Original Article

Check for updates

Available online at
www.heca-analitika.com/jeml

Journal of Educational Management and Learning

Vol. 1, No. 2, 2023

Leveraging Artificial Intelligence to Predict Student Performance: A Comparative Machine Learning Approach

Aga Maulana [1], Ghazi Mauer Idroes [2,3,*], Pati Kemala [2], Nur Balqis Maulydia [2], Novi Reandy Sasmita [4], Trina Ekawati Tallei [5], Hizir Sofyan [4] and Asep Rusyana [4]

[1] Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; agamaulana@usk.ac.id (A.M.)
[2] Graduate School of Mathematics and Applied Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; idroesghazi_k3@abulyatama.ac.id (G.M.I.); patikemala@mhs.usk.ac.id (P.K.); maulydiabalqis@gmail.com (N.B.M.)
[3] Department of Occupational Health and Safety, Faculty of Health Sciences, Universitas Abulyatama, Aceh Besar 23372, Indonesia;
[4] Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; novireandys@usk.ac.id (N.R.S.); hizir@usk.ac.id (H.S.); asep.rusyana@usk.ac.id (A.R.)
[5] Department of Biology, Faculty of Mathematics and Natural Sciences, Sam Ratulangi University, Manado 95115, Indonesia; trina_tallei@unsrat.ac.id (T.E.T.)

* Correspondence: idroesghazi_k3@abulyatama.ac.id

Article History

Received 3 November 2023
Revised 9 December 2023
Accepted 16 December 2023
Available Online 20 December 2023

Keywords:
Data mining
Random forest
Tabular data
Supervised learning
Education

Abstract

This study explores the application of artificial intelligence (AI) and machine learning (ML) in predicting high school student performance during the transition to university. Recognizing the pivotal role of academic readiness, the study emphasizes the need for tailored interventions to enhance student success. Leveraging a dataset from Portuguese high schools, the research employs a comparative analysis of six ML algorithms—linear regression, decision tree, support vector regression, k-nearest neighbors, random forest, and XGBoost—to identify the most effective predictors. The dataset encompasses diverse attributes, including demographic details, social factors, and school-related features, providing a comprehensive view of student profiles. The predictive models are evaluated using R-squared, Root Mean Square Error, and Mean Absolute Error metrics. Results indicate that the Random Forest algorithm outperforms others, displaying high accuracy in predicting student performance. Visualization and residual analysis further reveal the model's strengths and potential areas for improvement, particularly for students with lower grades. The implications of this research extend to educational management systems, where the integration of ML models could enable real-time monitoring and proactive interventions. Despite promising outcomes, the study acknowledges limitations, suggesting the need for more diverse datasets and advanced ML techniques in future research. Ultimately, this work contributes to the evolving field of educational AI, offering practical insights for educators and institutions seeking to enhance student success through predictive analytics.

1. Introduction

Education plays an important role in shaping the future of high school students, with academic performance serving as a critical indicator of their preparedness for the challenges that lie ahead [1, 2]. The academic journey from high school to university is a crucial transition period, and ensuring students are adequately prepared is essential for their success in higher education. In this context, predicting student performance becomes a key focus, as it enables educators and institutions to tailor interventions and support mechanisms to enhance the educational experience and outcomes for students entering university [3].

As high school students embark on their academic pursuits, the need to accurately predict their performance becomes imperative in guiding them towards a successful transition to university life [4]. Predictive models can assist in identifying areas of strength and weakness, enabling educators to offer targeted interventions that address specific academic needs. This not only enhances the overall learning experience for students but also ensures that they enter university with a solid foundation, better equipped to meet the academic challenges that await them [5].

Recent advancements in artificial intelligence (AI) and machine learning (ML) have revolutionized various domains [6–8], and education is no exception [9–11]. The application of AI and ML in predicting student performance has gained prominence, offering a data-driven approach to understanding and addressing academic challenges. These technologies allow for the analysis of vast amounts of data, including academic records, attendance patterns, and even social and emotional factors, providing a comprehensive view of a student's academic profile [12].

The benefits of leveraging machine learning and AI for predicting student performance are multifaceted [13, 14]. Beyond providing early intervention opportunities, these technologies enable educational institutions to allocate resources more efficiently, ensuring that support services are directed towards areas with the greatest need [15]. Moreover, the predictive analytics generated by ML algorithms offer valuable insights into the factors influencing student success, allowing for the development of targeted strategies to enhance overall educational outcomes [16].

This study seeks to address a specific research gap in the educational field: the need for a more accurate and efficient method to predict high school students' academic performance as they transition to university. The comparative analysis is crucial because it allows for a thorough evaluation of each algorithm's strengths and weaknesses in different aspects of educational data processing. This approach ensures that the chosen predictive model is not only robust but also tailored to meet the specific challenges and nuances of educational data, thereby providing educators and institutions with a more precise and effective tool for supporting students during this pivotal transition.

2. Materials and Methods

2.1. Dataset Description

The dataset originates from two Portuguese high schools, as documented in the study by Cortez and Silva, and is accessible through the University of California, Irvine (UCI) Machine Learning Repository [17, 18]. This comprehensive dataset incorporates a diverse range of attributes, including student grades, demographic information, social factors, and various school-related features from 395 students.

The data collection process involved the meticulous compilation of information through both school reports and questionnaires [19]. The inclusion of such a varied set of attributes enables a holistic examination of the factors influencing student performance, offering a solid foundation for our comparative analysis of machine learning algorithms in predicting academic success. The data specifically focuses on predicting mathematics performance as the target variable. The description of each feature in the dataset is presented in Table 1.

To process the categorical data types within this comprehensive dataset, we employ a label encoder. This step ensures that all the data is represented in a numeric format, aligning with the requirements of machine learning models. By transforming categorical variables into numerical equivalents, we enhance the model's capacity to extract meaningful patterns and relationships from the extensive set of attributes gathered during the data collection process.

2.2. Machine Learning Algorithms

To predict student performance, we selected six distinct machine learning models: linear regression, decision tree, support vector regression (SVR), k-nearest neighbors (KNN), random forest, and XGBoost, each chosen for their unique characteristics and proven effectiveness in various predictive modeling scenarios [20]. Linear regression offers a baseline for comparison due to its simplicity and interpretability. Decision trees and random forests are included for their ability to handle nonlinear relationships and feature interactions. SVR is chosen for its effectiveness in dealing with high-

Table 1. Dataset feature and description.

No.	Feature	Data Type	Description
1	school	Categorical	Student's school
2	sex	Categorical	Student's sex
3	age	Numeric	Student's age
4	address	Categorical	Student's home address type
5	famsize	Categorical	Family size
6	Pstatus	Categorical	Parent's cohabitation status
7	Medu	Numerical	Mother's education
8	Fedu	Numerical	Father's education
9	Mjob	Categorical	Mother's job
10	Fjob	Categorical	Father's job
11	reason	Categorical	Reason to choose this school
12	guardian	Categorical	Student's guardian
13	traveltime	Numeric	Home to school travel time
14	studytime	Numeric	Weekly study time
15	failures	Numeric	Number of past class failures
16	schoolsup	Categorical	Extra educational support
17	famsup	Categorical	Family educational support
18	paid	Categorical	Extra paid classes within the course subject
19	activities	Categorical	Extra-curricular activities
20	nursery	Categorical	Attended nursery school
21	higher	Categorical	Wants to take higher education
22	internet	Categorical	Internet access at home
23	romantic	Categorical	With a romantic relationship
24	famrel	Numeric	Quality of family relationships
25	freetime	Numeric	Free time after school
26	goout	Numeric	Going out with friends
27	Dalc	Numeric	Workday alcohol consumption
28	Walc	Numeric	Weekend alcohol consumption
29	health	Numeric	Current health status
30	absences	Numeric	Number of school absences
31	G1	Numeric	First period grade
32	G2	Numeric	Second period grade

dimensional spaces. KNN is included for its simplicity and efficacy in classification tasks. Finally, XGBoost is selected for its advanced capabilities in handling various types of data and its reputation for delivering high performance in prediction tasks.

To ensure robust evaluations, the dataset was split into an 80% training set and a 20% testing set [21, 22]. This division allowed us to train each model on the majority of the data and subsequently evaluate their predictive performance on unseen data, effectively gauging their generalization capabilities. The training-to-testing set ratio was maintained consistently across all models, fostering a fair and comparative analysis of their effectiveness in predicting high school student performance.

2.3. Model Evaluation

The evaluation of machine learning model performance is centered around three key metrics: R-squared (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). Each of these metrics offers a different perspective on the model's ability to predict student performance accurately and is critical for understanding the strengths and limitations of the applied algorithms [23].

R^2 measures the proportion of variance in the dependent variable that is predictable from the independent variables. It is a statistical metric that provides insight into the goodness of fit of the model. An R^2 score of 1 indicates a perfect fit, meaning the model predictions match the actual data exactly. In the context of this study, a high R^2 value would suggest that the model can account for a large portion of the variation in student performance.

RMSE is a metric to measure the error of a model in predicting quantitative data. It represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences. The RMSE metric is particularly useful because it gives a relatively high weight to large errors. This means the RMSE should be more useful when large errors are particularly undesirable.

Table 2. Performance of the machine learning models.

Model	R^2	RMSE	MAE
Linear Regression	0.88	1.64	1.14
Decision Tree	0.88	1.62	0.86
SVR	0.83	1.94	1.22
KNN	0.84	1.92	1.24
Random Forest	0.91	1.44	0.92
XGBoost	0.88	1.62	0.92

MAE is a measure of errors between paired observations expressing the same phenomenon. Comparisons between predictions and actual outcomes are often summarized using MAE. Unlike the RMSE, the MAE will not penalize large deviations as heavily as RMSE, making it more robust to outliers.

3. Results and Discussion

We successfully trained the machine learning model to predict the student performance, and the results of our machine learning models are presented in Table 2. The results demonstrate notable variations in the performance of the machine learning models across the evaluation metrics. Among the models, the Random Forest algorithm exhibited the highest performance, with an R^2 value of 0.91. This high R^2 score indicates a strong correlation between the predicted and actual student performance, suggesting that the model can explain a significant proportion of the variance in the data. Additionally, the Random Forest model achieved the lowest RMSE (1.44) and a relatively low MAE (0.92), highlighting its accuracy and precision in predicting student performance. The combination of these metrics suggests that the Random Forest model is particularly effective in capturing the complex relationships and patterns within the data, making it a robust tool for predicting student outcomes.

On the other hand, the SVR model showed the lowest performance among the evaluated models. With an R^2 of 0.83, it indicates a weaker correlation with the actual student performance compared to the other models. Furthermore, its higher RMSE (1.94) and MAE (1.22) values point to less accurate predictions. This could be due to the nature of the SVR algorithm, which might not be as effective in handling the specific characteristics and complexities of the dataset used in this study.

The results from the other models, such as Linear Regression, Decision Tree, and XGBoost, also provide valuable insights. Each of these models showed relatively high R^2 scores (0.88), suggesting their effectiveness in predicting student performance. However, variations in RMSE and MAE indicate differences in their accuracy and precision. For instance, the Decision Tree model had a notably lower MAE (0.86) compared to Linear Regression and XGBoost, which could make it a preferable option in scenarios where minimizing the absolute errors is crucial.

Conducting further analysis on the Random Forest model, which demonstrated superior performance in our study, we created a visualization to compare actual versus predicted student performance, as shown in Figure 1. The straight line represents the ideal scenario where the predicted grades exactly match the actual grades. Points that lie near this line are indicative of accurate predictions made by the Random Forest model. The distribution of points along the line of best fit demonstrates that, for the most part, the model's predictions are closely aligned with the actual performance of students, emphasizing its strong predictive capabilities.

The visualization reveals a high concentration of both training and testing set data points clustering near the line, which suggests that the Random Forest model has a consistent and reliable predictive performance across different subsets of the data. It's noteworthy that as we move towards the lower end of the grade spectrum, particularly at grade 0, there is a noticeable spread of points away from the line. This variance indicates that the model's predictions for students who received a grade of 0 are less accurate compared to other grades.

The deviation from the line at grade 0 might be attributed to various factors, such as imbalanced data concerning students with very low grades or the model's sensitivity to certain features that are not as prevalent or well-represented in these cases. This could suggest an area for improving the model, potentially by investigating the underlying features that lead to such outcomes or by incorporating more nuanced data that could help the model learn the patterns associated with these lower performance indicators more effectively.

Figure 2 presents the residual plot for the Random Forest model, which provides insights into the differences between the actual grades and the predicted grades by plotting the residuals on the y-axis against the actual grades on the x-axis. The residuals are calculated as the difference between the predicted values and the actual values, and ideally, they should be randomly distributed

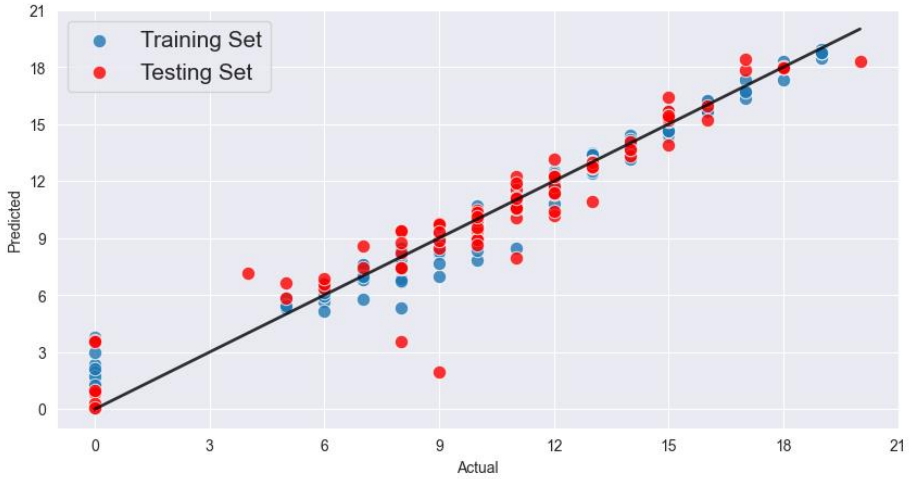

Figure 1. Actual vs. predicted plot of the Random Forest model.

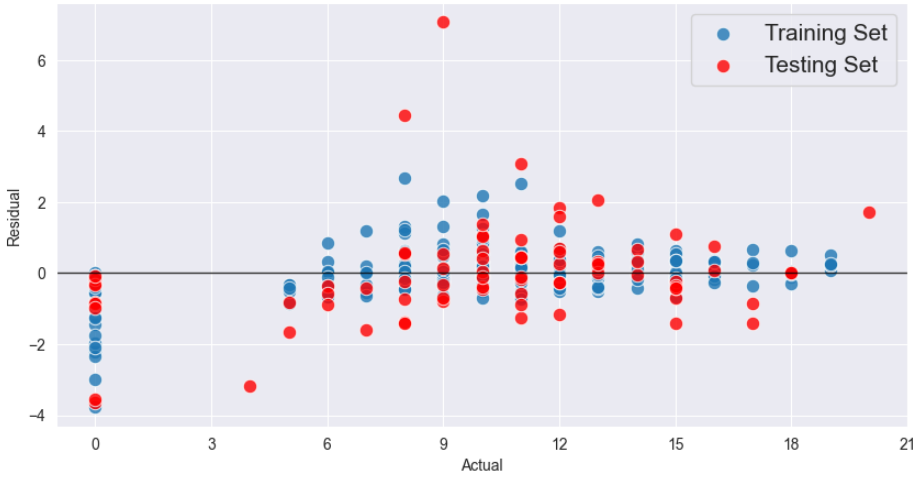

Figure 2. Residual plot of the Random Forest model.

around the horizontal line at zero, which indicates no error.

We observe that for most of the data points, the residuals fall within a relatively tight range around the zero line, both for the training set (blue) and the testing set (red). This indicates that the Random Forest model generally makes accurate predictions with small errors for the majority of cases. However, there are several points with larger residuals, particularly in the lower grade range (0-5), which suggest that the model's predictions deviate more significantly from the actual grades in these instances. It's notable that there aren't any clear patterns of systematic error; the residuals are spread above and below the zero line, indicating variability in the model's performance rather than a consistent bias.

Residuals are an important diagnostic tool because they help to reveal the presence of non-linear relationships that the model may not be capturing. The spread of residuals in the lower grade range could imply that the model is less adept at handling the underlying complexity within this segment of the data. This could potentially be addressed by exploring more complex models or feature engineering techniques that can better capture the nuances of lower-performing students' data. The absence of a discernible pattern in the residuals for higher grades suggests that the model is well-calibrated for these data points, making reliable predictions for students who perform within this range.

The implications of this study are far-reaching for the educational sector. By employing machine learning models, particularly the Random Forest algorithm, educators and policymakers can more accurately predict student performance and proactively implement tailored support strategies. This predictive capability is invaluable for identifying students who may be at risk of underperforming and allocating resources efficiently to areas where they are needed most.

In practical terms, the findings of this study can be integrated into educational management systems to monitor student progress in real-time. Schools and

educational institutions could utilize these predictive models to flag potential declines in student performance before they result in failure or dropout. This would enable timely interventions, such as tutoring or counseling, to assist students in overcoming academic challenges. Additionally, the insights gleaned from the model's predictions could inform curriculum development and teaching methodologies, leading to a more personalized learning environment.

Despite the promising results, this study is not without its limitations. The models were trained and validated on a dataset from two Portuguese high schools, which may not be representative of other educational contexts. Furthermore, the Random Forest model's predictive accuracy was less reliable for students receiving the lowest grades, indicating a potential bias towards students performing at an average or above-average level. This limitation could be due to the models not fully capturing the complexity of factors affecting student performance, such as psychological, behavioral, or external socioeconomic influences.

Future research should aim to address these limitations by incorporating a more diverse dataset that includes a broader demographic and socio-economic range. Including additional variables that could impact student performance, such as behavioral data or more nuanced socio-economic factors, may enhance the model's accuracy. Moreover, applying advanced machine learning techniques, such as deep learning or ensemble methods, could provide further improvements to the predictive models.

4. Conclusions

This study has demonstrated the potential of machine learning models to predict student performance effectively and underscores the importance of selecting the appropriate machine learning algorithm based on the specific requirements and characteristics of the dataset in educational settings. However, for these models to be practically applied in educational settings and to be truly transformative, future research will need to refine their predictive capabilities, ensure their applicability across diverse educational contexts, and explore their integration into real-time educational decision-making.

Author Contributions: Conceptualization, A.M. G.M.I. and H.S.; methodology, G.M.I., N.R.S. and A.R.; software, A.M.; validation, T.E.T., H.S. and A.R.; formal analysis, A.M., P.K. and N.B.M.; investigation, G.M.I. and N.R.S.; resources, G.M.I.; data curation, N.R.S.; writing—original draft preparation, A.M., P.K. and N.B.M.; writing—review and editing, G.M.I., T.E.T., H.S. and A.R.; visualization, N.R.S.; supervision, G.M.I.; project administration, G.M.I. All authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study is available at: https://archive.ics.uci.edu/dataset/320/student+performance.

Conflicts of Interest: All the authors declare that there are no conflicts of interest.

References

1. Foster, A., Shah, M., Barany, A., and Talafian, H. (2019). High school students' role-playing for identity exploration: findings from virtual city planning, *Information and Learning Sciences*, Vol. 120, No. 9/10, 640–662. doi:10.1108/ILS-03-2019-0026.
2. Idris, F., Hassan, Z., Ya'acob, A., Gill, S. K., and Awal, N. A. M. (2012). The Role of Education in Shaping Youth's National Identity, *Procedia - Social and Behavioral Sciences*, Vol. 59, 443–450. doi:10.1016/j.sbspro.2012.09.299.
3. Mengash, H. A. (2020). Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems, *IEEE Access*, Vol. 8, 55462–55470. doi:10.1109/ACCESS.2020.2981905.
4. Ha, D. T., Loan, P. T. T., Giap, C. N., and Huong, N. T. L. (2020). An empirical study for student academic performance prediction using machine learning techniques, *International Journal of Computer Science and Information Security (IJCSIS)*, Vol. 18, No. 3, 75–82.
5. Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., and Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review, *Applied Sciences*, Vol. 10, No. 3, 1042.
6. Noviandy, T. R., Maulana, A., Emran, T. B., Idroes, G. M., and Idroes, R. (2023). QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer's Disease Using Ensemble Machine Learning Algorithms, *Heca Journal of Applied Sciences*, Vol. 1, No. 1, 1–7. doi:10.60084/hjas.v1i1.12.
7. Maulana, A., Faisal, F. R., Noviandy, T. R., Rizkia, T., Idroes, G. M., Tallei, T. E., El-Shazly, M., and Idroes, R. (2023). Machine Learning Approach for Diabetes Detection Using Fine-Tuned XGBoost Algorithm, *Infolitika Journal of Data Science*, Vol. 1, No. 1, 1–7. doi:10.60084/ijds.v1i1.72.
8. Noviandy, T. R., Maulana, A., Idroes, G. M., Irvanizam, I., Subianto, M., and Idroes, R. (2023). QSAR-Based Stacked Ensemble Classifier for Hepatitis C NS5B Inhibitor Prediction, *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, IEEE, 220–225. doi:10.1109/COSITE60233.2023.10250039.
9. Idroes, G. M., Noviandy, T. R., Maulana, A., Irvanizam, I., Jalil, Z., Lensoni, L., Lala, A., Abas, A. H., Tallei, T. E., and Idroes, R. (2023). Student Perspectives on the Role of Artificial Intelligence in Education: A Survey-Based Analysis, *Journal of Educational Management and Learning*, Vol. 1, No. 1, 8–15. doi:10.60084/jeml.v1i1.58.
10. Maulana, A., Noviandy, T. R., Sasmita, N. R., Paristiowati, M., Suhendra, R., Yandri, E., Satrio, J., and Idroes, R. (2023). Optimizing University Admissions: A Machine Learning Perspective, *Journal of Educational Management and Learning*, Vol. 1, No. 1, 1–7. doi:10.60084/jeml.v1i1.46.
11. Noviandy, T. R., Idroes, G. M., Hardi, I., Emran, T. Bin, Zahriah, Z., Rahimah, S., Lala, A., and Idroes, R. (2023). Does Online Education Make Students Happy? Insights from Exploratory Data Analysis, *Journal of Educational Management and Learning*, Vol. 1, No. 2, 42–47. doi:10.60084/jeml.v1i2.124.

12. Katarya, R. (2019). A review: Predicting the performance of students using machine learning classification techniques, *2019 Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, IEEE, 36–41.

13. Namoun, A., and Alshanqiti, A. (2020). Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review, *Applied Sciences*, Vol. 11, No. 1, 237. doi:10.3390/app11010237.

14. Idroes, R., Subianto, M., Zahriah, Z., Afidh, R. P. F., Irvanizam, I., Noviandy, T. R., Sugara, D. R., Mursyida, W., Zhilalmuhana, T., and Idroes, G. M. (2023). Digital Transformations in Vocational High School: A Case Study of Management Information System Implementation in Banda Aceh, Indonesia, *Journal of Educational Management and Learning*, Vol. 1, No. 2, 48–54. doi:10.60084/jeml.v1i2.128.

15. Kour, S., Kumar, R., and Gupta, M. (2021). Analysis of student performance using Machine learning Algorithms, *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, IEEE, 1395–1403. doi:10.1109/ICIRCA51532.2021.9544935.

16. Asselman, A., Khaldi, M., and Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm, *Interactive Learning Environments*, Vol. 31, No. 6, 3360–3379. doi:10.1080/10494820.2021.1928235.

17. Cortez, P., and Silva, A. M. G. (2008). Using Data Mining to Predict Secondary School Student Performance.

18. Cortez, P. (2014). Student Performance, from https://archive.ics.uci.edu/dataset/320/student+performance, accessed 30-8-2023.

19. Fatimah, S., Farida, I., and Sukmawardani, Y. (2023). Interactive Learning for Water Pollution Awareness: A Game-Based Approach, *Journal of Educational Management and Learning*, Vol. 1, No. 1, 31–36. doi:10.60084/jeml.v1i1.52.

20. Caruana, R., and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms, *Proceedings of the 23rd International Conference on Machine Learning*, 161–168.

21. Idroes, R., Noviandy, T., Maulana, A., Suhendra, R., Sasmita, N., Muslem, M., Idroes, G. M., Kemala, P., and Irvanizam, I. (2021). Application of Genetic Algorithm-Multiple Linear Regression and Artificial Neural Network Determinations for Prediction of Kovats Retention Index, *International Review on Modelling and Simulations (IREMOS)*, Vol. 14, No. 2, 137.

22. Noviandy, T. R., Idroes, G. M., Maulana, A., Hardi, I., Ringga, E. S., and Idroes, R. (2023). Credit Card Fraud Detection for Contemporary Financial Management Using XGBoost-Driven Machine Learning and Data Augmentation Techniques, *Indatu Journal of Management and Accounting*, Vol. 1, No. 1, 29–35. doi:10.60084/ijma.v1i1.78.

23. Noviandy, T. R., Maulana, A., Idroes, G. M., Suhendra, R., Adam, M., Rusyana, A., and Sofyan, H. (2023). Deep Learning-Based Bitcoin Price Forecasting Using Neural Prophet, *Ekonomikalia Journal of Economics*, Vol. 1, No. 1, 19–25. doi:10.60084/eje.v1i1.51.