JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 9, No. 2, April 2025, pp. 467-481

Identification of Demographic Factors Affecting Student Performance using Tree-Based Machine Learning Models

Chatarina Enny Murwaningtyas
Department of Mathematics Education, Universitas Sanata Dharma, Yogyakarta, Indonesia
enny@usd.ac.id

Article History:
Received : 27-12-2024
Revised : 31-03-2025
Accepted : 03-04-2025
Online : 26-04-2025

Keywords:
Demographic Factors;
Academic Performance;
Educational Data Mining;
Tree-Based.

ABSTRACT

This study aims to identify key academic and demographic factors influencing student performance in the Logic and Set Theory course, particularly in the context of different learning modes during and after the COVID-19 pandemic. It adopts a quantitative exploratory design involving students from the 2020 to 2023 cohorts at Sanata Dharma University. Academic data (exam and assignment scores, course outcomes) and demographic data (e.g., parental education and income, region of origin, gender, and high school major) were collected from the academic system and supplemented via questionnaires. The dataset was cleaned, encoded, and normalized using RobustScaler, with class imbalance addressed through SMOTE. Descriptive statistics were used to explore initial data characteristics. Five tree-based machine learning models, Decision Tree, Random Forest, XGBoost, LightGBM, and CatBoost, were implemented within a pipeline that included preprocessing and model optimization using GridSearchCV with 5-fold cross-validation. Model evaluation employed multiple metrics, including accuracy, precision, recall, F1-score, AUC, and Average Precision. Results showed that XGBoost and CatBoost achieved the best performance (accuracy 92%, AUC 0.99) with balanced precision and recall across all four performance categories. Feature importance analysis indicated that exam and assignment scores were the strongest predictors, while demographic factors such as enrollment year, parental education, and income contributed moderately. Variables like gender, region, and high school major had minimal influence. This research demonstrates how machine learning can effectively integrate academic and demographic data, rather than analyzing them in isolation, to uncover nuanced patterns in student achievement. The findings support the development of data-driven educational interventions, such as preparatory learning modules, peer mentoring for underperforming groups, targeted academic advising for students from low-income or less-educated families, and flexible instructional strategies for cohorts affected by pandemic-related disruptions.

https://doi.org/10.31764/jtam.v9i2.28815

——————————— ◆ ———————————

A. INTRODUCTION

Student graduation is a crucial indicator of the success of higher education institutions and reflects their ability to educate and support students in their studies. A high graduation rate is often regarded as a marker of quality and competitiveness, demonstrating an institution's success in guiding students to complete their programs within the expected timeframe (Alhazmi & Sheneamer, 2023). Furthermore, student success during the first semester plays a significant role in determining educational persistence (Gil et al., 2021). Early academic performance not only reflects students' adaptation to the higher education environment, but also serves as an indicator of long-term academic success. Therefore, understanding the factors that influence students' academic achievement from the beginning of their studies is critical to supporting their educational continuity.

Various studies indicate that academic success and student graduation are influenced by a complex interplay of factors, including personal characteristics (Molnár & Kocsis, 2024), socioeconomic status (Bayirli et al., 2023), and demographics (Yusof et al., 2022), as well as other factors such as class attendance (Marshall, 2024), parental support (Jin, 2023), and academic achievement. Demographic factors, such as parents' education level (Isungset et al., 2022), family income (Marks & Pokropek, 2019), regional origin (Gimenez et al., 2018), gender (Lu et al., 2023), and age, have been proven to significantly impact students' academic performance. Students from better socioeconomic backgrounds or with parents with higher levels of education tend to have greater access to educational resources such as tutoring and supportive technology, which ultimately enhances their academic performance (Werang et al., 2024). Additionally, students' regional origins, whether urban or rural, influence the quality of the education they receive and their exposure to diverse learning opportunities (Zhao, 2022). Other sociodemographic factors, such as gender and age, also reveal disparities in academic achievement based on these factors (Early et al., 2023).

In recent years, the COVID-19 pandemic has posed new challenges to higher education institutions (Richards & Thompson, 2023). Online learning became the primary solution during the pandemic, followed by hybrid methods, and ultimately, full face-to-face learning. These changes significantly affected the students' learning experiences, especially in cohorts entering higher education during this period. Research has shown that online learning often reduces direct interaction between students and lecturers (Wut & Xu, 2021), limits access to physical facilities, and exacerbates technological disparities (Valentia, 2023). However, the transition to hybrid and full face-to-face learning has introduced new challenges in adaptation (Detyna et al., 2023) for both students and institutions. Therefore, it is important to explore how these changes have impacted academic outcomes (Sunarto, 2024), particularly in the first critical semester, which determines students' educational persistence.

In the context of higher education, Educational Data Mining (EDM) has become an increasingly popular tool for analysing student graduation patterns and their influencing factors (Barbeiro et al., 2024). EDM is an interdisciplinary field that develops methods for exploring unique data from educational environments to understand student characteristics and learning contexts. EDM techniques involve the application of machine learning algorithms to educational datasets, enabling the identification of patterns and trends that are not immediately visible. By leveraging EDM, educational institutions can design more effective interventions to support students and optimise teaching and learning strategies (Casillano & Cantilang, 2024).

Various classification methods in EDM have been used to predict students' academic performance, such as logistic regression (Zhang & Lu, 2024), K-nearest neighbours (Ritonga et al., 2024), and Naïve Bayes (Sembiring & Tambunan, 2021). However, tree-based models, such as decision trees (Bogdanov et al., 2024), random forests (Muminin et al., 2023), XGBoost (Hakkal & Lahcen, 2024), LightGBM (Huang, 2024), and CatBoost (Hancock & Khoshgoftaar, 2020), have garnered particular attention because of their combination of interpretability and predictive accuracy. Uddin and Lu (Uddin & Lu, 2024) support the superiority of tree-based algorithms, particularly Random Forest and Decision Tree, over non-tree algorithms in various performance metrics, such as accuracy, precision, recall, and F1-score. Additionally, Kumar et al. (2024) found that Random Forest and XGBoost consistently deliver high accuracy, whereas in certain contexts, CatBoost exhibits better stability in predictions. Mashagba et al. (2023) revealed that CatBoost achieved the highest accuracy of 92.16% in predicting students' final status compared with XGBoost and LightGBM, which each have strengths in speed and other aspects.

While the use of tree-based models in EDM has been widely explored, their application in analysing course-specific academic success—particularly in the first semester of Logic and Set Theory—remains limited. This is especially true when considering how learning mode transitions during the COVID-19 pandemic intersect with students' demographic characteristics. This research addresses this gap by focusing on a foundational course in mathematics education and integrating both academic and demographic dimensions.

This study aimed to identify the demographic factors that significantly influence students' academic success, particularly in Logic and Set Theory courses during their first semester. Additionally, it evaluated how learning conditions during and after the COVID-19 pandemic (online, hybrid, and full face-to-face) affected students' academic outcomes. Using tree-based machine learning algorithms, such as Decision Tree, Random Forest, XGBoost, LightGBM, and CatBoost, this study also seeks to determine the best model for predicting student graduation. This analysis is expected to offer comprehensive insights for designing targeted, data-driven educational interventions that support student success and improve graduation rates—both during regular academic years and in times of disruption.

B. METHODS

This study employed a quantitative exploratory research design to analyse the influence of demographic factors and learning conditions during and after the COVID-19 pandemic on the graduation outcomes of students in the Mathematics Education Study Program at Sanata Dharma University, particularly in the Logic and Set Theory course. The study involved students from the 2020 to 2023 cohorts with various learning modes: fully online for the 2020 cohort, hybrid for the 2021 cohort, and fully face-to-face for the 2022 and 2023 cohorts. The dataset comprised academic and demographic data from 207 students enrolled in the four cohorts.

The dataset included academic variables such as average assignment grades, average exam grades, and final course outcomes, as well as demographic variables, including region of origin, high school major, parents' education level (father's and mother's education), parents' occupations (father's and mother's occupations), parental income, number of siblings, gender, and age. Academic and demographic data were obtained from the university's information system. Missing demographic data were clarified using structured questionnaires sent to the students. Responses were validated by cross-checking administrative records when available. Cases with critical demographic information that could not be obtained or verified were excluded from the analysis.

Most demographic variables were categorical, except for the number of siblings and age, which were numerical variables. Parents' occupations were classified dichotomously into two groups: those working in education-related fields, such as teachers and lecturers, and those working in non-educational fields. The target variable, Course Outcome, was categorised into four levels: "Excellent" (grades A or A −), "Good" (grades B+, B, or B −), "Satisfactory" (grades C+ or C), and "Fail" (below passing grades).

After data collection, the information was integrated from multiple sources, combining grade lists presented per year with two classes each and merging academic data with demographic data based on student identifiers. Data cleaning involved handling missing data, as described above, by removing incomplete entries and identifying outliers in numerical variables using boxplots. Outliers were not removed but were normalised using RobustScaler (De Amorim et al., 2023) to reduce their influence while preserving the data integrity.

Categorical variables, including parents' occupations, region of origin, and high school major, were converted into numerical formats using encoding methods to ensure compatibility with machine learning algorithms. The cleaned dataset was subjected to descriptive analysis to determine the data characteristics. This analysis included calculating the means, medians, minimums, maximums, and standard deviations for numerical data, such as assignment and exam grades, and frequency distributions for the categorical data. Visualisations, such as histograms, depict the distribution of assignment and exam grades, whereas bar charts illustrate the distribution of categorical variables, including region of origin, high school major, parents' occupation, and gender.

The dataset was then split into two subsets, 70% for training and 30% for testing, using stratified sampling to maintain the class proportions of the target variable. Before applying SMOTE, the class distribution was imbalanced, with the "Fail" category underrepresented compared to "Good" and "Satisfactory." After applying SMOTE to the training set, each class was balanced to approximately equal proportions (25%), enabling better learning across all classes. Class imbalance was addressed using the Synthetic Minority Oversampling Technique (SMOTE), which was applied to the training data to enhance the representation of minority classes (Flores et al., 2022). SMOTE ensures that machine learning models can learn patterns equally from all classes.

A pipeline was developed to integrate data normalisation using RobustScaler, SMOTE oversampling, and model training. This pipeline was applied to five tree-based machine learning algorithms: Decision Tree, Random Forest, XGBoost, LightGBM, and CatBoost. These models were selected because of their robustness in handling tabular datasets with mixed data types, particularly in educational contexts where categorical variables dominate.

The decision tree served as the baseline model owing to its ability to map hierarchical patterns (Bogdanov et al., 2024), although it is prone to overfitting on complex datasets. Random Forest improves accuracy by aggregating multiple Decision Trees using ensemble methods, which also reduces the risk of overfitting (Sarker et al., 2024). Extreme Gradient Boosting (XGBoost) was implemented as a highly efficient boosting algorithm, iteratively correcting errors from previous models with regularisation features to prevent overfitting (Hakkal & Lahcen, 2024). A Light Gradient Boosting Machine (LightGBM) provides computational efficiency with its leaf-wise growth approach (Xi, 2024), whereas CatBoost directly handles categorical data without additional preprocessing, using ordered boosting to reduce bias during training (Hancock & Khoshgoftaar, 2020). The model parameters were optimised using GridSearchCV with 5-fold cross-validation to ensure the best performance of each algorithm (Odeh et al., 2023).

The model was evaluated using test data and various metrics to comprehensively assess its performance. The key metrics included accuracy, precision, recall, and F1-score, while additional metrics such as ROC-AUC (receiver operating characteristic (ROC) - area under curve (AUC) and Average Precision (AP) were employed to provide deeper insights into model prediction quality. The evaluation results were visualised using bar charts to facilitate comparisons of the algorithm performance across these metrics.

Once the best-performing model was identified, a feature importance analysis was performed to understand the contribution of each variable to the prediction. Feature importance techniques are tailored to specific algorithms. For the Decision Tree and Random Forest, feature importance was calculated based on impurity reduction (Gini impurity reduction). For XGBoost and LightGBM, feature importance was measured using the gain and average improvement in the objective function caused by the splitting of a particular feature. In contrast, CatBoost employs Shapley value-based feature importance to calculate the marginal contribution of each feature to the prediction process. This method was chosen because of its accuracy and relevance in handling categorical data, which dominate the research dataset. The research methodology is illustrated in Figure 1 to provide a visual representation of our workflow.

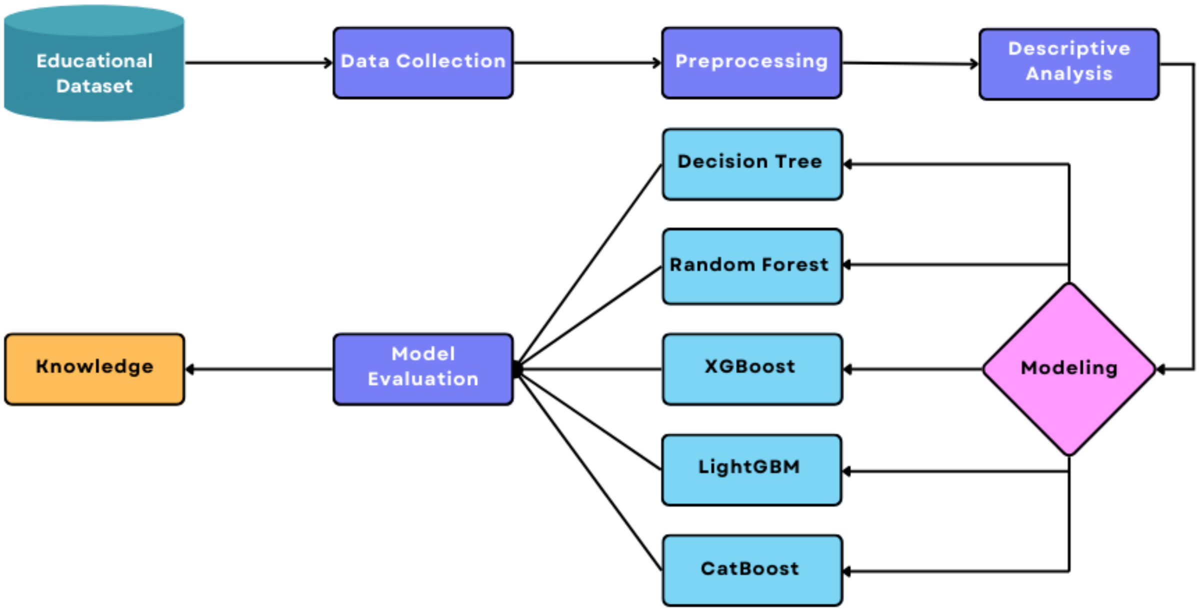

Figure 1. Research methodology workflow

This study aimed to identify the machine learning algorithm with the best performance in predicting student graduation outcomes and to determine the most significant independent variables influencing academic success. By combining rigorous preprocessing techniques, comprehensive evaluation metrics, and interpretable models, this methodology ensures that the resulting insights can effectively guide academic support policies and targeted interventions to improve educational outcomes.

C. RESULT AND DISCUSSION

1. Descriptive statistical analysis

This section presents an overview of the academic and demographic data distributions, which are essential for understanding the initial patterns before classification. Figure 2 presents a boxplot of academic scores, and Figure 3 presents the distribution of demographic characteristics.

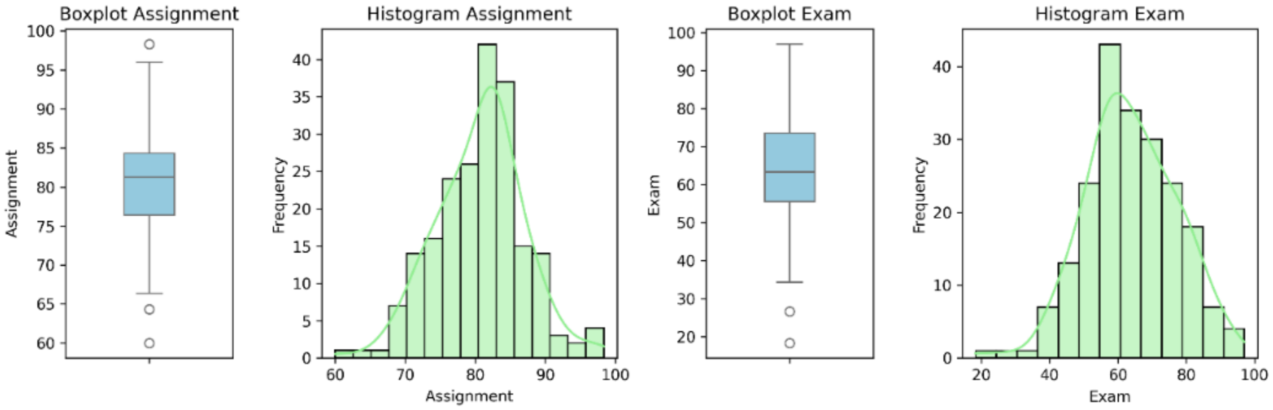

Figure 2. Boxplot of Assignment and Exam Score Distribution

The academic performance variables included assignment and exam scores, both of which are crucial for evaluating students' final outcomes. Assignment Scores represent the average of several tasks throughout the semester, whereas Exam Scores reflect the average of the midterm and final exams. As shown in Figure 2, the distribution of Assignment Scores was relatively symmetrical, with a median of 81.29 and a range of 66.36–96. The quartiles were 76.39 and 84.32. This suggests that most students performed well in their assignments, although the number and type of assignments varied across classes. Exam Scores, on the other hand, showed more dispersion, with a median of 63.33, a wider range (34.33 to 97), and lower quartiles (55.50–73.56), indicating higher variability. This indicates that exams posed greater challenges than assignments, possibly because of the complexity of the questions or time constraints during the examination period. The Shapiro-Wilk test results indicated that both variables exhibited distributions approaching normality, with test statistics of 0.992 (p = 0.347) for Assignment Scores and 0.994 (p = 0.512) for examination scores. However, Exam Scores displayed greater variability, reflecting the significant challenges associated with formal evaluations compared to assignments.

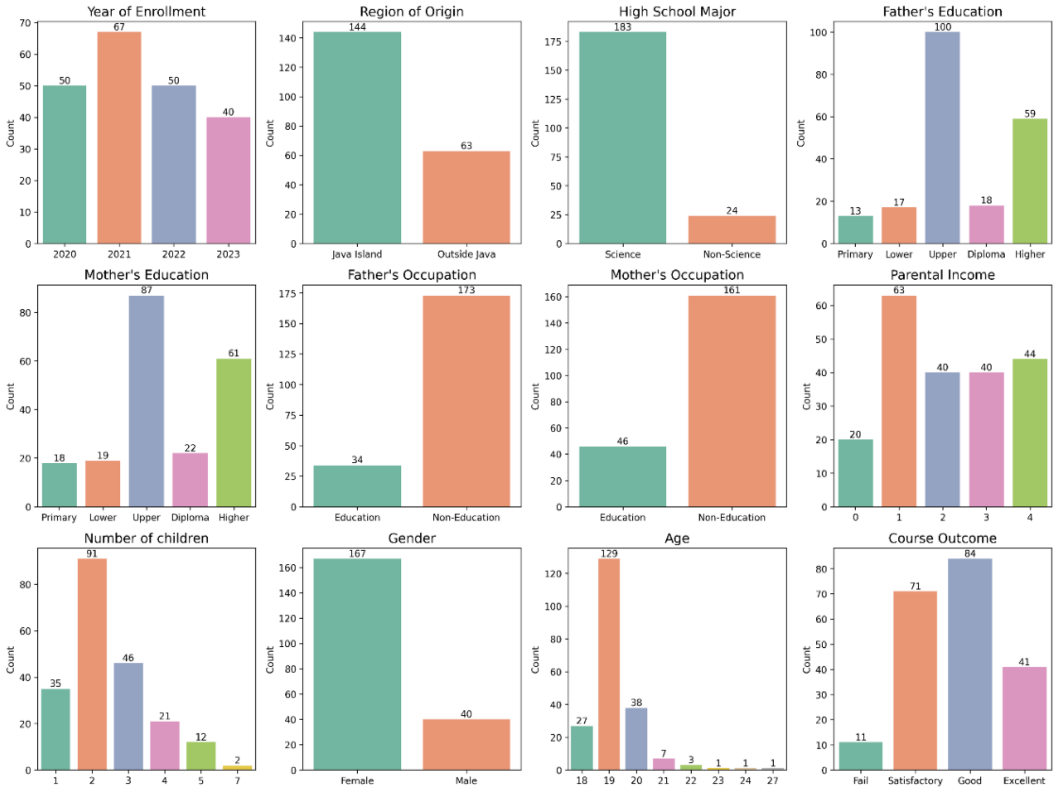

Figure 3. Bar Chart of Demographic Variables and Academic Outcomes

The demographic variables included cohort year, region of origin, high school major, parental education, parental occupation, family income, number of siblings, gender, and age. Figure 3 illustrates the distribution of these variables and their relationships with students' final outcomes. The majority of the students were from Java Island (69.6%), with the remaining 30.4% coming from outside Java. Most students had a science background (88.4%), whereas only 11.6% were non-science students. In terms of parental occupation, most students' parents did not work in the education sector, with only 16.4% of fathers and 22.2% of mothers employed in education-related fields. Family income was categorised into five groups, with 40.1% of students coming from families earning less than IDR 2,000,000 per month (categories 0 and 1). Another 38.6% belonged to the middle-income group (Categories 2 and 3), whereas 21.3% belonged to the high-income group (Category 4). Most students had one or two siblings (60.89%), while 39.01% had three or more siblings. Regarding age, the majority of students were between 18 and 19 years (75.36%), with an overall age range of 18–27. This reflects the dominance of recent high school graduates in the student population and suggests a relatively homogeneous age group that may influence group dynamics and peer learning.

When viewed together with academic outcomes (Figure 3), the distribution indicates that although most students passed, only 19.81% achieved "Excellent" while 5.31% were classified as "Fail". This highlights potential areas for institutional improvement, particularly among underperforming and high-achieving groups of students. The small percentage of students in the "Fail" category raises concerns about the specific barriers these students might face, such as a lack of academic support or socioeconomic disadvantage. Meanwhile, the fact that less than one-fifth of students reached "Excellent" suggests that the teaching strategy or assessment design may need refinement to better accommodate high achievers.

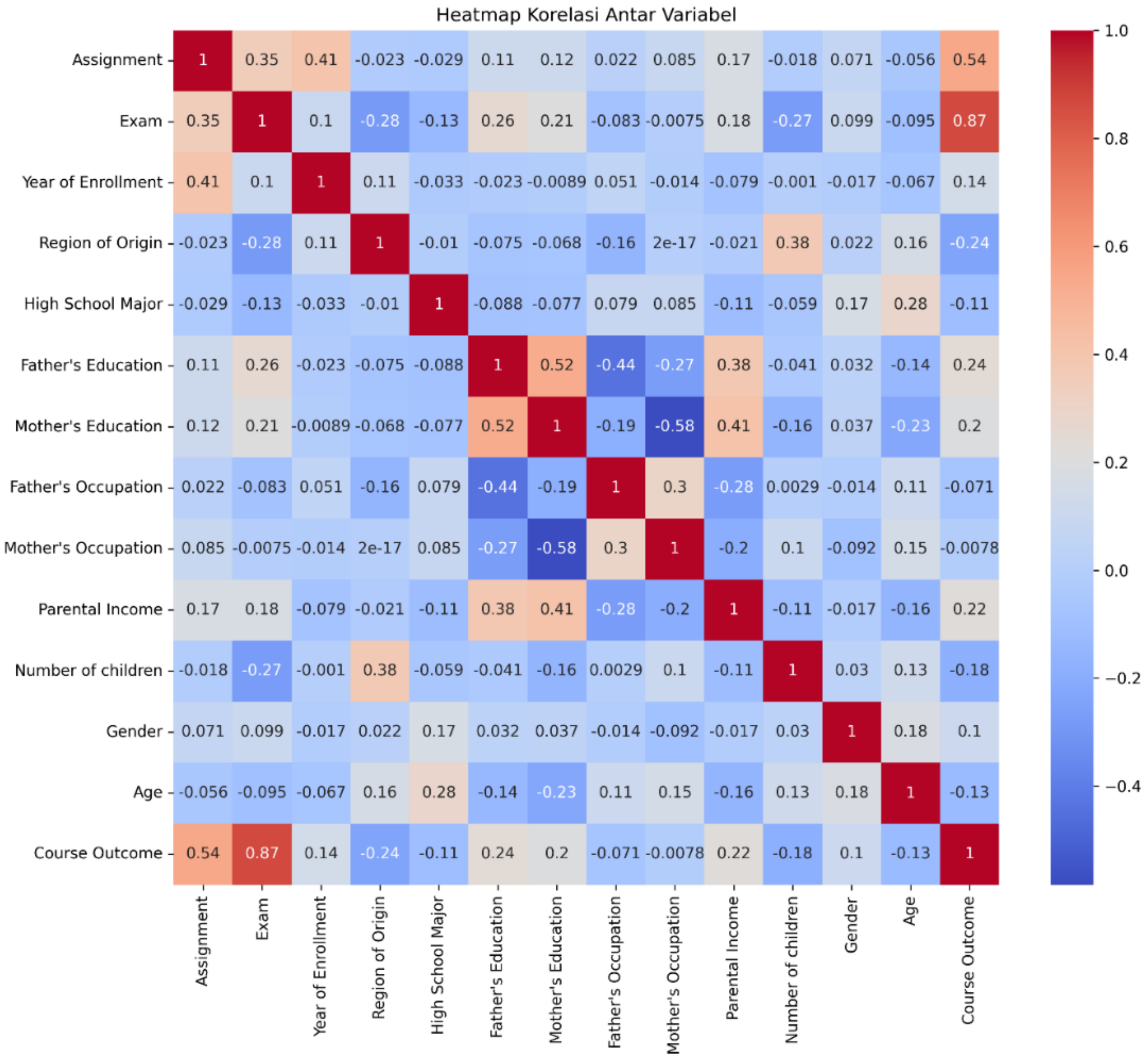

Figure 4. Correlation Heatmap of Academic and Demographic Variables

Figure 4 presents the correlations among variables, particularly with the target variable, Course Outcome, measured using Pearson's coefficients. Exam Scores showed the strongest correlation (r = 0.87), followed by Assignment Scores (r = 0.54). Weak negative correlations appeared for Region of Origin (r = -0.24) and Number of Siblings (r = -0.18), suggesting that students outside Java or with more siblings performed slightly worse. Other variables, such as Parental Income (r = 0.22), father's education (r = 0.24), and mother's education (r = -0.23), exhibited moderate correlations. While their direct impact is less than that of academic scores, they provide insights into socioeconomic influences. Variables with minimal direct influence (gender, father's and mother's occupation) remain integral to exploring interactions that potentially affect academic performance. These patterns reaffirm the importance of integrating demographic factors into academic risk analyses, as they often serve as underlying moderators of student outcomes. Overall, these demographic insights suggest targeted interventions, such as tailored academic support for students from lower-income families or regions outside Java to mitigate educational disparities and enhance equity.

2. Classification model analysis

To classify academic outcomes and explore variable contributions, this study applied five tree-based machine learning algorithms: Decision Tree, Random Forest, XGBoost, LightGBM, and CatBoost. These five tree-based algorithms were specifically chosen because of their robustness and superior performance in handling tabular and categorical data, effectively managing non-linear relationships and interactions between variables, and addressing class imbalances commonly found in educational datasets. Figure 5 compares the performance of the models across standard metrics.

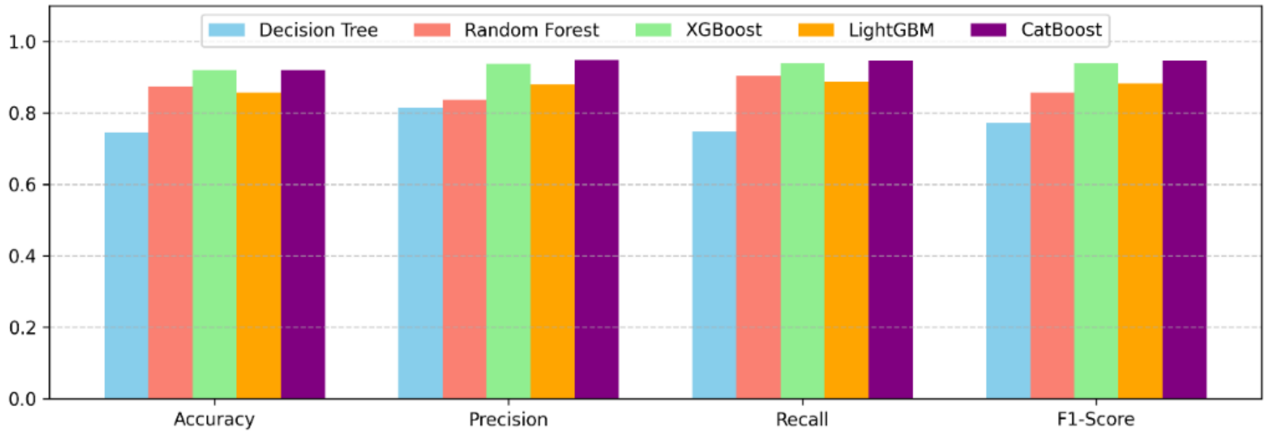

Figure 5. Comparison of Models Based on Accuracy, Precision, Recall, and F1-Score

The Decision Tree model provided an initial accuracy of 75% but demonstrated inconsistency in recall, particularly for the "Fail" category, which achieved a recall of only 0.67, despite having perfect precision (1.00) for this category. This indicates that although the Decision Tree effectively detects certain cases, it often misses predictions for minority classes. This pattern illustrates a trade-off: while Decision Trees are interpretable and simple to implement, they are prone to overfitting and poor generalisation in more complex datasets.

In contrast, the Random Forest model significantly improved the performance, achieving an accuracy of 87% and exhibiting more balanced metrics across all categories. Notably, the model demonstrated a strong recall for the "Excellent" and "Fail" categories, indicating its capability to capture important patterns in diverse datasets. It also handles categorical features more effectively than a single tree, making it a reliable model for moderately complex educational datasets. XGBoost was the best-performing model, with an accuracy of 92%. It achieved consistent precision, recall, and F1-scores across all categories, exceeding 0.90 in most cases. This performance was supported by a high Area Under Curve (AUC) ROC score of 0.99, reflecting the exceptional discrimination ability of the model.

Although LightGBM was slightly less accurate (86%) than XGBoost, it maintained a stable performance. The model excelled in the "Satisfactory" and "Good" categories but showed slight declines in the "Excellent" category. Similar to XGBoost, CatBoost achieved an accuracy of 92%

with consistent metrics across all categories. It recorded the highest F1-score (average 0.95) and an AUC of 0.99, making it one of the best models in this study.

A comparison of the accuracy, precision, recall, and F1-score metrics across all models (Figure 5) revealed that XGBoost and CatBoost consistently outperformed the other models. Random Forest demonstrated competitive performance, particularly in recall for minority classes. LightGBM performed nearly on par with Random Forest, whereas the Decision Tree lagged, highlighting its limitations in handling data complexity. The high performance of XGBoost and CatBoost can be attributed to their boosting framework and advanced regularisation techniques, which help prevent overfitting and improve generalisation, particularly in imbalanced datasets (Kumar et al., 2024; Odeh et al., 2023). However, these models involve more complex structures with less direct interpretability than simpler models, such as Decision Trees. Therefore, institutions prioritising transparent decision-making may prefer simpler models, despite their slightly lower predictive accuracy. Further visual evidence of the model performance differentiation is provided through the ROC and Precision-Recall curves (Figure 6), clearly demonstrating the superiority of the ensemble approaches.

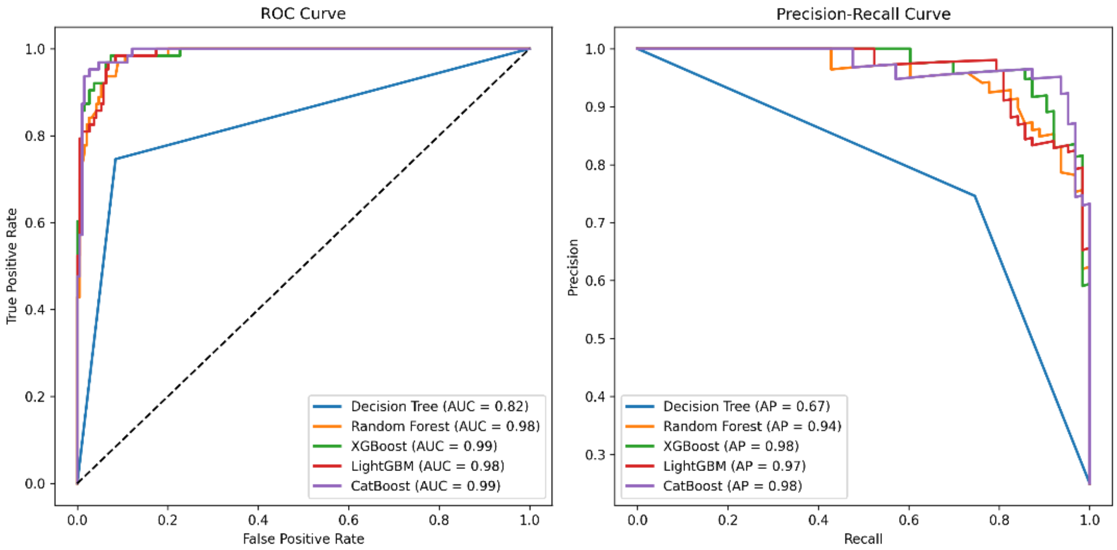

Figure 6. ROC and Precision-Recall Curves for Model Performance

On the ROC curve (Figure 6), XGBoost and CatBoost achieved the highest AUC scores (0.99), indicating their superior ability to differentiate between the classes. Random Forest and LightGBM followed closely, with AUC scores of 0.98, reflecting their competitive performance. The Decision Tree, with an AUC of 0.82, exhibited the lowest performance, underscoring the challenges in managing complex datasets. The Precision-Recall curve (Figure 6) displayed similar trends, where XGBoost, CatBoost, and Random Forest achieved high Average Precision (AP) scores of 0.98, 0.98, and 0.94, respectively, demonstrating their ability to maintain high precision even in imbalanced class distributions.

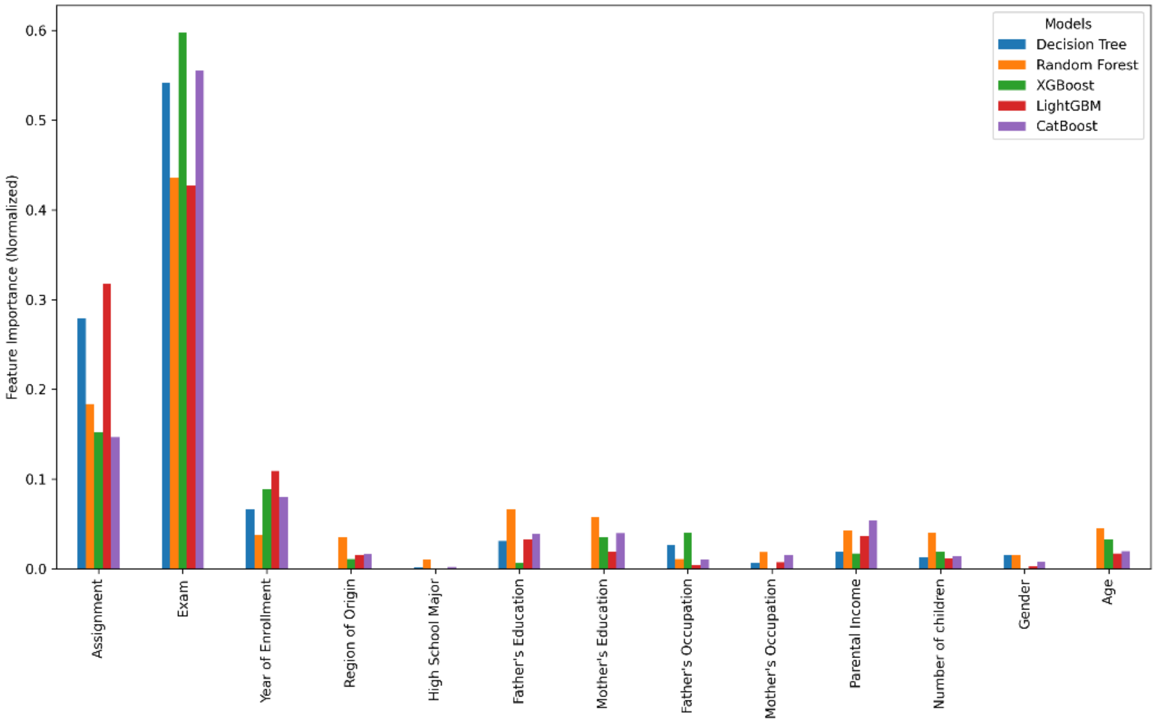

Figure 7. Comparison of Feature Importance Across Models

The analysis of feature importance (Figure 7) revealed that academic variables such as exams and assignments consistently emerged as the primary predictors of students' academic success across all machine learning models. Exam scores contributed the most to model predictions, particularly CatBoost (0.5558) and XGBoost (0.5978), underscoring the significance of exam performance as the strongest indicator of the learning outcomes. Assignment scores also played a crucial role, with the highest contribution observed in LightGBM (0.3174), reaffirming the impact of students' engagement in daily assignments on their learning achievements (Khairy et al., 2024).

Beyond academic features, socio-demographic variables, such as year of enrolment, mother's education, father's education, and Parental Income, provided moderate contributions in some models. The year of enrolment was particularly important in LightGBM (0.1093) and XGBoost (0.0887), reflecting the temporal relevance of learning contexts, particularly regarding the effects of the COVID-19 pandemic on learning patterns (Wang & Qing, 2023). Mothers' education (0.0578 in Random Forest and 0.0398 in CatBoost) and fathers' education (0.0663 in Random Forest) offer additional insights into the parental role in supporting educational success (Ludeke et al., 2021). Parental Income contributed moderately to CatBoost (0.0537) and LightGBM (0.0364), reflecting the influence of economic access on better learning opportunities. These findings align with those of Grätz and Wiborg (2020), who noted that socioeconomic disparities in academic performance are most pronounced among low-performing students, suggesting that students from low-income families face greater challenges in achieving academic success.

Conversely, features such as region of origin, high school major, gender, number of children, and age had very low or near-zero importance for all models. This indicates that these variables do not significantly contribute to the predictions and can be excluded to simplify the models without compromising their accuracy. Nevertheless, their presence may still be relevant in interaction terms or subgroup-specific analyses, especially when designing inclusive educational interventions. Retaining variables with low individual importance, such as gender and parental occupation, remains critical, as these variables may influence outcomes through interactions with other predictors, potentially uncovering hidden patterns or subgroup effects.

Models such as CatBoost and XGBoost excel because of their ability to capture complex patterns by effectively maximising the contributions of primary and supporting features, resulting in superior predictions for imbalanced datasets. Their boosting architecture allows for the iterative refinement of prediction errors, which is particularly useful when dealing with subtle, nonlinear relationships between features. Moreover, CatBoost's ability to handle categorical variables without extensive preprocessing and XGBoost's optimised gradient boosting approach provide both accuracy and efficiency. These strengths make them particularly suitable for educational datasets, which typically include mixed data types, missing values, and class imbalances.

These findings suggest that ensemble models, such as XGBoost and CatBoost, can classify students' academic performance, particularly for small and imbalanced datasets. Both models maintained high precision and recall for minority categories, such as "Fail" and "Excellent." Random Forest also showed competitive performance, particularly because of its ability to handle categorical features and prevent overfitting. In contrast, despite its simplicity and interpretability, the Decision Tree struggled with generalisation, as evidenced by its lower accuracy and inconsistent recall. Although slightly less performant than XGBoost and CatBoost, LightGBM remains a viable alternative because of its computational efficiency and speed. These observations highlight the critical balance that educational institutions must consider between interpretability and predictive accuracy when selecting machine learning models for practical implementation. Moreover, the consistency of academic features as dominant predictors across models confirms their role as core indicators of performance, whereas demographic factors, though less impactful individually, can offer valuable insight when interpreted contextually, especially in designing inclusive support systems.

D. CONCLUSION AND SUGGESTIONS

This study identifies exam and assignment scores as the primary factors influencing students' academic success in Logic and Set Theory courses, while demographic factors such as parental education, family income, and year of enrolment exert a moderate influence. Specifically, higher parental education positively affects students' academic outcomes by providing better academic support at home, fostering effective study habits, and creating a conducive learning environment. Family income also plays a crucial role, as students from lower-income families may face barriers such as limited access to resources, technology, or extracurricular learning support, ultimately affecting their academic performance. The year of enrolment, reflecting learning conditions during and after the COVID-19 pandemic (online, hybrid, or fully face-to-face), revealed that students from cohorts with online learning faced greater academic challenges than those in fully face-to-face learning environments. These findings underscore the significant impact of pandemic-induced changes in the learning environment and the subsequent transition to post-pandemic learning methods on students' academic outcomes.

Tree-based machine learning models, particularly XGBoost and CatBoost, demonstrated the best performance, with the highest accuracy of 92% and an AUC of 0.99, supporting previous theories regarding the superiority of ensemble algorithms in educational data analysis. XGBoost and CatBoost excel because of their capability to effectively handle complex interactions among variables, robustly manage imbalanced classes through iterative error correction, and reduce overfitting via regularisation and efficient handling of categorical features. These advantages make them particularly suitable for educational datasets, which often include a mixture of categorical and numerical variables, class imbalances, and subtle, non-linear relationships. This study contributes to educational data mining by revealing the short- and long-term impacts of the pandemic on student performance in foundational mathematics courses, while also demonstrating the effectiveness of tree-based machine learning models for designing data-driven interventions.

Educational institutions should consider designing adaptive policies to support student success, particularly during crises and post-pandemic recovery. Specifically, institutions could implement measurable interventions, such as providing targeted academic tutoring programs for students identified as vulnerable based on demographic profiles, enhancing technology access through subsidised or loaned equipment for lower income students, and conducting structured training sessions aimed at improving digital literacy and self-directed learning skills. Future research could explore additional variables, such as student motivation and engagement, or apply other machine learning models to further refine the prediction of academic outcomes. These steps can help build a more inclusive and effective educational environment for all students.

ACKNOWLEDGEMENT

This research was funded by the Institute for Research and Community Service of Universitas Sanata Dharma through the Master-Doctoral Research Grant Program. We extend our sincere gratitude for the invaluable support and funding, without which this research would not have been successfully conducted.

REFERENCES

Alhazmi, E., & Sheneamer, A. (2023). Early Predicting of Students Performance in Higher Education. *IEEE Access*, *11*, 27579–27589. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3250702

Barbeiro, L., Gomes, A., Correia, F. B., & Bernardino, J. (2024). A Review of Educational Data Mining Trends. *Procedia Computer Science*, *237*, 88–95. https://doi.org/10.1016/j.procs.2024.05.083

Bayirli, E. G., Kaygun, A., & Öz, E. (2023). An Analysis of PISA 2018 Mathematics Assessment for Asia-Pacific Countries Using Educational Data Mining. *Mathematics*, *11*(6), 1318. https://doi.org/10.3390/math11061318

Bogdanov, K., Gura, D., Khimmataliev, D., & Bogdanova, Y. (2024). Effectiveness of using Decision trees to increase student's analytical skills and cognitive development in education. *Interactive Learning Environments*, *33*(2), 1480–1489. https://doi.org/10.1080/10494820.2024.2372641

Casillano, N. F. B., & Cantilang, K. W. (2024). Employing educational data mining techniques to predict programming students at-risk of dropping out. *Indonesian Journal of Electrical Engineering and Computer Science*, *35*(2), 1219–1226. https://doi.org/10.11591/ijeecs.v35.i2.pp1219-1226

De Amorim, L. B. V., Cavalcanti, G. D. C., & Cruz, R. M. O. (2023). The choice of scaling technique matters for classification performance. *Applied Soft Computing*, *133*, 109924. https://doi.org/10.1016/j.asoc.2022.109924

Detyna, M., Sanchez-Pizani, R., Giampietro, V., Dommett, E. J., & Dyer, K. (2023). Hybrid flexible (HyFlex) teaching and learning: Climbing the mountain of implementation challenges for synchronous online and face-to-face seminars during a pandemic. *Learning Environments Research*, *26*(1), 145–159. https://doi.org/10.1007/s10984-022-09408-y

Early, E., Miller, S., Dunne, L., & Moriarty, J. (2023). The influence of socio-demographics and school factors on GCSE attainment: Results from the first record linkage data in Northern Ireland. *Oxford Review of Education*, *49*(2), 171–189. https://doi.org/10.1080/03054985.2022.2035340

Flores, V., Heras, S., & Julian, V. (2022). Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education. *Electronics*, *11*(3), 457. https://doi.org/10.3390/electronics11030457

Gil, P. D., Da Cruz Martins, S., Moro, S., & Costa, J. M. (2021). A data-driven approach to predict first-year students' academic success in higher education institutions. *Education and Information Technologies*, *26*(2), 2165–2190. https://doi.org/10.1007/s10639-020-10346-6

Gimenez, G., Martín-Oro, Á., & Sanaú, J. (2018). The effect of districts' social development on student performance. *Studies in Educational Evaluation*, *58*, 80–96. https://doi.org/10.1016/j.stueduc.2018.05.009

Grätz, M., & Wiborg, Ø. N. (2020). Reinforcing at the Top or Compensating at the Bottom? Family Background and Academic Performance in Germany, Norway, and the United States. *European Sociological Review*, *36*(3), 381–394. https://doi.org/10.1093/esr/jcz069

Hakkal, S., & Lahcen, A. A. (2024). XGBoost To Enhance Learner Performance Prediction. *Computers and Education: Artificial Intelligence*, *7*, 100254. https://doi.org/10.1016/j.caeai.2024.100254

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. *Journal of Big Data*, *7*(1), 94. https://doi.org/10.1186/s40537-020-00369-8

Huang, T. (2024). LightGBM model applied in the teaching course of Civic Education integrating red culture. *Applied Mathematics and Nonlinear Sciences*, *9*(1), 1–18. https://doi.org/10.2478/amns.2023.2.00173

Isungset, M. A., Conley, D., Zachrisson, H. D., Ystrom, E., Havdahl, A., Njølstad, P. R., & Lyngstad, T. H. (2022). Social and genetic associations with educational performance in a Scandinavian welfare state. *Proceedings of the National Academy of Sciences*, *119*(25), e2201869119. https://doi.org/10.1073/pnas.2201869119

Jin, X. (2023). Predicting academic success: Machine learning analysis of student, parental, and school efforts. *Asia Pacific Education Review*, 1–22. https://doi.org/10.1007/s12564-023-09915-4

Khairy, D., Alharbi, N., Amasha, M. A., Areed, M. F., Alkhalaf, S., & Abougalala, R. A. (2024). Prediction of student exam performance using data mining classification algorithms. *Education and Information Technologies*, *29*, 21621–21645. https://doi.org/10.1007/s10639-024-12619-w

Kumar, M., Singh, N., Wadhwa, J., Singh, P., Kumar, G., & Qtaishat, A. (2024). Utilizing Random Forest and XGBoost DataMining Algorithms for Anticipating Students'Academic Performance. *International Journal of Modern Education and Computer Science*, *16*(2), 29–44. https://doi.org/10.5815/ijmecs.2024.02.03

Lu, Y., Zhang, X., & Zhou, X. (2023). Assessing gender difference in mathematics achievement. *School Psychology International*, *44*(5), 553–567. https://doi.org/10.1177/01430343221149689

Ludeke, S. G., Gensowski, M., Junge, S. Y., Kirkpatrick, R. M., John, O. P., & Andersen, S. C. (2021). Does parental education influence child educational outcomes? A developmental analysis in a full-population sample and adoptee design. *Journal of Personality and Social Psychology*, *120*(4), 1074–1090. https://doi.org/10.1037/pspp0000314

Marks, G. N., & Pokropek, A. (2019). Family income effects on mathematics achievement: Their relative magnitude and causal pathways. *Oxford Review of Education*, *45*(6), 769–785. https://doi.org/10.1080/03054985.2019.1620717

Marshall, D. T. (2024). Student Attendance Patterns as Actionable Early Warning Indicators of High School Graduation Outcomes: Findings from an Urban Alternative Charter School. *Urban Science*, *8*(3), 78. https://doi.org/10.3390/urbansci8030078

Mashagba, E., Al-Saqqar, F., & Al-Shatnawi, A. (2023). Using Gradient Boosting Algorithms in Predicting Student Academic Performance. *2023 International Conference on Business Analytics for Technology and Security (ICBATS)*, 1–7. https://doi.org/10.1109/ICBATS57792.2023.10111325

Molnár, G., & Kocsis, Á. (2024). Cognitive and non-cognitive predictors of academic success in higher education: A large-scale longitudinal study. *Studies in Higher Education*, *49*(9), 1610–1624. https://doi.org/10.1080/03075079.2023.2271513

Muminin, R. S., Hadiana, A., & Natalia, N. (2023). The Study of Neural Network Algorithm, Random Forest for Classification of Student Graduation. *International Journal of Scientific Research in Science, Engineering and Technology*, *10*(3), 517–522. https://doi.org/10.32628/IJSRSET23103145

Odeh, A., Al-Haija, Q. A., Aref, A., & Taleb, A. A. (2023). Comparative Study of CatBoost, XGBoost, and LightGBM for Enhanced URL Phishing Detection: A Performance Assessment. *Journal of Internet Services and Information Security*, *13*(4), 1–11. https://doi.org/10.58346/JISIS.2023.I4.001

Richards, K., & Thompson, B. M. W. (2023). Challenges and instructor strategies for transitioning to online learning during and after the COVID-19 pandemic: A review of literature. *Frontiers in Communication*, *8*, 1–7. https://doi.org/10.3389/fcomm.2023.1260421

Ritonga, A., Masrizal, M., & Irmayanti, I. (2024). Analysis of Student Excellence Classes in Data Mining Using the KNN Method. *Sinkron*, *8*(2), 1148–1159. https://doi.org/10.33395/sinkron.v8i2.13627

Sarker, S., Paul, M. K., Thasin, S. T. H., & Hasan, Md. A. M. (2024). Analyzing students' academic performance using educational data mining. *Computers and Education: Artificial Intelligence*, *7*, 100263. https://doi.org/10.1016/j.caeai.2024.100263

Sembiring, M. T., & Tambunan, R. H. (2021). Analysis of graduation prediction on time based on student academic performance using the Naïve Bayes Algorithm with data mining implementation (Case study: Department of Industrial Engineering USU). *IOP Conference Series: Materials Science and Engineering*, *1122*(1), 012069. https://doi.org/10.1088/1757-899x/1122/1/012069

Sunarto, M. J. D. (2024). A Comparison of Students' Learning Outcomes in Advanced Mathematics Courses through Hybrid Learning. *Journal of Educators Online*, *21*(1), 1–11. https://doi.org/10.9743/JEO.2024.21.1.10

Uddin, S., & Lu, H. (2024). Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data. *PLOS ONE*, *19*(4), e0301541. https://doi.org/10.1371/journal.pone.0301541

Valentia, T. R. (2023). Digital Divide and Digital Literacy During the Covid-19 Pandemic. *Scriptura*, *13*(1), 69–78. https://doi.org/10.9744/scriptura.13.1.69-78

Wang, G., & Qing, X. (2023). Analyzing online and offline mixed teaching model for university students during and after COVID-19. *Interactive Learning Environments*, *32*(5), 1779–1794. https://doi.org/10.1080/10494820.2022.2127781

Werang, B. R., Agung, A. A. G., Sri, A. A. P., Leba, S. M. R., & Jim, E. L. (2024). Parental socioeconomic status, school physical facilities availability, and students' academic performance. *Edelweiss Applied Science and Technology*, *8*(5), 1–15. https://doi.org/10.55214/25768484.v8i5.1146

Wut, T., & Xu, J. (2021). Person-to-person interactions in online classroom settings under the impact of COVID-19: A social presence theory perspective. *Asia Pacific Education Review*, *22*(3), 371–383. https://doi.org/10.1007/s12564-021-09673-1

Xi, X. (2024). The role of LightGBM model in management efficiency enhancement of listed agricultural companies. *Applied Mathematics and Nonlinear Sciences*, *9*(1), 1–14. https://doi.org/10.2478/amns.2023.2.00386

Yusof, R., Hashim, N., Abdul Rahman, N., Mohd Yunus, S. Y., & Aziz Fadzillah, N. A. (2022). Academic Performance Prediction Model Using Classification Algorithms: Exploring the Potential Factors. *International Journal of Academic Research in Progressive Education and Development*, *11*(3), 706–724. https://doi.org/10.6007/ijarped/v11-i3/14753

Zhang, X., & Lu, H. (2024). Optimization of Practical Path of Teaching Reform in Higher Education—Based on Distributed Logistic Model Application. *Applied Mathematics and Nonlinear Sciences*, *9*(1), 1–17. https://doi.org/10.2478/amns-2024-1388

Zhao, K. (2022). Rural-urban gap in academic performance at a highly selective Chinese university: Variations and determinants. *Higher Education Research & Development*, *41*(1), 177–192. https://doi.org/10.1080/07294360.2020.1835836