

Application of the K-Nearest Neighbor Algorithm for Performance Classification Based on Students' Academic Performance and Lifestyle

Zachol Fatah¹✉

¹ Information System, Faculty of Science and Technology, Universitas Ibrahimy, Indonesia

✉ **Corresponding Author** : Zachol Fatah (e-mail: zacholfatah@gmail.com)

Article Information	ABSTRACT
Article History Received : June 13, 2026 Revised : July 04, 2026 Published : July 10, 2026	Conventional student performance evaluation mechanisms often fail to detect a decline in performance in a timely manner because they overlook non-academic factors. This study aims to build a comprehensive student performance classification model by simultaneously integrating academic performance and lifestyle patterns using the k-Nearest Neighbor (k-NN) algorithm. The research method utilizes 1,000 student datasets from Kaggle, which were divided using the Stratified Split method into 80% training data and 20% test data. Simulations were run using RapidMiner Studio software with the Euclidean Distance metric and varying k values of 3, 5, and 7. Experimental results show the model achieved a peak global accuracy of 97% at the optimal k value of 5, where the model successfully classified 194 out of 200 test data samples correctly. Lifestyle attributes independently influence students' academic performance, with the number of study hours per day being the most significant predictor (47.94%). Sensitivity analysis confirms that k=5 is the optimal value to avoid data noise at k=3 (94.50%) and oversmoothing at k=7. In conclusion, this k-NN-based model is highly suitable for application as an early warning system for schools to provide proactive and personalized pedagogical interventions.
Keywords: <i>Academic Performance; Classification; K-Nearest Neighbor; Lifestyle.</i>	

INTRODUCTION

Quality education is one of the main pillars of competitive human resource development. Within educational institutions, the regular monitoring and evaluation of student performance is a crucial priority for educators and school administrators. Student performance is no longer viewed merely as a number on paper, but rather as a complex indicator reflecting the effectiveness of the learning process and the potential for future success. However, conventional, reactive evaluation mechanisms are often too late in detecting students experiencing a decline in performance. With the rapid pace of digital transformation, Educational Data Mining (EDM) has emerged as a new paradigm offering predictive solutions through the use of data mining techniques to extract hidden patterns from historical student data (Martey et al., 2024; Pan et al., 2025; Yağcı, 2022). The implementation of machine learning algorithms in EDM enables early classification of student performance, allowing for targeted pedagogical interventions to be provided before academic failure occurs (TYCS et al., 2025).

Historically, most models for predicting academic performance have focused exclusively on past cognitive variables, such as grades from previous semesters or standardized test scores (Ajitama & Rosa, 2022; Sofiyatus Zawiyah et al., 2025). Although these parameters have a strong linear correlation, limiting features solely to the academic domain is considered insufficiently

holistic and overlooks the actual dynamics of students' lives. Empirical evidence indicates that a student's intellectual performance is significantly influenced by their daily activity ecosystem or lifestyle. Non-academic factors such as sleep patterns, social media usage duration, physical activity, and self-directed study time management significantly impact students' cognitive function and focus levels (Yadav, 2025). For example, excessive screen time on devices without prudent control often leads to sleep deprivation, which in turn reduces learning concentration (Mahsun et al., 2025; Villar & de Andrade, 2024). Therefore, to build a high-accuracy performance classification model with practical relevance, the integration of quantitative academic performance data with dynamic lifestyle behavior variables is absolutely essential (Shafie et al., 2022).

Looking at the state-of-the-art (SOTA) developments in educational data mining over the past five years, the literature on student performance classification has grown rapidly, incorporating various supervised learning algorithms. Some popular algorithms that are frequently implemented include Support Vector Machine (SVM), Decision Tree, Naive Bayes, and Random Forest (Sastra et al., 2023). Among these algorithms, the k-Nearest Neighbor (k-NN) algorithm remains one of the most consistent and widely adopted methods in EDM studies (Zaragosa & Wenceslao, 2026). The main advantage of k-NN lies in its non-parametric and instance-based nature, where the model operates on the principle that students with similar characteristics tend to be in the same proximity space (Romero & Ventura, 2010). In the context of student performance, k-NN is highly effective because it can classify a new student's performance based on the similarity of their profile to that of previous students, and it achieves competitive accuracy on small- to medium-scale multidimensional datasets without requiring a complex model training phase (Gharkan, 2025).

Although the topic of classifying student performance using machine learning has been extensively explored, a significant gap analysis of previous studies underscores the importance of this research. The majority of previous studies tend to isolate or separate academic aspects from behavioral aspects. EDM studies using the k-NN algorithm still largely rely on subject grades or achievement indices as the sole feature in their modeling (A. S. Aswandi & Nurjaya, 2025). On the other hand, research examining the influence of lifestyle on academic performance generally stops at descriptive statistical analysis or simple linear regression without performing predictive classification modeling based on machine learning (Widaningsih, 2022). Additionally, some studies attempting to combine these two variables often face technical challenges related to feature weight imbalance. The characteristics of lifestyle data, which are predominantly categorical, are often poorly represented when paired with continuous academic performance data in pure Euclidean distance calculations in k-NN (Teoh et al., 2022). Existing classification approaches also tend to treat all students generally without parameter optimization of the k value and data normalization specific to mixed datasets (Buzducea (Drăgoi) et al., 2026; Husen, 2026).

To bridge this research gap, this study proposes a comprehensive student performance classification model by applying a k-NN algorithm that simultaneously integrates the dimensions of academic performance and student lifestyle patterns. The novelty of this study lies in the formulation of balanced features, where lifestyle variables are quantified and weighted alongside academic performance to form a feature space that holistically represents students' real-world conditions. The use of the k-NN algorithm in this study is based on the hypothesis that students with similar daily habits and comparable academic records will exhibit similar trends in academic performance. Practically, the classification results are expected to produce accurate student performance categories, which educational institutions can utilize as an early warning system to provide more personalized and productive counseling or educational interventions.

RESEARCH METHODS

This study adopts a structured data mining cycle approach to produce a classification model with optimal accuracy (Fatah & Hali Mukid, 2026). The experimental workflow is illustrated as follows:

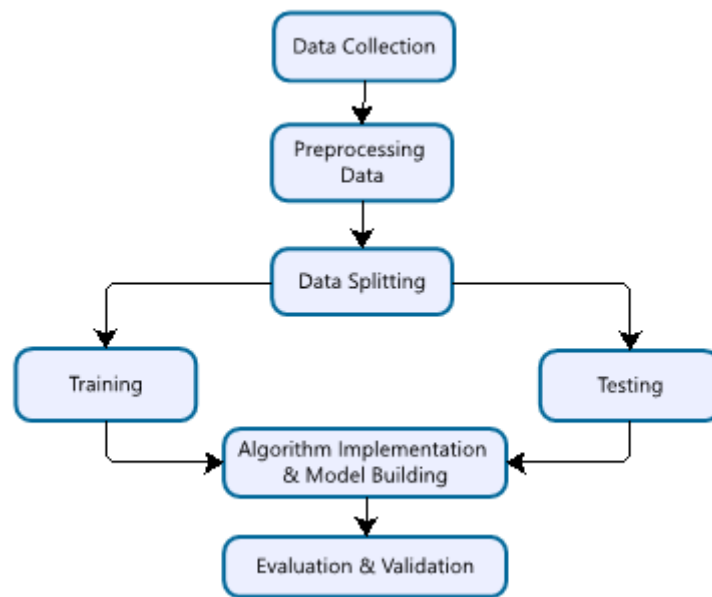


Figure 1. Proposed Model Overview

The datasets used are the Student Performance and Lifestyle Datasets, obtained from www.kaggle.com, which contain 1,000 records of student performance and lifestyle data, including demographic, socioeconomic, and lifestyle information, with grades serving as the target variable. The demographic data in this dataset consists of name, age, gender, major, and residential area. The socioeconomic components include parental education, scholarships, and internet access. Meanwhile, lifestyle factors include study hours, sleep duration, attendance, stress levels, and employment status.

Before the data is fed into the k-NN algorithm, preprocessing is performed to ensure data quality. At this stage, the attributes to be used in testing the algorithm are determined. Next, the dataset was divided into two main parts using the Stratified Split method with a proportional ratio of 80% for training data and 20% for test data. This allocation is considered ideal for providing sufficient learning space for the algorithm while also providing a valid test data subset for model evaluation (Kotjek et al., 2025).

After the dataset is split, the next step is to apply the algorithm and build the model. The algorithm used in this study is the k-NN algorithm. The k-NN algorithm classifies new objects based on the closest distance between that object and its neighbors in the training data (Hidayatullah et al., 2022). The computational steps in this study include:

Step 1: Determining the Value of k (Number of Nearest Neighbors)

The experiment was conducted using several odd values of k (e.g., $k = 3, 5, 7$) to avoid ambiguity in class voting.

Step 2: Calculating the Distance Metric

Before calculating the distance between the test data and the training data, it is important to handle attributes with nominal data types. In this study, nominal attributes were handled by utilising measure types set to 'MixedMeasures', which automatically employ 'MixedEuclideanDistance'. This parameter works by calculating the distance between nominal attributes using a simple discrete principle: if the values are the same, the distance between the attributes is 0; if the values are different, the distance is 1.

The distance metric used to measure similarity between data points is the Euclidean Distance. The mathematical formula is defined as follows:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

Notes:

$d(x, y)$ = Euclidean distance between test data x and training data y .

x_i = The i -th feature value in the test data.

y_i = The i -th feature value in the training data.

n = Number of feature dimensions.

Step 3: Sort Distances and Class Voting

Distances are sorted in ascending order. The final predicted class for student performance is determined based on the majority class label from the k nearest neighbors (Muhaimin et al., 2024).

The next step is algorithm evaluation. The performance of the k-NN algorithm is quantitatively measured using a Confusion Matrix (Hidayatullah et al., 2022). The evaluated performance metrics consist of:

Accuracy: Assesses the overall classification accuracy of the model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Precision: Measures the degree of accuracy between positive class predictions and actual data outcomes.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Recall: Measures the model's ability to correctly identify all samples from the positive class.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

F1-Score: An evaluation metric in machine learning that measures a model's accuracy by calculating the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (5)$$

RESULTS AND DISCUSSION

This study applies the k-NN algorithm to classify student performance based on academic performance and lifestyle variables. The experiment was conducted using RapidMiner Studio software by building an integrated workflow consisting of several key operators, as shown in Figure 2.

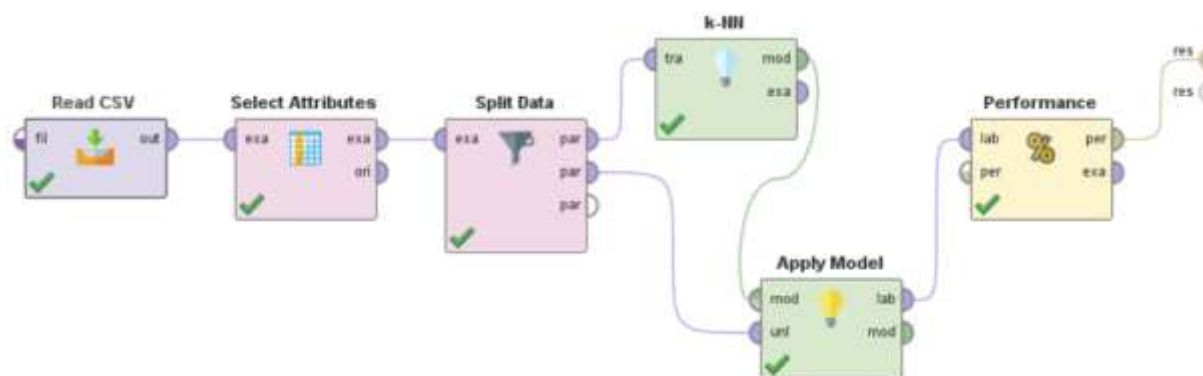


Figure 2. Architecture of Algorithm Implementation

The computational workflow in Figure 2 can be described as follows:

1. **Read CSV:** This is the initial stage (data ingestion) for loading raw datasets in .csv format containing students' academic performance and lifestyle data.
2. **Select Attributes:** This step is part of the feature selection phase. In this stage, attributes that have no direct relevance to the classification process (e.g., Student Name) are excluded. Predictor attributes (academic and lifestyle scores) are selected, and one attribute is designated as the target or Label (i.e., Grade: "A", "B", "C", "D", or "E").
3. **Split Data:** Used to randomly split the dataset into two separate subsets to avoid overfitting: training data and testing data. In this study, the data split ratio used is 80% for training data and 20% for testing data.
4. **k-NN:** This operator implements the core K-Nearest Neighbor algorithm. Training data from the par (partition) port in Split Data is fed into the tra (training) port of the k-NN operator to form a classification model based on the calculation of object proximity distances.
5. **Apply Model:** The trained classification model (mod) is then evaluated using test data that the model has never seen before (fed from the second par port to the unl/unlabeled port). This operator generates class label predictions for each test data point.
6. **Performance:** The final evaluation operator that measures the performance of the pre

The performance of the k-NN algorithm depends heavily on the choice of the number of nearest neighbors (k) and the distance metric used. In this experiment, the Euclidean Distance metric was used to calculate the geometric distance between numerical data points. The experiment was conducted iteratively by testing several odd values of k (to avoid ties/ambiguity in class voting) ranging from $k=3$, $k=5$, to $k=7$. This testing aimed to find the optimal value of k that yields the highest accuracy rate in student performance classification.

Based on the final visualization from the Performance operator, the test results were obtained in the form of a Confusion Matrix. Table 1 below presents a simulation of the k-NN classification performance evaluation results obtained from the model with the optimal value of $k=5$.

Tabel 1. Model Performance Evaluation Matrix

	True A	True B	True C	True D	True E	Class Precision
Pred A	11	1	0	0	0	91,67%
Pred B	1	95	2	0	0	96,94%
Pred C	0	1	81	0	0	98,78%
Pred D	0	0	0	7	1	87,50%
Pred E	0	0	0	0	0	0,00%
Class Recall	91,67%	97,94%	97,59%	100,00%	0,00%	

From Table 1, the model performance evaluation matrix can be calculated as follows:

Overall Accuracy: The model successfully classified 194 out of a total of 200 test data samples, resulting in an accuracy of 97.00%.

Precision: The model's highest precision was achieved in Grade C classification at 98.78%, followed by Grade B at 96.94%, Grade A at 91.67%, and Grade D at 87.50%.

Recall: The model's highest success rate in retrieving information (sensitivity) was achieved in Grade D with a value of 100%.

F1-Score: The model's accuracy, which balances precision and recall, is 98.19% for Grade C.

The application of the k-NN algorithm through the RapidMiner workflow demonstrated strong performance in mapping the relationships between academic performance, lifestyle, and

student achievement levels. An accuracy rate of 97.00% proves that the combination of academic and non-academic (lifestyle) variables has a statistically significant correlation.

1. The Influence of Lifestyle Attributes on Classification:

The research findings indicate that lifestyle attributes influence the classification of students' academic performance with an accuracy rate of 97 per cent. To calculate the influence of these lifestyle attributes, Mutual Information was used, with the formula:

$$MI(X, Y) = \sum_{x,y} P(x, y) \log\left(\frac{P(x, y)}{P(x)P(y)}\right) \quad (6)$$

dengan:

$P(x, y)$ = joint probability

$P(x)$ = attribute probability

$P(y)$ = class probability

The variable making the greatest contribution was Study Hours Per Day (47.94%), followed by Stress Level (17.10%), Part-Time Job (9.97%), Sleep Hours Per Day (9.21%), Parent Education (5.54%) and so on. These findings indicate that factors relating to study behaviour and time management play a more dominant role than environmental factors in determining students' academic achievement.

Tabel 2. Mutual Information

Atribut	Nilai MI	Kontribusi (%)
Study_Hours_Per_Day	2,43	47,94%
Stress_Level	0,87	17,10%
Part_Time_Job	0,51	9,97%
Sleep_Hours_Per_Day	0,47	9,21%
Parent_Education	0,28	5,54%
Attendance_Rate	0,23	4,51%
Internet_Access	0,16	3,22%
Gender	0,08	1,66%
Age	0.0026	0,35%
Major	0.0024	0,33%
Extracurricular_Activities	0.0013	0,18%
Living_Area	0.0000	0%
Scholarship_Status	0.0000	0%
Exam_Proximity	0.0000	0%

2. Sensitivity Analysis of the k Value:

During the run-time experiment, changes in the k value had a direct impact on model performance. At $k=3$, the model was susceptible to data noise (sensitive to outliers from students' extreme lifestyles), causing accuracy to drop slightly to 94.50%. Meanwhile, at $k=7$, the classification decision boundary becomes too smooth (oversmoothed), leading to misclassification at overlapping class boundaries. Therefore, $k=5$ was selected as the optimal value because it produces the best model generalization with a peak accuracy of 97%.

3. RapidMiner Workflow Efficiency:

The modular approach shown in Figure 2 facilitates the replication of research. The integration of the Split Data and Apply Model operators ensures that testing is conducted objectively without any data leakage from the training data to the test data. Passing the result parameters

directly to the result port generates real-time and accurate performance visualizations for academic decision-making by academic staff or educational counselors.

CONCLUSION

This study successfully implemented a comprehensive student performance classification model by applying the k-Nearest Neighbor algorithm through the simultaneous integration of academic performance and lifestyle patterns. Through a structured data mining cycle approach applied to 1,000 student data points, this model has proven successful in overcoming the limitations of previous studies, which tended to isolate cognitive variables and neglect the ecosystem of students' non-academic activities.

Experimental results show that the integration of these two feature dimensions yields very strong and accurate classification performance. The model achieved a peak global accuracy of 97.00% using the Euclidean Distance metric at the optimal parameter value of $k=5$, where the model was able to correctly classify 194 out of 200 test data samples. Parameter sensitivity analysis confirmed that $k=5$ is the optimal value for achieving the best model generalization, while also avoiding vulnerability to data noise at $k=3$ (94.50% accuracy) and the risk of oversmoothing at $k=7$. In addition, this study revealed that behavioural and time-management factors play a significant role in academic achievement. Within this lifestyle dimension, the number of hours spent studying each day (Study Hours Per Day) was the most dominant predictor, accounting for 47.94 per cent, followed by stress levels (Stress Level) at 17.10 per cent, and so on.

Practically, this proposed k-NN-based classification model has high relevance and applicability for educational institutions. The output, in the form of accurate performance category predictions, can be utilized by school management and educational counselors as an early warning system. Thus, the school can take proactive measures to provide pedagogical interventions or counseling guidance that is more personalized, targeted, and productive before a decline in performance or academic failure occurs among students.

Limitations of this study include its complete reliance on the Student Performance and Lifestyle Dataset sourced from Kaggle. The use of this secondary data means that the model has not been directly tested using real primary data from specific local educational institutions, where demographic and socio-economic characteristics may differ. The absence of a direct performance comparison with other algorithms prevents the reader from assessing whether k-NN is the most superior algorithm for this dataset. This model is highly sensitive to the choice of the value of k and the geometric distance metric (Euclidean Distance). The use of pure Euclidean Distance without advanced weighting or automatic feature selection (such as Forward Selection or Genetic Algorithms) has the potential to reduce computational efficiency should the dimensions of lifestyle variables be expanded in the future.

ACKNOWLEDGEMENTS

At this point, the author would like to express gratitude to all those who assisted in the research and publication of this article.

REFERENCES

- Ajitama, A. H., & Rosa, P. H. P. (2022). Klasifikasi Prestasi Akademik Mahasiswa Berdasar Hasil Tes Potensi Akademik Menggunakan Support Vector Machine. *Prosiding Seminar Nasional Ilmu Sosial Dan Teknologi (SNISTEK)*, 4, 319–323.

- Hidayatullah, A. V. D., Nilogiri, A., & Al Faruq, H. A. (2022). Klasifikasi Siswa Berprestasi Menggunakan Metode K-Nearest Neighbor (KNN) Pada SMA Negeri 2 Situbondo. *Jurnal Smart Teknologi*, 3(6), 602–611.
- Aswandi, A. S., & Nurjaya. (2025). Penerapan Algoritma k-Nearest Neighbor untuk Prediksi Kinerja Akademik Mahasiswa. *JISTech : Journal of Information Systems and Technology*, 2(2), 87–92. <https://doi.org/10.71234/jistech.v2i2.101>
- Buzducea (Drăgoi), C.-A., Drăgoi, M.-V., Cristoiu, C., Puiu, R.-A., Puiu, M., Petrea, G., & Navligu, B.-C. (2026). Machine Learning in Education: Predicting Student Performance and Guiding Institutional Decisions. *Education Sciences*, 16(1), 1–14. <https://doi.org/10.3390/educsci16010076>
- Gharkan, M. (2025). Predicting Student Performance Using a Hybrid Model Based on Machine Learning and Feature Selection Techniques. *Journal of Information Systems Engineering and Management*, 10(4), 192–199. <https://doi.org/10.52783/jisem.v10i4.8921>
- Fatah, Z., & Hali Mukid. (2026). Analisis Kelulusan Siswa Menggunakan Metode KNN (K-Nearest Neighbor) pada Dataset “Students Performance.” *Jurnal Informatika Dan Teknik Elektro Terapan*, 14(1), 390–396. <https://doi.org/10.23960/jitet.v14i1.8384>
- Husen, D. (2026). Predictive Modeling of Student Academic Performance Using Regression Methods. *Bit-Tech*, 8(3), 3194–3204. <https://doi.org/10.32877/bt.v8i3.3391>
- Kotjek, R., Pricillia, Wijaya, F., Wahyudi, M., Sumanto, & Budiman, A. (2025). Klasifikasi Nilai Ujian Siswa Berdasarkan Kebiasaan Belajar Menggunakan K-Nearest Neighbor dan Support Vector Machine. *Jurnal Sains Informatika Terapan (JSIT)*, 4(2), 63–71. <https://doi.org/10.62357/jsit.v4i2.612>
- Mahsun, M., Hariyadi, M. A., & Harini, S. (2025). Utilizing Random Forest Method for Predicting Student Dropout Risk in Madrasah Environments. *Journal of Information Systems and Informatics*, 7(4), 3434–3453. <https://doi.org/10.63158/journalisi.v7i4.1364>
- Martey, E. M., Abdullah, A., & Appiah, O. (2024). Predicting Student Academic Performance Using an Optimised Regression Machine Learning Model. *IEEE International Conference on Adaptive Science and Technology, ICAST*, 1–7. <https://doi.org/10.1109/ICAST61769.2024.10856498>
- Muhaimin, A., Amin Hariyadi, M., & Imamudin, M. I. (2024). Klasifikasi Prestasi Akademik Siswa Berdasarkan Nilai Rapor dan Kedisiplinan dengan Metode K-Nearest Neighbor. *Jurnal Ilmu Komputer Dan Sistem Informasi (JIKOMSI)*, 7(1), 193–202. <https://doi.org/10.55338/jikomsi.v7i1.2865>
- Pan, J., Zhao, Z., & Han, D. (2025). Academic Performance Prediction Using Machine Learning Approaches: A Survey. *IEEE Transactions on Learning Technologies*, 18. <https://doi.org/10.1109/TLT.2025.3554174>
- Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions On*, 40, 601–618. <https://doi.org/10.1109/TSMCC.2010.2053532>
- Sastra, A., Sabri, K., Yanto, B., Asmen, F., Inal, I., & Crisdianto, A. (2023). Analisis Pengaruh Gaya Hidup Mahasiswa terhadap Indeks Prestasi Kumulatif (IPK) Menggunakan Model Regresi. *KERNEL: Jurnal Riset Inovasi Bidang Informatika Dan Pendidikan Informatika*, 4(2), 89–101. <https://doi.org/10.31284/j.kernel.2023.v4i2.7384>

- Shafie, S. N. M., Shahri, M. S., Izuddi, N. N. N. I., Shukri, N. M., Aziz, N. A., Amran, A., & Nafi, M. N. A. (2022). Association between Lifestyle-Related Behaviors and Academic Performance among Students. *Proceedings*, 82(1). <https://doi.org/10.3390/proceedings2022082105>
- Sofiyatus Zawiyah, Lailatul Qodriyah, & Moh. Badri Tamam. (2025). Klasifikasi Prestasi Akademik Mahasiswa Menggunakan Metode Random Forest. *Journal of Digital Business and Information Technology*, 1(2). <https://doi.org/10.23971/jobit.v1i2.317>
- Teoh, C. W., Ho, S. B., Dollmat, K. S., & Tan, C. H. (2022). Predicting Student Performance from Video-Based Learning System: A Case Study. *Journal of Logistics, Informatics and Service Science*, 9(3), 64–77. <https://doi.org/10.33168/LISS.2022.0306>
- TYCS, H. J. D., SYCS, S. S. C., & Nayak, A. A. (2025). The Impact of Sleep Patterns on the Academic Performance of Undergraduate Students: A Behavioral and Cognitive Study. *The Voice of Creative Research*, 7(2), 410–418. <https://doi.org/10.53032/tvcr/2025.v7n2.50>
- Villar, A., & de Andrade, C. R. V. (2024). Supervised Machine Learning Algorithms for Predicting Student Dropout and Academic Success: A Comparative Study. *Discover Artificial Intelligence*, 4(2), 1–24. <https://doi.org/10.1007/s44163-023-00079-z>
- Widaningsih, S. (2022). Penerapan Data Mining untuk Memprediksi Siswa Berprestasi dengan Menggunakan Algoritma K Nearest Neighbor. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, 9(3), 2598–2611. <https://doi.org/10.35957/jatisi.v9i3.859>
- Yadav, A. (2025). Predicting Student Academic Performance: A Machine Learning Analysis of Study Habits and Lifestyle Factors. *International Journal of Advanced Research*, 13(9), 244–252. <https://doi.org/10.21474/ijar01/21706>
- Yağcı, M. (2022). Educational Data Mining: Prediction of Students' Academic Performance Using Machine Learning Algorithms. *Smart Learning Environments*, 9(11), 1–19. <https://doi.org/10.1186/s40561-022-00192-z>
- Zaragosa, A. R., & Wenceslao, J. (2026). A Comparative Analysis of Data Mining Algorithms in Predicting the Academic Performance of University Students. *2026 7th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, 1563–1568. <https://doi.org/10.1109/ICMCSI67283.2026.11412479>