector Classifier, Naive Bayes, and Gaussian as methods to analyse and predict the academic performance of students of the Information Systems Study Program, Faculty of Computer Science at the University of Jember. The results showed that the Decision Tree algorithm has the highest accuracy value of 0.9264 followed by Random Forest and K-Nearest Neighbour. Meanwhile, the prediction results show that the Decision Tree, K-nearest neighbour, and Random Forest algorithms can predict the same results.

KEYWORDS: machine learning, educational data mining, student academic performance, prediction

I. INTRODUCTION

Student academic performance is a crucial indicator reflecting the quality of education at higher education institutions [1] and serves as a foundational basis for evaluation processes aimed at improving educational quality. Good academic performance is influenced by various factors, such as student engagement in extracurricular activities [2]. In this context, a deep understanding of the factors influencing academic performance becomes crucial, especially in today's digital era. Various previous studies have shown that good academic performance is not only influenced by students' intellectual abilities but also by external factors such as students' involvement in extracurricular activities [3] The author found that one-third of students perform well academically and actively participate in campus activities. Additionally, factors such as place of origin [4] also impact academic performance. These external factors are utilized in this study to examine whether there is an influence between external factors and students' academic performance.

In today's digital era, leveraging information technology and data analytics has become increasingly critical to understanding the factors that impact academic success. Machine learning, a branch of artificial intelligence[5], [6], offers various methods for analyzing [7], [8] and predicting student performance with high accuracy [9], [10].

Processing student academic performance data requires machine learning. Previous research has demonstrated the effectiveness of machine learning in predicting academic performance. For example, Wiyono et al. (2019) found that the Decision Tree and KNN algorithms could achieve up to 92% accuracy in predicting student performance, underscoring these methods' potential in educational contexts [11]. Jain et al. (2022) also reported that K-Nearest Neighbors (KNN) and Naive Bayes performed well, with KNN yielding the highest accuracy at 82% in academic performance analysis [12].

Research by Chen (2023) demonstrated that the Random Forest algorithm achieves the most accurate predictions, making it one of the most effective methods for analyzing academic performance [13]. Burman et al. (2019) and Wahed et al. (2020) also found that Support Vector Machine (SVM) effectively predicts student performance [14], [15]. In a study by Albreiki (2021), various machine learning (ML) techniques were shown to

help identify students at risk and predict dropout rates [16]. Therefore, using ML in this study is expected to accurately predict student academic performance based on external factors and GPA.

Based on the previously mentioned studies, it is evident that various algorithms exhibit distinct strengths depending on the parameters and datasets employed. These insights motivate the author to undertake a more in-depth analysis comparing machine learning models applicable in the educational sector. The objective of this research is to investigate the application of several machine learning models, including Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Decision Tree, Random Forest, and Gaussian, to analyze and predict the academic performance of students in the Information Systems program at Jember University.

The analysis of students' academic performance is anticipated to yield valuable insights that can enhance teaching and learning strategies within the Information Systems program at Jember University. By pinpointing the factors that lead to academic success, educational institutions can develop more targeted interventions to assist students who need extra support. This research aims not only to advance knowledge in the education field but also to offer practical recommendations for decision-making at the institutional level.

With a solid foundation in this research area, the author aspires to significantly contribute to the understanding and enhancement of students' academic performance through the application of information technology and data analytics. This study is expected to serve as a reference for future research and lay the groundwork for the development of more effective academic prediction systems going forward..

II. METHOD

This research method stage describes dataset collection, data pre-processing, and data analysis with machine learning algorithms.

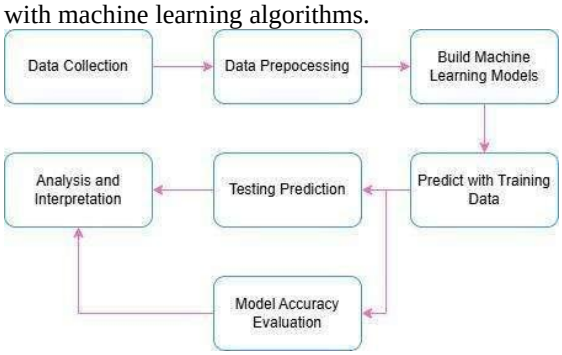

FIGURE 1. Research Method

Data Collection

In this study, data collection was conducted through the distribution of an online questionnaire using Google Forms. This questionnaire was designed to collect comprehensive information regarding the factors that may influence the academic performance of students in the Information Systems study program at the Faculty of Computer Science, Jember University. The data collection process took place during the 2022/2023 and 2023/2024 academic years, with a total of 203 respondents consisting of 98 male students and 104 female students.

The distributed questionnaire included various questions grouped into several categories, namely academic data and external factors. The academic data collected included the Cumulative Grade Point Average (GPA) and the Semester Credit System (SKS). Meanwhile, external factors are used in this research because several studies have proven that external factors influence the academic performance of students. The education level of parents, the occupations of parents, and the time spent studying can provide important insights into the social context and environment that affect academic performance. Research shows that the educational background of parents can significantly influence the motivation and academic support received by students [4]. In addition, previous studies have also indicated that students who are active in organizations or extracurricular activities tend to have better academic performance [3]. By combining this data, researchers can identify which variables are most influential in improving academic performance.

External factors can also help in understanding the variability among individual students. Data shows that there are significant differences in academic performance based on social and economic backgrounds, which can be measured through factors such as the number of family members and access to technology (e.g., Wi-Fi usage and smartphones). By considering these factors, the analysis can be more adaptive and responsive to the specific needs of students. Therefore, the external data used in this study includes the education level of parents, the occupations of parents, the amount of time spent studying, the place of origin, gender, the number of heads in a family, high school background, whether they use Wi-Fi, how often they use smartphones, whether they are active in organizations, how often they go out with friends, and whether they receive support from their parents..

Each question in the questionnaire was designed to provide deep insights into the conditions and habits of students that could affect their academic performance. For example, questions regarding study time were divided into several categories, such as less than 2 hours, 2-4 hours, 4-6 hours, and more than 6 hours. This aimed to understand the extent of time commitment that students devote to studying. Additionally, questions about involvement in organizations were measured with answer options ranging from inactive to active in various student organizations.

The use of Google Forms as a data collection tool provided several advantages, including ease of distribution and real-time data collection. Respondents could fill out the questionnaire at any time and from anywhere, thereby increasing the likelihood of participation. Once data collection was completed, all information obtained from the questionnaire was analyzed and categorized. Categorical data was converted into binary format to facilitate further analysis using machine learning algorithms. This process is a crucial step in ensuring that the analyzed data is accurate and reliable.

Through this systematic and structured data collection, this research aims to explore the relationship between external factors and students' academic performance, as well as to develop predictive models that can assist educational institutions in taking appropriate intervention steps for students who need additional support. Thus, the results of this study are expected to make a significant contribution to teaching and learning strategies in the academic environment.

Data Pre-processing

The data pre-processing phase is an essential part of evaluating students' academic performance, focusing on preparing the data for efficient processing with machine learning algorithms. In this research, data gathered from 203 students enrolled in the Information Systems program at the Faculty of Computer Science, University of Jember, encompasses various pertinent attributes, such as the Cumulative Grade Point Average (GPA), Semester Credit System (SKS), and other external factors that could impact academic success.

One of the initial steps in pre-processing is converting categorical data into numerical format. Categorical data, such as gender, parental education, and organizational activities, need to be expressed in numerical form to be processed by machine learning algorithms. For example, for the gender attribute, the value "female" can be coded as 1 and "male" as 0. Similarly, the education levels of the mother and father can be categorized using a numerical scale, where 1 represents elementary school, 2 represents junior high school, and so on up to higher education levels.

After converting categorical data, the next step is to standardize the data into binary format. This process involves changing all values that can be interpreted as "yes" to 1 and "no" to 0. For instance, an attribute indicating whether a student has access to Wi-Fi or not is changed to 1 for "yes" and 0 for "no." This step is crucial because many machine learning algorithms, such as K-Nearest Neighbors and Decision Trees, require data in numerical format for analysis.

Next, the standardized data is then organized into a table that includes 18 attributes for each student. This table provides a comprehensive description of the characteristics of students and the factors influencing their academic performance. For example, attributes such as study time, family size, and parental support are also converted into numerical format to facilitate further analysis.

The pre-processing process also includes data normalization, which aims to ensure that all attributes have the same scale. Normalization is important, especially when attributes have very different ranges of values, such as GPA ranging from 0.00 to 4.00 and study time measured in hours. By normalizing the data, we can prevent machine learning algorithms from giving disproportionate weight to certain attributes that have larger values.

After all data has been transformed into numerical format and normalized, the pre-processing phase concludes with the splitting of the data into training and testing sets. This division is crucial for assessing the performance of the machine learning model that will be developed. The training set is utilized to train the model, whereas the testing set is employed to evaluate the model's accuracy and effectiveness in predicting students' academic performance.

TABLE 1. Description of Student Dataset

No	Atributes	Indicators
1.	SKS	Semester credit system (numeric: from 20 to 144)
2	IPK	Grade point average numeric: (from 0,00 to 4,00)
3	Mom_edu	Mother's education (numeric : 1 – Elementary School, 2 – Junior High School, 3 – Senior High School, 4- Undergraduate Degree, 5- Master Degree, 6- Doctoral Degree)
4	Fa_edu	Father's education (numeric : 1 – Elementary School, 2 – Junior High School, 3 – Senior High School, 4- Undergraduate Degree, 5- Master Degree, 6- Doctoral Degree)
5	Study_time	Weekly study time (numeric : 1- <2 hours, 2- 2-4 hours, 3- 4-6 hours, 4- >6 hours)
6	Address	Student's home address province (numeric : 1-jawa timur, 2-jawa tengah, 3- jawa barat, 4-dki jakarta, 5-bali, 6- sumatra, 7- kalimantan, 8-ntb, 9- ntb, 10 - maluku. 11 - papua)
7	Gender	Student's Gender (biner: 1- female, 0- male)
8	fam_size	Family size (numeric : 1- 3 people, 2- 4 people, 3- 5 people, 4- 6 people, 5- >6 people)
9	high_scholl	High school origin (numeric : 1-SMA, 2- MA, 3- SMK)
10	wifi	Use wifi or not (biner: 1- yes, 0- no)
11	ormawa/ukm	Active in organisation (numeric : 0-no, 1-bpm, 2-bem, 3-etalase, 4-binary, 5-laos, 6-al azhar, 7-ukm o)
12	has phone	Student's has a phone or not (biner : 1- yes, 0- no)
13	hangout	Going out with friends (numeric : 1-never, 2- sometimes, 3-often, 4-very often)
14	usage smartphone	Usage time of smartphone in hours (numeric : 1- 1-3 hours, 2- 3-6 hours, 3- 6-9 hours, 4- >9 hours)

15	fam_support	Family educational support (numeric : 1- yes, 0- no)
16	class	Class (numeric : 1- Excellent > 3.5, 2- Good 3.0 - 3.5, 3- Poor < 3.0)

Analysis

The data collected has been processed using various machine learning algorithms. The bar graph in Figure 2 illustrates the comparison of model accuracy, highlighting the performance of several algorithms utilized in this study. The Decision Tree model achieved the highest accuracy among the others, with an accuracy value of 0.9264, indicating its potential effectiveness in classifying or predicting student academic performance.

In contrast, both the Random Forest and K-Nearest Neighbors (KNN) algorithms exhibited the same accuracy of 0.8037, followed by the Support Vector Classifier (SVC), which had an accuracy of 0.7791. The Naive Bayes and Gaussian models demonstrated lower accuracy, suggesting that these algorithms may not be the most suitable for predicting student academic performance within the context of this dataset.

From the results of this comparison, we can conclude that the Decision Tree is the most accurate model to use in predicting student academic performance in the Information Systems Study Program student data.

```
Perbandingan Akurasi Semua Model:
Naive Bayes: 0.2025
KNN: 0.8037
SVC: 0.7791
Decision Tree: 0.9264
Random Forest: 0.8037
Gaussian: 0.2025
```

FIGURE 2. Accuration of machine learning algorihtms

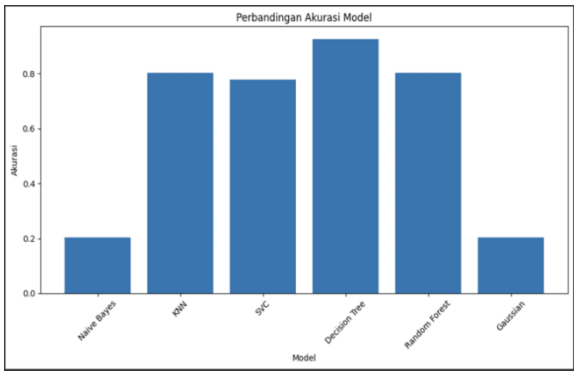

FIGURE 3. Graph of the accuration of machine learning algorihtms

Prediction is done by entering new data as follows, SKS: 0, GPA: 0, Mother's Education: High School, Father's Education: S1, Mother's occupation: housewife, father's job: lecturer, length of study: 4-6 hours, address: Central Java, gender: female, number of family members: 5 people, school origin: High school, have wifi: yes, join UKM/ormawa: no, have a smartphone: yes, often leave the house: sometimes, smartphone usage: >6 hours, get family support: yes. Based on the results in Figure 4 and 5, the algorithms that can predict new data accurately are the K-Nearest Neighbor, Decision Tree, and Random Forest algorithms. While other algorithms misclassify new data.

```
Prediksi untuk data baru:

Naive Bayes:
Prediksi: [2]
Probabilitas: [[8.92517382e-109 1.00000000e+000 0.00000000e+000]]

KNN:
Prediksi: [3]
Probabilitas: [[0.2 0.  0.8]]

SVC:
Prediksi: [1]
Probabilitas: [[0.08018408 0.14743384 0.77238209]]

Decision Tree:
Prediksi: [3]
Probabilitas: [[0. 0. 1.]]

Random Forest:
Prediksi: [3]
Probabilitas: [[0.28 0.21 0.51]]

Gaussian:
Prediksi: [2]
Probabilitas: [[8.92517382e-109 1.00000000e+000 0.00000000e+000]]
```

FIGURE 4. New Data Prediction Results of Student Academic Performance

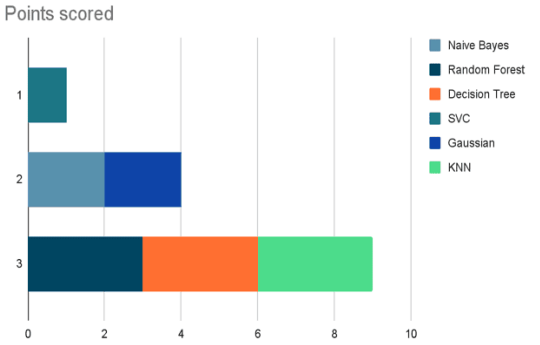

FIGURE 5. Graph of Data Prediction Results of Student Academic Performance

Decision Tree and Random Forest showed strong probabilities on specific classes, indicating a high level of confidence in their predictions. This indicates that both models have better generalisation capabilities on this data. The Random Forest model had probabilities that were slightly scattered among several classes, indicating that it considered several possibilities before making a final decision, while Decision Tree gave a more definitive decision on one class with a probability of 1. However, the predicted value for the new data across all models had the same value of 3.

III. RESULT AND DISCUSSION

The results of this study in Table 2 show that the Decision Tree and Random Forest algorithms have higher accuracy than other models in predicting student academic performance. This is also in line with the findings of several previous studies which also show that the Decision Tree algorithm [17]–[20] and Random Forest [21], [22] are generally superior in data analysis and prediction

[23], [24]. Research by Imran et al. (2019) found that the Decision Tree model has a high accuracy of 95.78% in academic performance classification, mainly due to its ability to handle variables with many levels and identify important patterns among them [17].

Random Forest which is a collection of many decision trees has the advantage of reducing the overfitting problem that is common in single Decision Tree models. Research by Chen & Ding (2023) supports this finding, where they mention that ensemble methods such as Random Forest often produce more stable and reliable performance in the context of academic prediction[23]. Random Forest is also more resilient to data variability [22], which is important when working with academic data that typically has large variations between individual students.

The K-Nearest Neighbors (KNN) algorithm is one of the most popular machine learning algorithms used for classification and regression. However, KNN has some limitations that need to be considered, especially related to computation, K parameter selection, and nearest neighbour search and selection. Choosing the right K value and the nearest neighbour search process can affect the performance of this algorithm, especially on large datasets that require higher computational resources [25]. In large datasets, the KNN algorithm may be less effective in reducing prediction uncertainty due to the high amount of data that needs to be considered [26]. Uncertainty is something that cannot be avoided, but it can and must be minimized to produce more precise decisions [27], [28].

The KNN algorithm remains a good choice for small datasets with less than 100 samples, as it has a shorter learning time than other more complex algorithms, such as Artificial Neural Networks[26], [29]. In situations where the interpretability of results is a priority, other algorithms such as Cubist, Multiple Linear Regression, or Random Forest may be more appropriate than KNN, as they offer greater transparency in explaining the factors that influence the prediction [26]. Likewise, a Support Vector Classifier (SVC) can perform better in small samples and improve generalisation [30] and it is necessary to combine it with other analysis techniques such as Principal Component Analysis (PCA) to increase the speed of the classifier generation [31]. SVC can provide good results on structured data with clear patterns but may be less efficient in capturing complex relationships among variables without deep preprocessing or feature engineering. Meanwhile, KNN is often sensitive to data scale and inter-class variability, which may affect its accuracy results in complex academic data.

The Naive Bayes classifier is widely used for its simplicity and robustness, but it has limitations, particularly when dealing with dependent features [32]. The assumption of feature independence often doesn't hold in real-world scenarios, leading to classification errors [32], [33]. To address this, researchers have proposed various improvements. Ou et al., (2022) introduced a mixed-attribute fusion-based NBC that uses an autoencoder to generate independent encoded attributes [33]. Yadav et al., (2019) suggested a perplexed Bayes classifier to handle feature dependencies [34]. Stephens et al., (2018) developed a framework to quantify error cancelation and predict NBC performance [32]. Despite its limitations, NBC can be effective for specific applications, such as rare disease detection in electronic medical records [35]. Understanding NBC's strengths and weaknesses allows for better implementation and adaptation to various problem domains.

Recent research has explored the limitations and applications of Gaussian algorithms in machine learning, particularly in dependent feature environments. Jollans et al. (2019) compared various machine learning regression methods, including Gaussian Process Regression, finding that their performance varied based on sample size, feature set size, and predictor effect size [36]. Naiem et al. (2023) proposed enhancements to the Gaussian Naïve Bayes classifier for DDOS detection in cloud computing, addressing issues related to feature independence and zero-frequency problems [37]. Caron et al. (2021) studied Gaussian linear model selection in dependent contexts, developing penalty functions for both short-range and long-range dependent error processes [38]. These studies collectively highlight the importance of considering feature dependencies and error process characteristics when applying Gaussian algorithms in machine learning.

In the context of academic performance prediction, this discovery has important practical implications. Decision Trees and Random Forest can be the first choice for building academic prediction systems that can help educational institutions provide early intervention for students who need additional academic support. The implementation of a precise and accurate model like this can support efforts to improve academic success in a more targeted way.

Overall, the results of this study show that the Decision Tree and Random Forest models are the most promising models for academic performance prediction analysis in the dataset used. Future research can optimize these two models by making adjustments to the hyperparameters and exploring feature engineering techniques to improve prediction accuracy. The addition of student data with a wider variety of variables can help generalise the findings and improve the practical applicability of these models in an educational context, potentially resulting in better strategies and interventions to improve student success. Spatial aspects of the analysis may need to be added to

increase the complexity of the data [39] so that the results of the analysis of student academic performance are more adaptive, flexible, and in-depth.

TABLE 2. Accuracies of Academic Performance Model

Classifier	Error Rate	Accuracy
Naive Bayes	0.7975	0.2025
Decision Tree	0.0736	0.8027
Random Forest	0.1963	0.7791
K-Nearest Neighbor	0.1963	0.9264
Support Vector Classifier	0.2209	0.8037
Gaussian	0.7975	0.2025

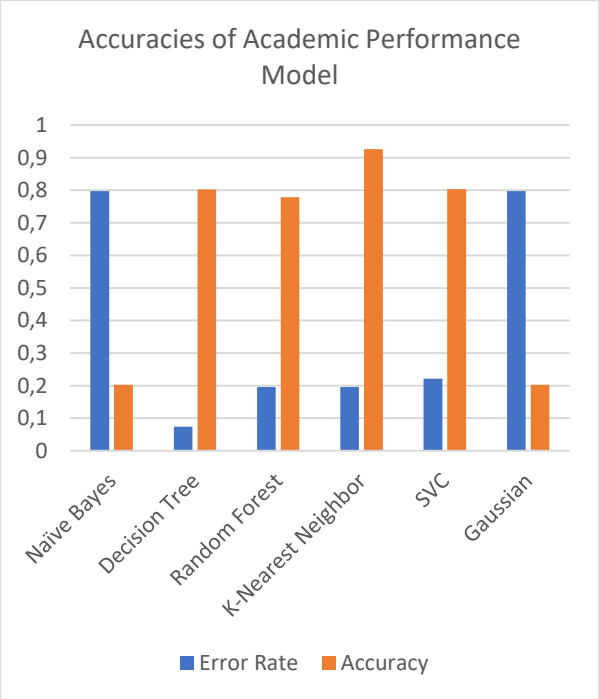

FIGURE 6. Comparison between Accuracies and Error Rate of Academic Performance Model

IV. CONCLUSION

This study concludes that external factors significantly influence the academic performance of students, informing the predictive capabilities of machine learning algorithms. External factors can affect students' academic performance, thereby impacting the predicted grades based on the combination of external factors and students' GPAs. The classification accuracy values of the algorithms found that Decision Tree, Random Forest, and K-Nearest Neighbor achieved the highest accuracy, allowing these algorithms to predict new data accurately. Decision Tree, Random Forest, and K-Nearest Neighbor are capable of building predictive systems for students who require additional academic support. This research underscores the potential benefits of utilizing machine learning to implement timely interventions aimed at enhancing student success through data-driven insights and tailored academic support. Recommendations for future research include adding external factors and analyzing which ones have the most significant impact on students' academic performance. The analysis of factors influencing students' academic performance should incorporate both quantitative and qualitative research methods and subjects.

ACKNOWLEDGMENT

The authors wish to convey their heartfelt thanks to the Faculty of Computer Science at the University of Jember for offering the resources and support essential for carrying out this research. We also appreciate the Information Systems Study Program for allowing access to the academic data used in this study, along with the committed faculty members who provided invaluable insights and guidance throughout the research process.

Special thanks go to the students who participated in the study and to the administrative staff who facilitated data collection and ensured the accuracy and confidentiality of student information. We would also like to acknowledge the contributions of our peer reviewers, whose feedback helped refine and improve this research.

REFERENCE

[1] C. S. Sarrico, "Quality management, performance measurement and indicators in higher education institutions: between burden, inspiration and innovation," *Qual. High. Educ.*, vol. 28, no. 1, pp. 11–28, 2022, doi: 10.1080/13538322.2021.1951445.

[2] A. E. King, F. A. E. McQuarrie, and S. M. Brigham, "Exploring the Relationship Between Student Success and Participation in Extracurricular Activities," *Sch. A J. Leis. Stud. Recreat. Educ.*, vol. 36, no. 1–2, pp. 42–58, 2021, doi: 10.1080/1937156X.2020.1760751.

[3] K. Afandi, M. H. Arief, N. Faizatul Laily, and D. Maulana Nugroho, "Analisis Performa Akademik Mahasiswa Menggunakan Social Network Analysis (Studi Kasus: Prodi Bisnis Digital Universitas dr. Soebandi)," *J. Technol. Informatics*, vol. 5, no. 2, pp. 64–69, 2024, doi: 10.37802/joti.v5i2.514.

[4] G. A. Gemechu, "Family Socio-economic Status Effect on Students' Academic Achievement at College of Education and," *J. Teach. Educ. Educ.*, vol. 7, no. 3, pp. 207–222, 2018.

[5] E. Barbierato and A. Gatti, "The Challenges of Machine Learning: A Critical Review," *Electron.*, vol. 13, no. 2, 2024, doi: 10.3390/electronics13020416.

[6] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, *Machine learning towards intelligent systems: applications, challenges, and opportunities*, vol. 54, no. 5. Springer Netherlands, 2021. doi: 10.1007/s10462-020-09948-w.

[7] P. Balaji, S. Alelyani, A. Qahmash, and M. Mohana, "Contributions of machine learning models towards student academic performance prediction: A systematic review," *Appl. Sci.*, vol. 11, no. 21, 2021, doi: 10.3390/app112110007.

[8] I. Issah, O. Appiah, P. Appiahene, and F. Inusah, "A systematic review of the literature on machine learning application of determining the attributes influencing academic performance," *Decis. Anal. J.*, vol. 7, no. March, p. 100204, 2023, doi:

10.1016/j.dajour.2023.100204.
[9] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and predicting students' performance by means of machine learning: A review," *Appl. Sci.*, vol. 10, no. 3, 2020, doi: 10.3390/app10031042.
[10] L. Zhao, J. Ren, L. Zhang, and H. Zhao, "Quantitative Analysis and Prediction of Academic Performance of Students Using Machine Learning," *Sustain.*, vol. 15, no. 16, 2023, doi: 10.3390/su151612531.
[11] S. Wiyono and T. Abidin, "Comparative Study of Machine Learning Knn, Svm, and Decision Tree Algorithm To Predict Student'S Performance," *Int. J. Res. -GRANTHAALAYAH*, vol. 7, no. 1, pp. 190–196, 2019, doi: 10.29121/granthaalayah.v7.i1.2019.1048.
[12] K. Jain and N. Choudhary, "Comparative analysis of machine learning techniques for predicting production capability of crop yield," *Int. J. Syst. Assur. Eng. Manag.*, vol. 13, pp. 583–593, 2022, doi: https://doi.org/10.1007/s13198-021-01543-8.
[13] Y. Chen and L. Zhai, "A comparative study on student performance prediction using machine learning," *Educ Inf Technol*, vol. 28, pp. 12039–12057, 2023, doi: https://doi.org/10.1007/s10639-023-11672-1.
[14] H. Waheed, S. U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Comput. Human Behav.*, vol. 104, 2020, doi: 10.1016/j.chb.2019.106189.
[15] I. Burman and S. Som, "Predicting Students Academic Performance Using Support Vector Machine," *Proc. - 2019 Amity Int. Conf. Artif. Intell. AICAI 2019*, pp. 756–759, 2019, doi: 10.1109/AICAI.2019.8701260.
[16] B. Albreiki, N. Zaki, and H. Alashwal, "Revisión bibliográfica sistemática de la predicción del rendimiento de los estudiantes mediante técnicas de aprendizaje automático," *Educ. Sci.*, vol. 11, no. 9, 2021.
[17] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, "Student academic performance prediction using supervised learning techniques," *Int. J. Emerg. Technol. Learn.*, vol. 14, no. 14, pp. 92–104, 2019, doi: 10.3991/ijet.v14i14.10310.
[18] I. O. Muraina, E. Aiyegbusi, and S. Abam, "Decision Tree Algorithm Use in Predicting Students' Academic Performance in Advanced Programming Course," *Int. J. High. Educ. Pedagog.*, vol. 3, no. 4, pp. 13–23, 2023, doi: 10.33422/ijhep.v3i4.274.
[19] A. D. Vergaray, C. Guerra, N. Cervera, and E. Burgos, "Predicting Academic Performance using a Multiclassification Model: Case Study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, pp. 881–889, 2022, doi: 10.14569/IJACSA.2022.01309102.
[20] N. Walia, M. Kumar, N. Nayar, and G. Mehta, "Student's Academic Performance Prediction in Academic using Data Mining Techniques," *SSRN Electron. J.*, pp. 1–5, 2020, doi: 10.2139/ssrn.3565874.
[21] D. Jacob and R. Henriques, "Educational Data Mining to Predict Bachelors Students' Success," *Emerg. Sci. J.*, vol. 7, no. Special Issue 2, pp. 159–171, 2023, doi: 10.28991/ESJ-2023-SIED2-013.
[22] J. Yu, "Academic Performance Prediction Method of Online Education using Random Forest Algorithm and Artificial Intelligence Methods," *Int. J. Emerg. Technol. Learn.*, vol. 16, no. 5, pp. 45–57, 2021, doi: 10.3991/ijet.v16i05.20297.
[23] S. Chen and Y. Ding, "A Machine Learning Approach to Predicting Academic Performance in Pennsylvania's Schools," *Soc. Sci.*, vol. 12, no. 3, 2023, doi: 10.3390/socsci12030118.
[24] M. H. Setiono, "Komparasi Algoritma Decision Tree, Random Forest, SVM dan K-NN dalam Klasifikasi Kepuasan Penumpang Maskapai Penerbangan," *Inti Nusa Mandiri*, vol. 17, no. 1, pp. 32–39, 2022, doi: 10.33480/inti.v17i1.3420 KOMPARASI.
[25] S. Zhang, "Challenges in KNN Classification," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 10, pp. 4663–4675, 2022, doi: 10.1109/TKDE.2021.3049250.
[26] Y. Khaledian and B. A. Miller, "Selecting appropriate machine learning methods for digital soil mapping," *Appl. Math. Model.*, vol. 81, pp. 401–418, 2020, doi: 10.1016/j.apm.2019.12.016.
[27] M. H. Arief and F. Ramdani, "Uncertainty analysis in spatial planning application based on geographical information system (GIS) in Malang City," in *IOP Conference Series: Earth and Environmental Science*, Kuala Lumpur, Malaysia: OP Publishing, 2022, pp. 1–11. doi: 10.1088/1755-1315/1064/1/012041.
[28] M. H. Arief, F. Ramdani, and F. A. Bachtiar, "A Conceptual Framework for Uncertainty Analysis in Map-Based Urban Spatial Planning," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Sep. 2021, pp. 197–202. doi: 10.1145/3479645.3479683.
[29] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," *Decis. Anal. J.*, vol. 3, no. May, p. 100071, 2022, doi: 10.1016/j.dajour.2022.100071.
[30] F. Gao, J.-G. Hsieh, and J.-H. Jeng, "Support Vector Classifier Trained by Gradient Descent," in *2021 International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD)*, Nanjing, China: IEEE, 2021, pp. 1–5. doi: 10.1109/ICSMD53520.2021.9670839.
[31] A. Yaicharoen, K. Hashikura, M. A. S. Kamal, and K. Yamada, "Improving Efficiency of Support Vector Mclassifier With Feature Selection," *ICIC Express Lett. Part B Appl.*, vol. 13, no. 5, pp. 479–486, 2022, doi: 10.24507/icicelb.13.05.479.
[32] C. R. Stephens, H. F. Huerta, and A. R. Linares, "When is the Naive Bayes approximation not so naive?," *Mach. Learn.*, vol. 107, no. 2, pp. 397–441, 2018, doi: 10.1007/s10994-017-5658-0.
[33] G. Ou, Y. He, P. Fournier-Viger, and J. Z. Huang, "A Novel Mixed-Attribute Fusion-Based Naive Bayesian Classifier," *Appl. Sci.*, vol. 12, no. 20, pp. 1–16, 2022, doi: 10.3390/app122010443.
[34] S. K. Yadav, D. K. Tayal, and S. N. Shivhare, "Perplexed Bayes classifier-based secure and intelligent approach for aspect level sentiment analysis," *Int. J. Adv. Intell. Paradig.*, vol. 13, no. 1–2, pp. 15–31, 2019, doi: 10.1504/IJAIP.2019.099941.
[35] B. Ehsani-Moghaddam, J. A. Queenan, J. MacKenzie, and R. V. Birtwhistle, "Mucopolysaccharidosis type II detection by Naïve Bayes Classifier: An example of patient classification for a rare disease using electronic medical records from the Canadian Primary Care Sentinel Surveillance Network," *PLoS One*, vol. 13, no. 12, pp. 1–17, 2018, doi: 10.1371/journal.pone.0209018.
[36] L. Jollans *et al.*, "Quantifying performance of machine learning methods for neuroimaging data," *Neuroimage*, vol. 199, no. December 2018, pp. 351–365, 2019, doi: 10.1016/j.neuroimage.2019.05.082.
[37] S. Naiem, A. E. Khedr, A. M. Idrees, and M. I. Marie, "Enhancing the Efficiency of Gaussian Naïve Bayes Machine Learning Classifier in the Detection of DDOS in Cloud Computing," *IEEE Access*, vol. 11, pp. 124597–124608, 2023, doi: 10.1109/ACCESS.2023.3328951.
[38] E. Caron, J. Dedecker, and B. Michel, "Gaussian linear model selection in a dependent context," *Electron. J. Stat.*, vol. 15, no. 2, pp. 4823–4867, 2021, doi: 10.1214/21-EJS1885.
[39] M. H. Arief, K. Afandi, Kustin, I. F. Arifin, and N. F. Laily, "Analisis Spasial Aksesibilitas Fasilitas Kesehatan di Kabupaten Jember," *J. Minfo Polgan*, vol. 12, no. September, pp. 1764–1771, 2023, doi:

https://doi.org/10.33395/jmp.v12i2.12984.

KHOIRUNNISA AFANDI Biography of Khoirunnisa Afandi

Khoirunnisa Afandi, born on December 28, 1994, is a dedicated lecturer in the Information Systems study program at the Faculty of Computer Science, Jember University. She completed her undergraduate studies (S1) at Jember University, where she developed a strong foundation in information technology and systems. Driven by her passion for knowledge and research, she pursued her master's degree (S2) at Institut Teknologi Sepuluh Nopember (ITS), specializing in data mining.

With a solid academic background, Khoirunnisa has also gained practical experience in the industry. She previously worked as a systems analyst at PT Sebangsa Bersama, where she honed her skills in analyzing and designing information systems to meet organizational needs. Her experience in the corporate world has enriched her teaching, allowing her to provide real-world insights to her students.

Khoirunnisa's research interests lie primarily in the field of data mining, where she explores innovative techniques and methodologies to extract valuable insights from large datasets. Her contributions to this field have been recognized through various publications and presentations at academic conferences. As a passionate educator, she is committed to fostering a learning environment that encourages critical thinking and innovation among her students.

M. HABIBULLAH ARIEF was born in Jember on February 11, 1992. He pursued his Associate Degree in Informatics Management at Telkom University, Bandung, followed by a Bachelor's Degree in Information Systems at Universitas Brawijaya, Malang, and a Master's Degree in Computer Science, also from Universitas Brawijaya, Malang. He has experience as a lecturer in the Information Systems Study Program at the Faculty of Computer Science, Universitas Jember, and as a Quality Assurance Staff at CV. Sinergi Spasial Indonesia. His research interests include Geographic Information Systems, focusing on applying spatial analysis in information systems to study phenomena.

MARTIANA KHOLILA FADHIL was born in Jember on July 19, 1999. She has experience as a lecturer in the Information Systems Study Program at the Faculty of Computer Science, Universitas Jember. She studied bachelor's degree in Information Systems at the University of Jember, and a master's degree in Information Systems at the Sepuluh November Institute of Technology. She has experience as an SPBE Surveyor at PT Tata Cipta Teknologi Indonesia. Her research interests include Information Systems, with a focus on IT adoption conducting the introduction and implementation of new IT applications.