Predicting student academic success using entry test, language, and spiritual formation data with ensemble learning

Evander Banjarnahor[1], Budi Wibawanta[2], Ronald Belferik[3], Rijanto Purbojo[4]
[1]Department of Mathematics, Faculty of Science and Technology, Universitas Pelita Harapan, Tangerang, Indonesia
[2]Professional Teacher Education Program, Faculty of Education, Universitas Pelita Harapan, Tangerang, Indonesia
[3]Department of Informatics, Faculty of Information Technology, Universitas Pelita Harapan, Tangerang, Indonesia
[4]Department of Psychology, Faculty of Psychology, Universitas Pelita Harapan, Tangerang, Indonesia

Article Info

Article history:

Received May 19, 2025
Revised Apr 30, 2026
Accepted May 17, 2026

Keywords:

Ensemble learning
Machine learning
Naïve Bayes
Student academic success
Support vector machine

ABSTRACT

Student academic success is influenced by various factors, both academic and non-academic. This study aims to examine the correlation between key student attributes and final grade point average (GPA), as well as to develop a machine learning model to predict academic success. The correlation analysis involved academic variables such as admission scores (Mathematics, English, Indonesian, and academic aptitude test/TPA), English ability test (EAT), spiritual formation (SF), and first-year GPA (GPA_1). The results indicate that GPA_1 has the highest correlation with final GPA (0.63), followed by SF (0.44), while other variables exhibit lower correlations. To enhance prediction accuracy, a machine learning approach using three primary models was employed: Naïve Bayes, support vector machine (SVM), and an ensemble learning method based on a stacking classifier that combines SVM and Naïve Bayes. The evaluation used five train-test split ratios and performance metrics, including accuracy, precision, recall, and F1-score. Experimental results reveal that the SVM model achieves the highest accuracy at 88.40%, followed by the ensemble model combining SVM and Naïve Bayes (88.00%) and the Naïve Bayes model (87.10%). These findings confirm that the machine learning approaches could effectively predict student academic success, providing a foundation for academic decision-making and educational intervention strategies.

Corresponding Author:

Budi Wibawanta
Professional Teacher Education Program, Faculty of Education, Universitas Pelita Harapan
Tangerang, Indonesaia
Email: budi.wibawanta@uph.edu

1. INTRODUCTION

Data mining is defined as discovering new and useful information from large amounts of data. This method aims to identify new patterns and trends using various classification algorithms [1]. Educational data mining (EDM) is a rapidly growing field of science that focuses on developing and applying methods to analyze data sets generated in educational environments [2], [3]. The analysis aims to gain insights into student preferences and behavior [4]. There are two main tasks in the education sector: analyzing student data and predicting their performance, which will contribute to determining their academic future [5]. Therefore, maintaining students' academic performance is very important [6], [7].

Based on this foundation, research in higher education is increasingly moving towards developing predictive models that integrate various aspects of assessment, such as entrance test data, language skills, and spiritual formation (SF). Recent studies have shown that incorporating these factors can improve prediction accuracy and provide a more comprehensive understanding of the factors influencing students' academic performance. Previous research indicates that language ability and other non-cognitive factors significantly influence academic success [8]. Zhou *et al*. [9], showed that emotional and spiritual intelligence can encourage student academic achievement. In addition to educational and technical factors, language proficiency also plays an important role in academic success, particularly among non-native speaker and international students. Previous studies have shown that language skills significantly influence academic performance, participation, engagement, and peer integration, highlighting the importance of language support in higher education environments [10].

Religion and education are closely related and can influence each other. Ahmed *et al*. [11] also found that good spirituality significantly and positively impacts student achievement motivation. Moreover, spiritual character building, including ethical values, discipline, and intrinsic motivation, influence student resiliency in facing academic challenges [12]. Our previous studies have also demonstrated that machine learning and learning analytics approaches can effectively predict student success using academic, behavioral, and sociocultural data. The first study employed random forest, support vector machine (SVM), and decision tree algorithms to analyze Moodle-based learning activity data and found that assignment attempt frequency had the strongest correlation with academic achievement, with random forest achieving prediction accuracy above 81% [13]. Subsequent studies developed predictive models for first-year GPA by integrating cognitive, socioeconomic, and sociocultural factors, including cultural connectedness, social acceptance, and cross-ethnic community support, which were found to significantly contribute to academic achievement and student resilience in multicultural higher education contexts [14], [15]. These findings suggest that student success prediction can be enhanced by integrating academic, behavioral, and contextual variables into machine learning model development.

This study highlights the importance of proper algorithm selection and the use of diverse data, including factors such as language ability and SF, to improve prediction accuracy. Sánchez-Sánchez compared four different models: i) linear regression, ii) support vector regression, iii) random forest, and iv) XGBoost in identifying first-year students' academic performance determinants had resulted similar predictive performance and successfully identified the main variables affecting educational performance [16]. A study done by Beseiso [17] showed that the XGB classifier, integrated with convolutional neural networks (CNN) from student data, consistently outperformed other models across two different data sets.

The main challenge in predicting student performance is identifying the model with the highest predictive accuracy. Ensemble learning shows a good level of predictive accuracy in providing strong evidence for the effectiveness of the model in scientific investigations [18]. Research by Wang *et al.* [19] showed that the graph-based machine learning ensemble method was able to improve and higher prediction accuracy up to 14.8% compared to other algotirma models. Ensemble learning effectively handles complex data and can predict student success with an accuracy of up to 97.88% [20]. The ensemble approach produces the best predictions by combining attributes such as attention in theory classes, Moodle quiz scores, and Moodle forums activity level [21]. In addition, ensemble learning is more resistant to variability and noise in the data, resulting in more reliable predictions [22]. This study uses an ensemble learning model to predict students' academic success based on data from entrance exams, language skills, and spiritual character formation. Studies integrating language-related variables and spiritual character formation with ensemble learning-based prediction models are limited.

2. RESEARCH METHOD

2.1. Research data

This study analyzed 452 academic data of the scholarship students admitted in year of 2020 and 2021. These data are classified as secondary data that extracted from three resources: i) the data from new student selection test for scholarship program during admission process that cosist of test scores of math, English, bahasa Indonesia and academic aptitute test ii) the data from the first year of study that consist of english abilty test (EAT), spritual formation score, and the first-year grade point average (GPA) and iii) the final GPA that reflects the student academic success. The first two data sources are considered as the predictor of the academic success of the scholarship program. We use the data of scholarship students because they are expected to graduate with higher academic standards compared to regular students. This will allow us to create a higher standard prediction model of academic success. Table 1 is the academic status criteria of the final year GPA for academic success.

Based on this table, a student with the final GPA below 2.75 are considered ineligible for graduation and are at risk of academic warning or remediation; the GPA of 2.75-2.99 is categorized as “Average,” indicating minimal academic achievement; the GPA of 3.00-3.29 is classified as “Good,” reflecting sufficient understanding of the course; and the GPA of 3.30-4.00 is categorized as “Excellent,” indicating excellent academic achievement. We used these criteria as the target class for the prediction model in our analysis.

Table 2. below illustrates how we apply the academic status criteria to the student final GPA. Later, we use this classification as the prediction model training process, where various machine learning methods will be applied to identify factors contributing to student academic success. The developed model is expected to provide better insights into identifying academically at-risk students early on and provide strategic recommendations to improve their academic achievements during their study.

Table 1. Final GPA academic status criteria

Interval GPA	Status
3,30 – 4,00	Excellent
3.00 – 3.29	Good
2.75 – 2.99	Average
0 – 2.74	Failed

Table 2. Illustration of student academic data based on graduation status

Cohort	Nama	Math	Bing	BIndo	TPA	EAT1	SF	GPA_1	GPA	Status
2020	Student_001	50,00	62,00	65,71	66,67	580,00	69,35	2,89	2,84	Failed
2020	Student_002	33,33	70,00	82,86	47,22	550,00	91,30	2,05	3,07	Average
2020	Student_003	53,33	76,00	65,71	58,33	490,00	80,32	3,44	3,30	Excellent
2020	Student_004	50,00	78,00	62,86	58,33	510,00	79,34	3,38	3,40	Excellent
2020	Student_005	56,67	84,00	71,43	72,22	523,00	83,45	3,39	3,48	Excellent
…	….	…	…	…	…	….	…	…	…	…
2021	Student_449	53,33	82,00	82,86	55,56	477,00	95,18	3,66	3,67	Excellent
2021	Student_450	36,67	66,00	62,86	66,67	480,00	88,71	3,43	3,26	Excellent
2021	Student_451	60,00	58,00	82,86	72,22	477,00	79,81	2,91	2,87	Average
2021	Student_452	53,33	86,00	57,14	63,89	586,67	95,95	3,82	3,88	Excellent

2.2. Method

The methodological steps of this research include collecting the dataset, training the model on the training dataset, categorizing the testing dataset, and evaluating the accuracy of the prediction results. The dataset used in this study is the student dataset, which is divided into training and testing data sets using the train-test split method with five different proportions: 10:90, 15:85, 20:80, 25:75, and 30:70. The dataset comprises both quantitative variables (Math, Bing, BIndo, TPA, EAT, SF, GPA_1, GPA) and qualitative variables from the academic status criteria based on the student final GPA (average, failed, excellent).

Several base classifiers are employed during the training phase for the ensemble stacking method. The clustering-based training process consists of two steps. In the first step, each base classifier is trained on the same dataset to generate the corresponding prediction results. In the second step, a meta-classifier utilizes these prediction results to determine the most probable class for the test data. The flowchart illustrating the ensemble stacking method is presented in Figure 1.

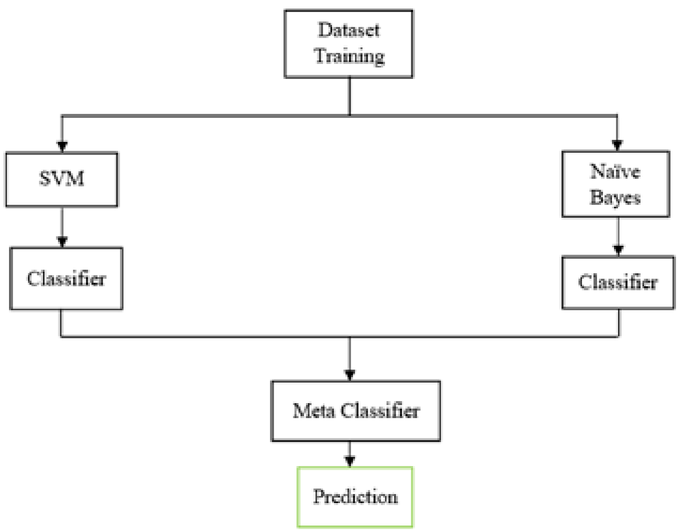

Figure 1. SVM and Naïve Bayes stacking method

2.2.1. Machine learning

Machine learning is a subfield of computer science that focuses on developing computer systems capable of learning and optimizing their performance based on experience [23]. The fundamental concept of machine learning is that computers should be able to learn from data without requiring explicitly detailed programming. During the training process of a machine learning model, the model is exposed to data and learns to perform specific tasks by analyzing patterns within the data [24]. Once trained, the machine learning model can make predictions or decisions based on new, previously unseen data. This ability to learn and adapt to new data has facilitated the widespread application of machine-learning methods in various studies, including those focused on classifying or predicting rainfall [25].

2.2.2. Support vector machine

SVMs are a strong group of supervised learning models that are commonly used for tasks like regression and classification [26]. SVM belongs to the category of supervised learning. SVM techniques are used in machine learning to classify linear and non-linear data. The basic idea of SVM is to maximize the distance of the hyperplane from the data it classifies. The training process of an SVM is formulated as a convex optimization problem, in which the goal is to identify the optimal hyperplane that minimizes classification errors while maximizing the margin. This is achieved by introducing lagrange multipliers to transform the constrained optimization problem into an unconstrained one, followed by solving its dual formulation to obtain the optimal solution. In a two-dimensional feature space, the hyperplane can be mathematically expressed as follows [27]:

$$f(x) = w.x + b \tag{1}$$

where x represents the input vector, w denotes the weight vector of the SVM, and b is the bias or offset term.

2.2.3. Naïve Bayes

Naïve Bayes is a widely used probabilistic algorithm that has proven to be particularly effective in text classification and sentiment analysis tasks due to its simplicity, computational efficiency, and scalability [28]. The Naïve Bayes classifier is a supervised learning algorithm designed for multi-class classification problems. Its construction is based on Bayes' theorem, which assumes conditional independence among all feature variables. Mathematically, the theorem can be expressed as follows [29].

$$Pr(C_i \mid X_j) = \frac{Pr(X_j|C_i)\, Pr(C_i)}{Pr(X_j)} \tag{2}$$

Where $i, j = 1,2, \dots, N$. Here, $Pr(C_i \mid X_j)$ represents the posterior probability, $Pr(X_j \mid C_i)$ denotes the likelihood, and both $Pr(C_i)$and $Pr(X_j)$correspond to the prior probabilities. By assuming that the features $X_{i1}, X_{i2}, \dots, X_{in}$are mutually independent, the Naïve Bayes model can be further simplified. The application of the fundamental Naïve Bayes theorem indicates that the probability of a data sample with specific attributes belonging to class C(Posterior) is determined by the prior probability of class C(Prior), multiplied by the likelihood of those attributes occurring within class C(Likelihood), and divided by the overall probability of the attributes across all possible classes (Evidence). The simplified form of the Naïve Bayes theorem is expressed as follows:

$$Posterior = \frac{Prior \times Likelihood}{Evidence} \tag{3}$$

The evidence value remains constant across all classes within the dataset. The posterior probabilities of each class are then compared to determine the most likely class label for a given data sample. This process computes posterior probabilities for classification, which can serve as informative features in sampling-based predictive modeling [29].

3. RESULTS AND DISCUSSION

In this study, we analyzed the performance of SVM, Naïve Bayes, and stacking-based ensemble models for data classification. The dataset was divided into training and test sets using five different ratios (10:90, 15:85, 20:80, 25:75, and 30:70) via the train-test split method. Subsequently, the data were normalized using the z-score normalization method to ensure uniform feature scaling. The SVM model was optimized with GridSearchCV to determine the optimal parameters, whereas the Naïve Bayes model was applied without parameter tuning. These two models were combined using a StackingClassifier ensemble, employing a linearly kernelized SVM as the final estimator.

Performance evaluation was conducted on the test data for each model and ratio using accuracy, precision, recall, and F1 score metrics. The testing process was iteratively executed for all combinations of models and data ratios. The evaluation results demonstrate the impact of training data ratios on model performance and underscore the effectiveness of ensemble models in leveraging the strengths of SVM and Naïve Bayes to enhance classification accuracy.

Figure 2 presents a correlation heatmap illustrating the linear relationships among several student academic variables, including the EAT score, Mathematics score, English (BIng) score, Indonesian (BIndo) score, Test of academic potential (TPA) score, SF score, first-year GPA (GPA_1), and final GPA (GPA). The correlation analysis indicates that GPA_1 exhibited the strongest positive correlation with the final GPA, with a correlation coefficient 0.63. This finding suggests that first-year academic performance serves quantitatively as the primary predictor of students' overall academic success.

In addition, the SF score demonstrated a moderate positive correlation (0.44) with the final GPA, indicating that the dimension of SF contributes to students' academic achievement. Lower correlations were observed for Math (0.13), English (0.11), and Indonesian (0.22) scores, suggesting that, although statistically significant, the contribution of these variables to the final GPA was more minor compared to GPA_1 and SF. The TPA and EAT scores exhibited very weak correlations with final GPA (0.03 and 0.07, respectively), implying that initial academic potential and analytical thinking test scores had minimal association with students' ultimate educational success. These findings underscore the importance of initial academic performance as a critical determinant in students' progression toward graduation.

This research employed an ensemble learning approach utilizing a stacking classifier, combining SVM and Naïve Bayes as base learners, with a linear-kernel SVM as the meta-learner. These models were selected based on their respective algorithmic characteristics: SVM is recognized for effectively managing high-dimensional and complex data, while Naïve Bayes offers advantages in probabilistic data processing and computational efficiency. By integrating these two models within a stacking framework, this study aims to enhance model accuracy and generalization capability beyond that achievable through single-model approaches.

The training process used various training-to-test data split ratios, specifically 10:90, 15:85, 20:80, 25:75, and 30:70, to evaluate model performance across different scenarios. The feature data were normalized using the StandardScaler method to achieve a more stable and uniform distribution. GridSearchCV was employed to identify optimal hyperparameters for the SVM model. After training and evaluating each model individually, an ensemble model was developed by combining predictions from the SVM and Naïve Bayes base learners, which then served as inputs for the meta-learner. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics to determine the effectiveness of the ensemble approach relative to individual models. Detailed evaluation results for each model are presented in Table 3, and Figure 3 shows a visualization of the accuracy of comparing the models.

Table 3. Performance of the model

Train/test ratio	Model	Accuracy	Precision	Recall/sensitivity	F1 score
10:90		81,80%	81,80%	81,80%	81,80%
15:85		84,40%	84,40%	84,40%	84,40%
20:80	Naïve Bayes	84,50%	84,50%	84,50%	84,50%
25:75		85,00%	85,00%	85,00%	85,00%
30:70		87,10%	87,10%	87,10%	87,10%
10:90		69,00%	69,00%	69,00%	69,00%
15:85		79,00%	79,00%	79,00%	79,00%
20:80	SVM	88,40%	88,40%	88,40%	88,40%
25:75		83,50%	83,50%	83,50%	83,50%
30:70		85,20%	85,20%	85,20%	85,20%
10:90		70,30%	70,30%	70,30%	70,30%
15:85		80,00%	80,00%	80,00%	80,00%
20:80	SVM + Naïve Bayes ensemble	85,10%	85,10%	85,10%	85,10%
25:75		85,30%	85,30%	85,30%	85,30%
30:70		88,00%	88,00%	88,00%	88,00%

In this study, we evaluated the performance of three machine learning models Naïve Bayes, SVM, and an ensemble approach combining SVM and Naïve Bayes in classification tasks. The model evaluation used four primary metrics: accuracy, precision, sensitivity (recall), and F1-score. Based on the simulation results, the Naïve Bayes model performed effectively, increasing accuracy from 81.80% to 87.10%. The SVM model demonstrated a broader performance range, achieving the lowest accuracy at 69.00% and peaking at 88.40%. Meanwhile, the ensemble approach (SVM + Naïve Bayes) produced competitive results, attaining a maximum accuracy of 88.00%. These findings indicate that combining the strengths of both models in an ensemble can effectively enhance classification performance.

Overall, the SVM model exhibited the best individual performance, reaching the highest accuracy of 88.40%, closely followed by the SVM + Naïve Bayes ensemble at 88.00%. The Naïve Bayes model also yielded commendable results, achieving a maximum accuracy of 87.10%. The implementation of an ensemble approach thus demonstrates considerable potential to further enhance classification accuracy by addressing the limitations inherent in individual models. Consequently, selecting an optimal model should align with the specific requirements and characteristics of the classification task.

While the predictive performance of the models demonstrates statistical robustness, it is equally essential to examine the interpretability and pedagogical relevance of the results within the broader scope of learning analytics. The strong association between first-year academic performance (GPA_1) and final academic outcomes, along with the meaningful contribution of non-academic indicators such as SF, provides a basis for developing data-informed strategies in educational settings. These insights can support early academic interventions, student support systems, and curriculum enhancements aimed at fostering holistic student development. Higher Education Institutions could impelement serveral actionable strategies to improve the equality, retention , and degree complation [30]. Furthermore, the integration of learning analytics with machine learning (i.e in an LMS) enables educators and administration to translate predictive outcomes into actionable pedagogical insights, allowing more personalized academic guidance and data-driven curriculum design [31]. This connection between prediction and interpretation enhances the practical value of the study and deepens the understanding of how machine learning can support evidence-based decision-making in education.

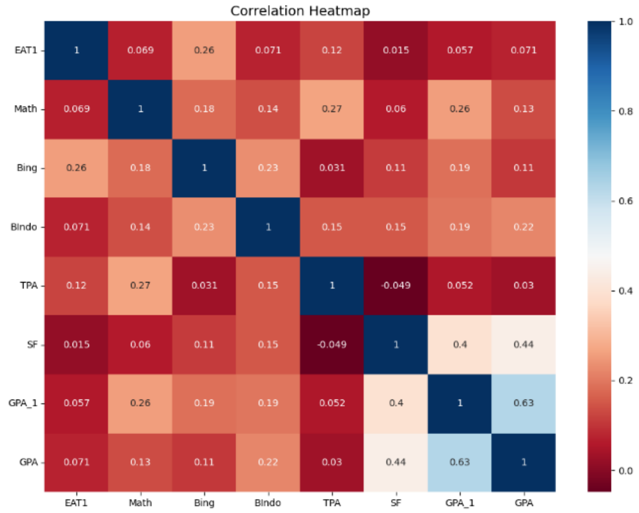

Figure 2. Correlation heatmap of the dataset

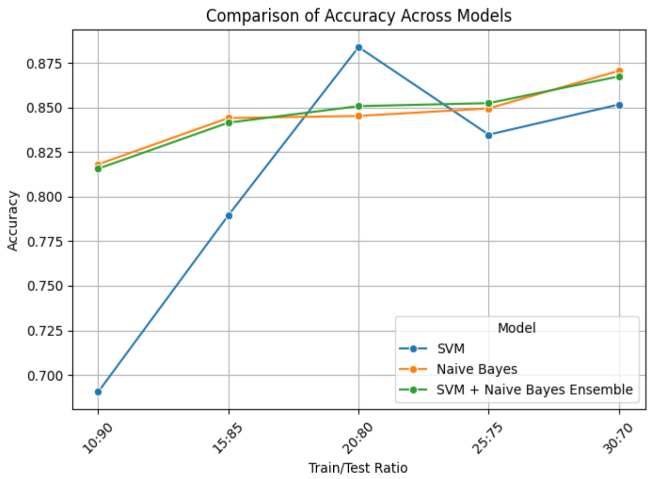

Figure 3. Comparison of accuracy across models

4. CONCLUSION

The results of this study indicate that first-year academic performance (GPA_1) exhibits the strongest positive correlation with the final GPA, having a correlation coefficient of 0.63, thus serving as the primary predictor of students' academic success. Additionally, non-academic factors, particularly SF scores, significantly contribute to academic achievement with a correlation of 0.44. In contrast, Mathematics, English, and bahasa Indonesia scores demonstrated weaker correlations with the final GPA. Furthermore, the TPA and EAT scores showed weak correlations with students' academic success. These findings suggest that initial academic performance, alongside specific non-academic factors, significantly influences academic achievement more than purely cognitive variables, such as essential academic ability and analytical potential.

Regarding the classification models, this research compared three approaches: Naïve Bayes, SVM, and a stacking classifier-based ensemble method combining both models. Evaluation results revealed that the SVM model achieved the highest accuracy of 88.40%, closely followed by the ensemble model (SVM + Naïve Bayes) with an accuracy of 88.00%. In comparison, the Naïve Bayes model obtained a maximum accuracy of 87.10%. The ensemble approach demonstrates the potential for enhancing classification performance by integrating the strengths of each model. Therefore, machine learning-based methods, particularly the SVM and ensemble approaches, can effectively predict student academic success based on the various educational and non-academic factors examined in this study. Future work integrating explainable AI techniques such as SHapley Additive exPlanations (SHAP) or Local interpretable model-agnostic explanations (LIME) is recommended to support transparent and actionable learning analytics, enabling educational stakeholders to better interpret and apply model insights in targeted academic interventions.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to the Faculty of Education, Universitas Pelita Harapan, for the academic support, facilities, and contributions provided throughout this research. The authors also appreciate all individuals who contributed through discussions and technical assistance during the completion of this study.

FUNDING INFORMATION

The publication of this article was supported by the Institute for Research and Community Service (LPPM), Universitas Pelita Harapan.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Evander Banjarnahor	✓	✓		✓	✓	✓		✓	✓	✓			✓	
Budi Wibawanta		✓			✓	✓		✓	✓	✓	✓			
Ronald Belferik	✓	✓	✓	✓			✓			✓			✓	
Rijanto Purbojo				✓	✓					✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

The raw data used in this research are confidential academic records and are not available in a public repository. The processed data that support the conclusions of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, Dec. 2022, doi: 10.1186/s40561-022-00192-z.

[2] I. Papadogiannis, M. Wallace, and G. Karountzou, "Educational data mining: a foundational overview," *Encyclopedia*, vol. 4, no. 4, pp. 1644–1664, Oct. 2024, doi: 10.3390/encyclopedia4040108.

[3] S. S. Bhaskaran, "Mathematical model design to integrate EDM with KDDM framework," *Heliyon*, vol. 11, no. 10, p. e42076, May 2025, doi: 10.1016/j.heliyon.2025.e42076.

[4] X. Sun, "Educational technology based on data mining: mining student behavior patterns to optimize teaching strategies," *International Journal of High Speed Electronics and Systems*, Dec. 2024, doi: 10.1142/s0129156425401019.

[5] S. M. Dol and P. M. Jawandhiya, "Systematic review and analysis of EDM for predicting the academic performance of students," *Journal of The Institution of Engineers (India): Series B*, vol. 105, no. 4, pp. 1021–1071, Aug. 2024, doi: 10.1007/s40031-024-00998-0.

[6] P. Porouhan and W. Premchaiswadi, "Process mining and learners⇔ behavior analytics in a collaborative and web-based multi-tabletop environment," *International Journal of Online Pedagogy and Course Design*, vol. 7, no. 3, pp. 29–53, Jul. 2017, doi: 10.4018/IJOPCD.2017070103.

[7] R. Z. Tan, P. C. Wang, W. H. Lim, S. H. C. Ong, and K. Avnit, "Early prediction of students' Mathematics performance," in *Proceedings of 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering, TALE 2018*, IEEE, Dec. 2018, pp. 651–656. doi: 10.1109/TALE.2018.8615289.

[8] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, p. 3, Dec. 2020, doi: 10.1186/s41239-020-0177-7.

[9] Z. Zhou, H. Tavan, F. Kavarizadeh, M. Sarokhani, and K. Sayehmiri, "The relationship between emotional intelligence, spiritual intelligence, and student achievement: a systematic review and meta-analysis," *BMC Medical Education*, vol. 24, no. 1, p. 217, Mar. 2024, doi: 10.1186/s12909-024-05208-5.

[10] M. A. Hamade, N. A. Ismail, and S. N. A. Mohamed, "The role of language proficiency in academic success for international students: A case study approach," *International Journal of Studies in International Education*, vol. 1, no. 1, pp. 14–19, Feb. 2024, doi: 10.62951/ijsie.v1i1.107.

[11] M. A. Ahmed, S. Hashim, and N. R. N. Yaacob, "Islamic spirituality, resilience and achievement motivation of yemeni refugee students: A proposed conceptual framework," *International Journal of Learning, Teaching and Educational Research*, vol. 19, no. 4, pp. 322–342, Apr. 2020, doi: 10.26803/ijlter.19.4.19.

[12] S. Rahayu, "The impact of artificial intelligence on education: Opportunities and challenges," *Jurnal Educatio FKIP UNMA*, vol. 9, no. 4, pp. 2132–2140, 2023, doi: doi.org/10.31949/educatio.v9i4.6110.

[13] E. Banjarnahor, D. P. Sibarani, B. Wibawanta, D. A. G. Sihotang, and Y. A. S. Abraham, "A machine learning approach to predicting student success through data mining of LMS moodle activity data," in *Proceedings - 2025 4th International Conference on Electronics Representation and Algorithm: Artificial Intelligence: Creating Tomorrow's World Today, ICERA 2025*, IEEE, Jun. 2025, pp. 233–238. doi: 10.1109/ICERA66156.2025.11086633.

[14] B. Wibawanta and E. Banjarnahor, "Predicting first-year academic success: The influence of socioeconomic and educational backgrounds among first-generation scholarship students," *IJORER : International Journal of Recent Educational Research*, vol. 6, no. 5, pp. 1703–1717, Sep. 2025, doi: 10.46245/ijorer.v6i5.1015.

[15] E. Banjarnahor, B. Wibawanta, R. Belferik, T. A. P. Sadeva, and N. Andriani, "Implementation of machine learning models to predict first-year student achievement based on ethnicity and entrance test scores," in *2025 1st International Conference on Data Science and Geoinformatics, ICDSG 2025*, IEEE, Nov. 2025, pp. 370–375. doi: 10.1109/ICDSG67714.2025.11381374.

[16] A. M. Sánchez-Sánchez, J. D. Mello-Román, M. Segura, and A. Hernández, "Identifying the determinants of academic success: a machine learning approach in Spanish higher education," *Systems*, vol. 12, no. 10, p. 425, Oct. 2024, doi: 10.3390/systems12100425.

[17] M. Beseiso, "Enhancing student success prediction: A comparative analysis of machine learning technique," *TechTrends*, vol. 69, no. 2, pp. 372–384, Mar. 2025, doi: 10.1007/s11528-025-01044-6.

[18] M. Kumar, V. Bhardwaj, D. Thakral, A. Rashid, and M. T. Ben Othman, "Ensemble learning based model for student's academic performance prediction using algorithms," *Ingenierie des Systemes d'Information*, vol. 29, no. 5, pp. 1925–1935, Oct. 2024, doi: 10.18280/isi.290524.

[19] Y. Wang, A. Ding, K. Guan, S. Wu, and Y. Du, "Graph-based ensemble machine learning for student performance prediction," Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.07893

[20] A. Al-Ameri, W. Al-Shammari, A. Castiglione, M. Nappi, C. Pero, and M. Umer, "Student academic success prediction using learning management multimedia data with convoluted features and ensemble model," *Journal of Data and Information Quality*, vol. 17, no. 3, pp. 1–16, Sep. 2025, doi: 10.1145/3687268.

[21] W. Chango, R. Cerezo, and C. Romero, "Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses," *Computers and Electrical Engineering*, vol. 89, p. 106908, Jan. 2021, doi: 10.1016/j.compeleceng.2020.106908.

[22] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Applied Sciences (Switzerland)*, vol. 11, no. 1, pp. 1–28, Dec. 2021, doi: 10.3390/app11010237.

[23] A. Kumar and S. Singh, "A review on Indian summer monsoon rainfall prediction using machine learning techniques," in *ICSCCC 2021 - International Conference on Secure Cyber Computing and Communications*, IEEE, May 2021, pp. 524–528. doi: 10.1109/ICSCCC51823.2021.9478104.

[24] O. S. Riza and A. Nuryadi, "Bibliometric study : Rainfall classification - prediction using machine learning methods," *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi*, vol. 14, no. 2, pp. 206–218, Nov. 2023, doi: 10.31849/digitalzone.v14i2.16618.

[25] A. Gupta, H. K. Mall, and S. Janarthanan, "Rainfall prediction using machine learning," in *2022 1st International Conference on Artificial Intelligence Trends and Pattern Recognition, ICAITPR 2022*, IEEE, Mar. 2022, pp. 1–5. doi: 10.1109/ICAITPR51569.2022.9844203.

[26] M. Jabardi, "Support vector machines: Theory, algorithms, and applications," *Infocommunications Journal*, vol. 17, no. 1, pp. 66–75, 2025, doi: 10.36244/ICJ.2025.1.8.

[27] Z. Luo *et al.*, "A method for prediction and analysis of student performance that combines multi-dimensional features of time and space," *Mathematics*, vol. 12, no. 22, p. 3597, Nov. 2024, doi: 10.3390/math12223597.

[28] E. Febriyanto and W. Wasilah, "Optimization of naive bayes and decision tree algorithms through the application of bagging and adaboost techniques for predicting student study success," *Jurnal Infotel*, vol. 17, no. 1, pp. 136–149, May 2025, doi: 10.20895/infotel.v17i1.1258.

[29] G. Y. Sheu and N. R. Liu, "Symmetrical and asymmetrical sampling audit evidence using a naive bayes classifier," *Symmetry*, vol. 16, no. 4, p. 500, Apr. 2024, doi: 10.3390/sym16040500.

[30] S. Sodagari, "Predictive modeling and cohort data analytics for student success and retention," *Evaluation and Program Planning*, vol. 113, p. 102689, Dec. 2025, doi: 10.1016/j.evalprogplan.2025.102689.
[31] M. Shoaib, N. Sayed, J. Singh, J. Shafi, S. Khan, and F. Ali, "AI student success predictor: Enhancing personalized learning in campus management systems," *Computers in Human Behavior*, vol. 158, p. 108301, Sep. 2024, doi: 10.1016/j.chb.2024.108301.

BIOGRAPHIES OF AUTHORS

Evander Banjarnahor, S.Pd., M.Si. Received his bachelor's degree in mathematics education from Universitas Pelita Harapan, Indonesia in 2016, master's degree in mathematics from Universitas Indonesia in 2021. His research focus is Data Science and Bioinformatics. Currently, he is working as a Lecturer at the Department of Mathematics, Faculty of Science and Technology, Universitas Pelita Harapan. He can be contacted at email: evander.banjarnahor@uph.edu.

Dr. Budi Wibawanta, S.Sos., M.Si. Received his bachelor's degree in social science from Universitas Gadjah Mada, Indonesia, in 1998, his master's degree in science from Universitas Gadjah Mada in 2002, and his doctoral degree in Education from Universitas Negeri Jakarta in 2018. His research focuses on education and evaluation. He is currently a lecturer at the Professional Teacher Education Program, Faculty of Education, Universitas Pelita Harapan, Tangerang, Indonesia. He can be contacted at email: budi.wibawanta@uph.edu.

Ronald Belferik, S.Kom., M.Kom. Received his bachelor's degree in computer science from Universitas Mikroskill, Indonesia, in 2013 and his master's degree in computer science from Universitas Sumatera Utara in 2020. His research focuses on programming and algorithms. He is currently a lecturer at the Department of Informatics, Faculty of Information Technology, Universitas Pelita Harapan, Tangerang, Indonesia. He can be contacted at email: ronald.belferik@uph.edu.

Dr. Rijanto Purbojo, S.Psi., M.Sc. Received his bachelor's degree in Clinical Psychology from Universitas Padjadjaran, Bandung, Indonesia in 1996, master's and doctoral degree in Educational Technology from Universitas of Twente in 2005. His research focus is Educational Technology. Currently, he is working as a Lecturer at the Department of Psychology, Faculty of Psychology, Universitas Pelita Harapan. He can be contacted at email: rijanto.purbojo@uph.edu.