ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 3, December 2024

DATA MINING ANALYSIS OF MOODLE LEARNING DATA AND STUDENT PERCEPTIONS DURING AND AFTER THE COVID-19 PANDEMIC

Maria Fatima Dineri De Jesus[1], Chatarina Enny Murwaningtyas[2*]

[1,2] Departement of Mathematics Education, Sanata Dharma University

email: mariafddejesus@gmail.com[1], enny@usd.ac.id[2*]

Abstract

This study examines the academic performance of students from the 2020 and 2023 cohorts, highlighting differences in activity, attendance, task completion, midterm and final exam scores, and perceptions of educational metrics. A data mining approach was applied to predict students' GPA using Decision Tree, Random Forest, Multinomial Naïve Bayes, and Gaussian Naïve Bayes algorithms. The Gaussian Naïve Bayes model showed the highest accuracy of 0.93 for the 2020 cohort and 0.92 for the 2023 cohort, with the lowest error rate making it the most effective predictor. Feature importance analysis revealed that task completion and exam scores were the most influential factors, while students' perceptions had a lesser impact. The findings suggest that direct academic metrics should be the focus for improving student performance. This study emphasizes the need for further refinement of predictive models and suggests incorporating both academic metrics and student perceptions for a holistic understanding of student performance.

Keywords : Moodle Data, Student Perceptions, Decision Tree, Random Forest, Naïve Bayes.

Received: 08-08-2024 | **Revised:** 30-10-2024 | **Accepted:** 05-11-2024
DOI: https://doi.org/10.23887/janapati.v13i3.84005

INTRODUCTION

Indonesia is one of the countries affected by the emergence of the 2019 Coronavirus Disease (COVID-19). One of the impacts of COVID-19 in Indonesia is the transformation of the education system, which previously occurred face-to-face in classrooms, to online learning. However, online learning faced several implementation challenges as it had never been adequately prepared [1]. Nonetheless, previous online learning methods had already combined electronic learning with face-to-face learning [2]. This has prompted various educational institutions, particularly higher education, to start enhancing diverse learning methods and modifying existing online learning.

E-learning is one of the widely applied learning methods in education, utilizing information and communication technology to support the teaching and learning process [3]. Al-Smadi et al. [4] described E-learning as a modern learning method because educational information is delivered to students through technology by educational institutions. Through this digital learning method, educational institutions that initially conducted face-to-face learning in classrooms began transitioning to using digital learning methods.

One of the platforms that support digital-based learning methods is the Learning Management System (LMS). LMS is software that facilitates and automates various learning management functions, from participant registration and material delivery to evaluation and reporting, in one integrated digital platform [5]. The purpose of creating an LMS is to manage learning using technology and information systems, where learning is more integrated with a web-based platform, and LMS is easily accessible as both open-source and commercial options [6], [7].

LMS continues to be used in learning even though the COVID-19 pandemic has passed. Istiqomah et al. [8] stated that although Indonesia has passed the COVID-19 pandemic, it does not mean that direct learning methods revert to conventional methods as before the pandemic. Instead, face-to-face learning is still combined with digital learning methods. Furthermore, LMS has evolved into blended learning after the pandemic, and the use of LMS-based e-learning media has proven to be feasible for future learning.

LMS has several open-source applications, including Moodle, which users can access to participate in learning via the web on a computer or mobile device [9]. Moodle offers facilities to support learning, such as learning modules, assignments, quizzes, chat features between teachers and participants, or among participants [10]. Moodle generates a vast amount of educational information data [11].

Research by Chen et al. [12] demonstrated that using a Moodle-based E-learning environment positively impacts improving collaborative learning and the academic achievement of engineering students. Romero et al. [13] also found that using Moodle and Socrative quizzes as formative aids greatly helps students prepare through learning materials, and students' initial impressions of using Moodle and Socrative quizzes significantly influence their future learning outcomes. Suparwito's [14] research found that LMS is very user-friendly, and activities in Moodle, such as discussion forums, are very useful for students, impacting their enthusiasm for learning and completing assignments. Therefore, students' perspectives on using Moodle are necessary as a reference for improving learning that positively impacts student learning outcomes.

Information in the Moodle LMS can be analyzed with data mining. Data mining can be applied in various fields to uncover hidden patterns and make predictions based on available data [15]. It involves various large-scale data exploration techniques with technological assistance, aiming to identify recurring patterns, trends, or rules that describe data characteristics in a specific context [16].

Dol & Jawandhiya [17] explained that classification is a part of data mining involving mapping data to predefined classes, also known as supervised learning. This process typically divides data into two parts: training data containing a set of attributes and classes, and test data used to evaluate the model. They found that classification techniques are often used in Educational Data Mining (EDM) to analyze student performance. Algorithms such as Naïve Bayes, Random Forest, and Support Vector Machine (SVM) are frequently employed due to their high effectiveness.

Research by Tamada et al. [18] indicated that the Random Forest algorithm provided the best results in predicting students at risk of dropping out from college based on student record data, achieving an F1 score of 84.47%. The Decision Tree algorithm followed closely, demonstrating nearly similar performance. Both Random Forest and Decision Tree have proven to be highly effective in educational data mining, particularly when dealing with high-dimensional datasets. Sianturi & Yuhana [6] further confirmed the strength of the Decision Tree algorithm, which achieved the highest accuracy (0.96 using an 80:20 data split) in detecting student learning styles within the Moodle LMS, outperforming Naive Bayes and K-Nearest Neighbor (KNN).

Meanwhile, research by Kika et al. [19] demonstrated that Naive Bayes, specifically in classifying student learning styles using Moodle log data, achieved an accuracy of 71.18%, higher than J48 Decision Tree and PART. For Naive Bayes, the selection of the specific variant depends on the nature of the data. Multinomial Naive Bayes is particularly suitable for handling categorical data with more than two categories [20], such as Likert scale responses, and not just limited to text data. Malhotra et al. [21] successfully applied Multinomial Naive Bayes in a non-text context, specifically in the classification of software defects, proving its applicability in handling categorical data with more than two categories. This demonstrates the flexibility of Multinomial Naive Bayes beyond text classification and its potential for use in various domains where categorical data is present, including but not limited to educational data mining. On the other hand, Gaussian Naive Bayes is used for continuous numerical data, such as test scores and attendance. The flexibility of these two variants allows the Naive Bayes algorithm to efficiently handle both types of data, making it a robust choice for this study.

The Mathematics Education Study Program at Sanata Dharma University continues to use Moodle in learning activities. However, an evaluation of learning outcomes using Moodle based on student surveys in this study program has never been conducted. This study involves data from 90 students, including Moodle data: activity, attendance, assignments, midterm exams, final exams, and perceptions: accessibility, participation, understanding, preparation, discipline, and responsiveness. The target variable is the final grade average of three courses, categorized into two groups: GPA below 3.0 and GPA above 3.0. The research focuses on three first-semester courses: Logic and Set Theory, Algebra and Trigonometry, and Plane Geometry, which are challenging to teach through online systems due to their high mathematical content.

This research employs four classification algorithm models: Decision Tree, Random Forest, Multinomial Naive Bayes, and Gaussian Naive Bayes. These algorithms are used to develop predictive models of student learning outcomes in two different groups: the 2020 cohort during the COVID-19 pandemic and the 2023 cohort after the pandemic. The aim is to compare these four algorithms and determine the dominant features in the predictive models. This study seeks to perform both classification and prediction analysis. The classification focuses on categorizing students into two groups based on their GPA (below 3.0 and above 3.0), while the prediction aspect seeks to forecast future student performance using key features identified through data mining. By comparing the results from different classification algorithms, this research not only classifies but also predicts which factors most significantly influence student outcomes. Identifying these features can be used as a consideration in the implementation of future learning. The classification results from this research provide valuable insights for educators in making informed decisions regarding the use of Moodle to support the learning process in the Mathematics Education Study Program.

The use of Learning Management Systems (LMS) such as Moodle has proven effective in supporting digital and blended learning, particularly during the COVID-19 pandemic. Previous studies have demonstrated improvements in student collaboration and academic outcomes through LMS use. This study aims to fill a gap by evaluating the impact of Moodle on student learning outcomes in the Mathematics Education program, applying data mining techniques to predict student performance. The results of this analysis will provide valuable insights for improving future teaching strategies.

METHOD

This research employs data mining techniques, focusing specifically on Decision Tree, Random Forest, Multinomial Naïve Bayes, and Gaussian Naïve Bayes model algorithms. These techniques are utilized to analyze Moodle data and student perceptions from undergraduate students in the Mathematics Education program at Sanata Dharma University, specifically from the 2020 and 2023 cohorts, totaling 90 students. The 2020 cohort represents students who experienced online learning during the COVID-19 pandemic, while the 2023 cohort represents students post-pandemic.

Moodle data includes student activity logs, attendance records, assignment submissions, midterm exam scores, and final exam scores. The activity data measures the percentage of completion of tasks or learning content provided on Moodle. This data was collected from three core courses: Algebra and Trigonometry, Plane Geometry, and Logic and Set Theory. These courses are first-semester courses that contain substantial mathematical content, making them challenging to teach online.

Student perceptions were collected using a validated questionnaire, which had been previously validated using data from the 2021 cohort. Key perception metrics include accessibility, participation, understanding, preparation, discipline, and responsiveness. Accessibility evaluates how easily students can access Moodle and course materials. Participation measures engagement in online discussions and activities. Understanding assesses students' comprehension of the course content. Preparation reviews the adequacy of materials and resources provided. Discipline looks at the ability to maintain study schedules, and responsiveness evaluates the quality of interaction between students and instructors. This comprehensive set of metrics provides a holistic view of the student learning experience during and after the COVID-19 pandemic.

The initial stage of the research involves data preprocessing to prepare the data for the mining process. This includes several critical steps, such as data cleaning and data transformation. Data cleaning addresses missing values, corrects inconsistencies, and removes irrelevant data entries that may skew the analysis results. Following data cleaning, data transformation steps are undertaken to code courses for a streamlined analysis process.

The data is subsequently split into training and testing sets, with 70% allocated for training and 30% for testing. This separation is conducted distinctly for the 2020 and 2023 cohorts to preserve the integrity of cohort comparisons. The split enables the training of predictive models on one subset of data while the other subset is used for model validation and performance assessment. The performance of the models is evaluated by calculating key metrics such as accuracy, precision, recall, F1-score, and confusion matrix values using standard equations, which provide a comprehensive assessment of each model's predictive capabilities [22], [23].

To implement these models, pipelines were constructed to streamline the process of data transformation, oversampling, and classification. Each pipeline includes steps for scaling the data using RobustScaler and MinMaxScaler, addressing class imbalance using SMOTE (Synthetic Minority Over-sampling Technique), and applying the respective classifier. This approach ensures that the models are well-prepared to handle the dataset's intricacies and improve predictive accuracy. RobustScaler scales the data by removing the median and scaling according to the interquartile range, making it robust to outliers [24]. MinMaxScaler scales each feature to a given range (often between zero and one), ensuring that all features contribute equally to the model's performance [25]. SMOTE addresses class imbalance by generating synthetic samples for the minority class, thereby balancing the class distribution and improving the model's ability to learn from imbalanced data [26], [27].

SMOTE is preferred over ADASYN and SVM-SMOTE due to its simplicity and ability to create synthetic samples through linear interpolation, balancing the dataset without adding unnecessary complexity. ADASYN, while adaptive, can lead to model overfitting by focusing on difficult-to-classify samples [28], and SVM-SMOTE increases computational demands by integrating SVM optimization with oversampling [29]. SMOTE provides a balance between effectiveness and ease of implementation, making it ideal for general use in handling imbalanced datasets.

GridSearchCV was used for hyperparameter tuning with Leave-One-Out Cross-Validation (LOO-CV) to find the optimal parameters for each model. This method is highly reliable for model selection, though computationally intensive, it ensures robust and unbiased performance evaluation [30]. Key hyperparameters tuned included the maximum depth, minimum samples split, and minimum samples leaf for Decision Tree; the number of estimators, maximum depth, minimum samples split, and minimum samples leaf for Random Forest; the alpha for Multinomial Naive Bayes; and the var_smoothing for Gaussian Naïve Bayes.

To identify the most influential features in predicting student outcomes, feature importance analysis was performed. This helps in understanding which features significantly impact the models' predictions and can inform future educational strategies. For Decision Tree and Random Forest models, feature importances were directly extracted from the models' attributes, reflecting the importance of each feature in making predictions. Features that result in the largest information gain or Gini impurity reduction are considered the most important. For Gaussian Naive Bayes, feature importance is measured using the variance of the means of the features across classes [31]. Features with higher variance are considered more important as they contribute significantly to the model's predictions. For Multinomial Naive Bayes, feature importance is determined by the range of log probabilities of the features across classes [32]. Features with larger ranges are more influential in making predictions.

These feature importance analyses are visualized using bar plots to provide a clear understanding of the most influential features in the dataset. The visualization helps in identifying which variables play the most significant roles in the models' predictions. This information is crucial for educators and administrators to enhance the learning experience and improve student outcomes using Moodle.

The methodology used in this research can be broadly described in the steps depicted in Figure 1. By following these steps, the research aims to provide a comprehensive analysis of student performance and engagement using data mining techniques, ultimately contributing to the improvement of educational strategies and learning environments.

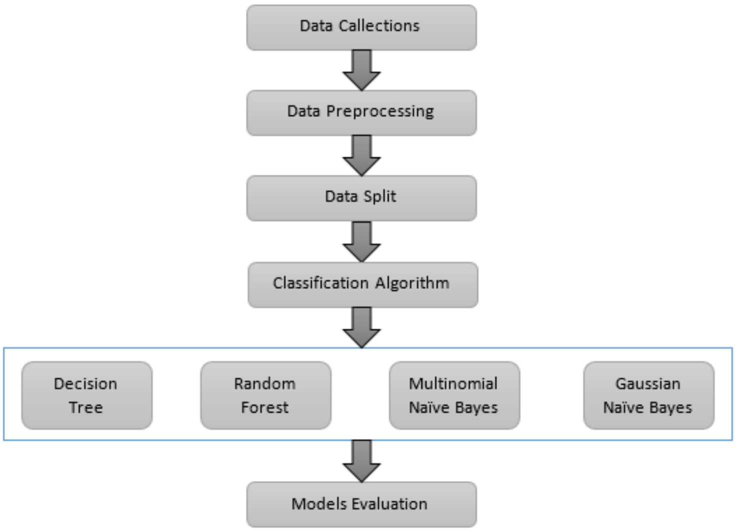


Figure 1. Methodology

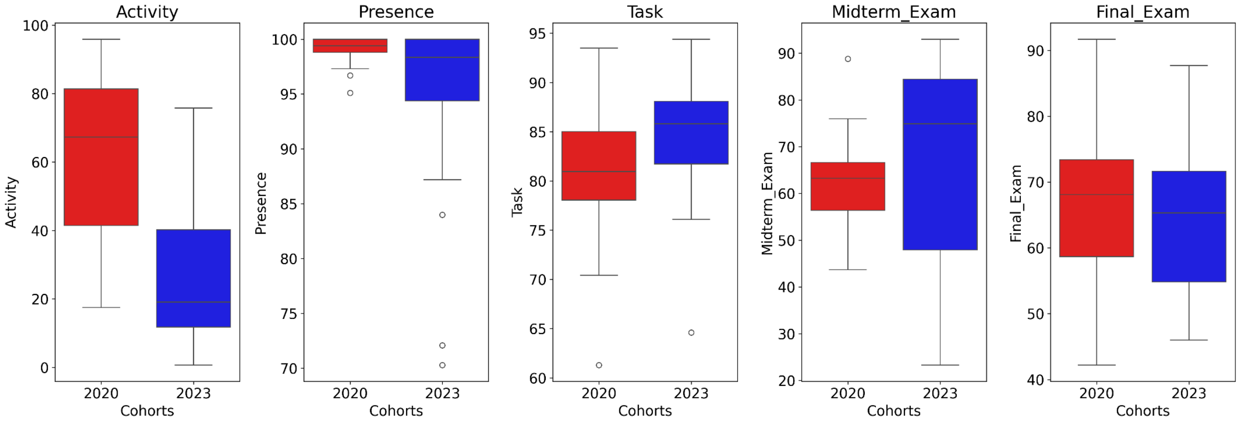


Figure 2. Boxplots comparing various educational metrics between the 2020 and 2023 cohorts.

RESULT AND DISCUSSION

The boxplots in Figure 2 provide a comparative analysis of various educational metrics between the 2020 and 2023 cohorts, offering insights into student engagement and performance during and after the COVID-19 pandemic.

The activity levels, represented by the percentage of task completions, show a significant difference between the two cohorts. The 2020 cohort, which experienced online learning during the pandemic, has a higher median activity level compared to the 2023 cohort. This indicates that students in the 2020 cohort were more engaged in completing tasks on Moodle. Additionally, the range of activity levels in 2020 is broader, suggesting more varied engagement among students. Conversely, the 2023 cohort has lower and more consistent activity levels, indicating that students post-pandemic may not be as engaged with the Moodle platform.

The presence metric, which measures attendance, also shows a noticeable difference between the two cohorts. The 2020 cohort has a higher median attendance rate, with a narrower interquartile range, indicating more consistent attendance among students. This can be attributed to the flexibility of attending classes online; students could participate in sessions via Zoom even if they were unwell or had other commitments. In contrast, the 2023 cohort, required to attend classes in person, shows a wider range of attendance rates. If students were sick or had other obligations, they were marked absent, leading to lower and more varied attendance.

In terms of task completion rates, the 2023 cohort has a higher median task completion rate compared to the 2020 cohort. This indicates that while the 2020 cohort was more active overall, the 2023 cohort was more consistent in completing assigned tasks. The wider interquartile range for the 2020 cohort suggests a greater variation in task completion among students during the pandemic.

The midterm exam scores show a substantial difference between the two cohorts. The 2023 cohort has higher median scores and a broader range of scores, indicating a wider distribution of performance. The 2020 cohort, on the other hand, has lower and more consistent midterm scores. This may suggest that the challenges of online learning during the pandemic affected students' performance in midterm exams.

Similarly, the final exam scores display a significant difference, with the 2023 cohort having higher median scores and a wider range of scores. Like the midterm exam results, the 2020 cohort shows lower and more consistent scores. This further supports the observation that the pandemic impacted students' performance in exams, potentially due to the lack of direct interaction and support from instructors.

The bar plots in Figure 3 illustrate the distribution of student perceptions on various metrics between the 2020 and 2023 cohorts. These metrics include accessibility, participation, understanding, preparation, discipline, and responsiveness. By comparing these distributions, we can gain insights into how students perceived their learning experiences during and after the COVID-19 pandemic.

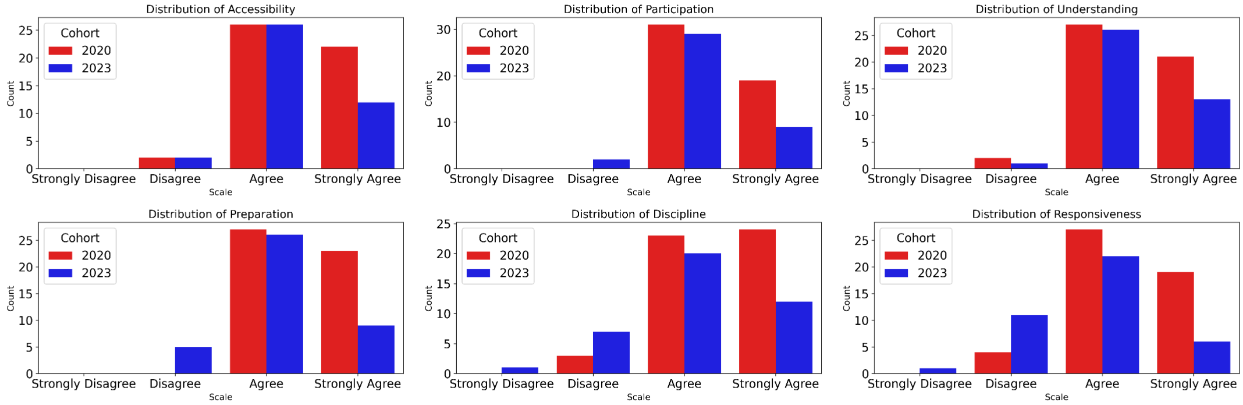


Figure 3. Bar plots comparing student perceptions on various metrics between the 2020 and 2023 cohorts.

The distribution of accessibility shows that both cohorts predominantly agreed that Moodle was accessible, with a significant number of students strongly agreeing. However, the 2020 cohort has a slightly higher count of students who strongly agree compared to the 2023 cohort. This suggests that during the pandemic, students might have found online platforms more accessible due to the necessity of relying on them for all educational activities.

In terms of participation, the 2020 cohort shows a higher count of students agreeing and strongly agreeing with their engagement in online discussions and activities compared to the 2023 cohort. The pandemic may have prompted higher participation as students sought to stay connected and engaged through online platforms. The lower participation in the 2023 cohort may indicate a shift back to in-person interactions, where online participation was less critical.

The understanding metric reveals that both cohorts have a high count of students agreeing with their comprehension of course content. However, the 2020 cohort again has a slightly higher count of students strongly agreeing compared to the 2023 cohort. This suggests that the extensive use of online resources and recorded lectures during the pandemic might have enhanced students' understanding of the material, as they could revisit content as needed.

The distribution of preparation shows that both cohorts generally agree and strongly agree that the materials and resources provided were adequate. However, similar to other metrics, the 2020 cohort has a slightly higher count of students strongly agreeing. This might be due to the comprehensive online resources made available during the pandemic, which were essential for remote learning.

For discipline, the 2020 cohort shows a higher count of students agreeing and strongly agreeing with their ability to maintain study schedules. The structured environment of online learning, with scheduled classes and deadlines, might have contributed to better discipline among students. In contrast, the 2023 cohort has a slightly higher count of students disagreeing with maintaining discipline, possibly due to the transition back to less structured in-person learning environments.

The responsiveness metric, which evaluates the quality of interaction between students and instructors, shows a higher count of students in the 2020 cohort agreeing and strongly agreeing compared to the 2023 cohort. During the pandemic, instructors may have been more proactive in engaging with students through online platforms to compensate for the lack of face-to-face interaction, leading to higher perceived responsiveness.

The analysis of these metrics indicates that the 2020 cohort, which experienced online learning during the pandemic, generally had higher levels of agreement and strong agreement across all perception metrics. This suggests that despite the challenges of remote learning, students found online platforms like Moodle accessible, engaging, and supportive of their learning needs.

However, the transition back to face-to-face learning in the 2023 cohort shows a slight decline in these metrics, possibly indicating that while traditional learning environments provide essential in-person interactions, they may lack some of the structured support and accessibility features that online platforms offer.

Educational institutions should consider integrating the strengths of both online and offline learning methods to provide a balanced and supportive learning environment. This blended approach can help maintain high levels of student engagement, understanding, and discipline while ensuring accessibility and responsiveness from instructors.

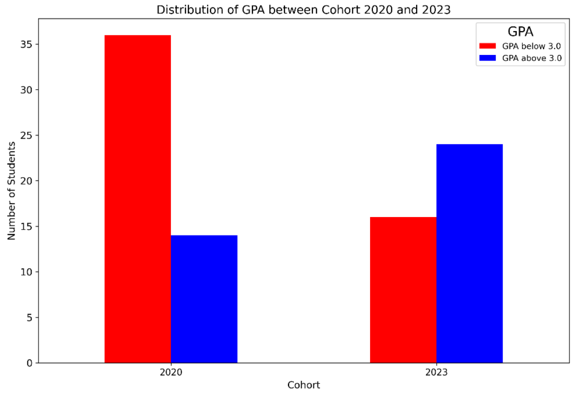


Figure 4. Distribution of GPA between Cohort 2020 and 2023.

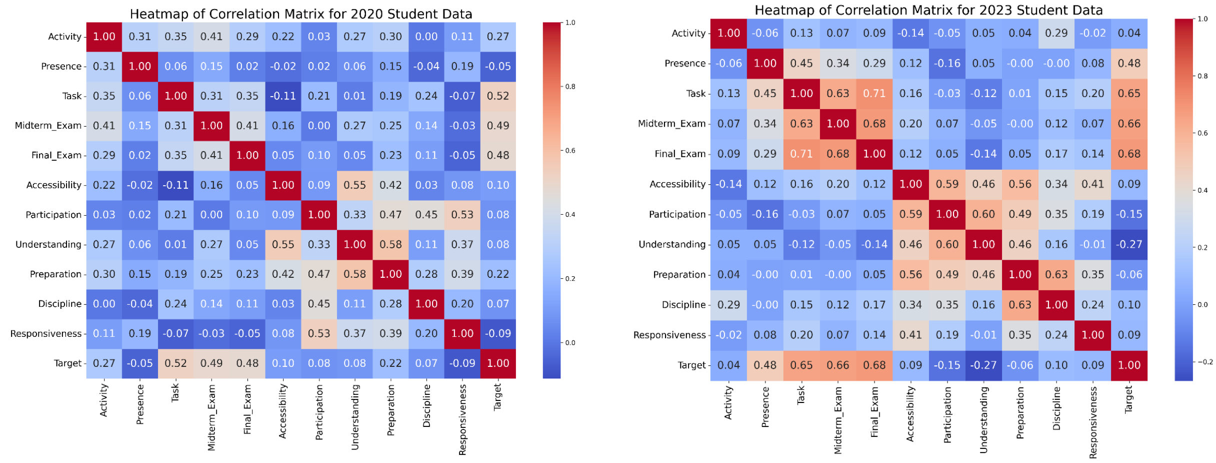
The analysis of student data from the 2020 and 2023 cohorts provides valuable insights into the impact of different learning environments on academic performance. Before delving into the details of the correlation analysis and model evaluations, it is essential to understand the distribution of GPAs across these cohorts.

The bar plot in Figure 4 illustrates the distribution of GPA for the 2020 and 2023 cohorts. It is evident that in the 2020 cohort, a significant majority of students have a GPA below 3.0, with a smaller proportion achieving a GPA above 3.0. Conversely, in the 2023 cohort, the distribution shows a higher number of students with a GPA above 3.0 compared to those with a GPA below 3.0.

This distribution indicates that the 2020 cohort, which experienced remote learning due to the COVID-19 pandemic, had more students struggling to achieve a GPA above 3.0 compared to the 2023 cohort, which returned to face-to-face learning. The higher proportion of students with lower GPAs in 2020 could be attributed to the challenges of online learning, such as difficulties in maintaining discipline, engagement, and effective comprehension of the course content.

The heatmaps in Figures 5 display the correlation matrices for the 2020 and 2023 student data, respectively, providing insights into the relationships between various features and student performance. These matrices help identify which features are most influential in determining student outcomes

In the 2020 student data heatmap (Figure 5.a), several features exhibit notable correlations with the target variable. Task Completion shows a strong positive correlation (0.52) with the target variable, indicating that the completion of tasks was a significant determinant of student performance during the pandemic. Similarly, Midterm Exam and Final Exam scores also display strong positive correlations with the target variable, at 0.49 and 0.48, respectively. These correlations suggest that exam scores were crucial indicators of student success in 2020. Additionally, Activity has a moderate positive correlation (0.27) with the target variable, implying that student engagement in Moodle activities was moderately linked to their performance. Other features, such as Presence, Accessibility, Participation, Understanding, Preparation, Discipline, and Responsiveness, exhibit weaker correlations with the target variable, indicating they might be less influential in predicting student performance.



(a) (b)

Figure 5. Heatmaps of correlation matrices (a) for the 2020 student data and (b) for the 2023 student data.

In contrast, the 2023 student data heatmap (Figure 5.b) reveals different patterns. Task Completion demonstrates an even stronger positive correlation (0.65) with the target variable compared to 2020, suggesting that completing tasks remained a critical predictor of student performance in the post-pandemic period. Midterm Exam and Final Exam scores also show strong positive correlations with the target variable, at 0.66 and 0.68, respectively, emphasizing the continued importance of exam performance in determining student success. Presence has a moderate positive correlation (0.48) with the target variable, higher than in 2020, indicating that attendance played a more significant role in student performance post-pandemic. However, Activity shows a weak positive correlation (0.04) with the target variable, suggesting that student engagement in Moodle activities had minimal impact on performance in 2023. Other features, such as Accessibility, Participation, Understanding, Preparation, Discipline, and Responsiveness, exhibit weaker or even negative correlations with the target variable, further indicating their lesser influence in predicting student performance.

From these correlation matrices, it is evident that certain features are more influential in determining student performance across both cohorts. Task Completion, Midterm Exam, and Final Exam scores consistently show strong positive correlations with the target variable, making them key predictors of student success. In the 2020 cohort, Activity is also moderately correlated with performance, highlighting the importance of engagement during online learning. Conversely, in the 2023 cohort, Presence becomes more significant, reflecting the shift back to in-person learning and the increased importance of attendance.

These findings suggest that predictive models for student performance should prioritize Task Completion, Midterm Exam, and Final Exam scores as primary features. Additionally, Presence should be considered for post-pandemic data, while Activity may be more relevant for data during the pandemic.

The analysis of the correlation matrices indicates that while several features are relevant in predicting student performance, Task Completion, Midterm Exam, and Final Exam scores are consistently the most influential. This insight can guide the development of predictive models, ensuring they focus on the most impactful features to accurately forecast student outcomes. Educational institutions can use this information to enhance learning strategies, emphasizing the importance of task completion and exam preparation to improve student performance.

The previous descriptive statistical analysis provided insights into the distribution of various educational metrics between the 2020 and 2023 cohorts. This analysis highlighted differences in student activity, presence, task completion, midterm exam scores, and final exam scores. To further understand how these metrics can predict student performance, data mining techniques were employed.

The previous descriptive statistical analysis revealed differences in educational metrics between the 2020 and 2023 cohorts, including student activity, attendance, task completion, exam scores, and perceptions of various factors. The analysis showed a higher number of students with a GPA below 3.0 in the 2020 cohort compared to the 2023 cohort, indicating an improvement in academic performance post-pandemic. Correlation analysis highlighted significant relationships between direct metrics like attendance, task completion, and exam scores with students' GPA, emphasizing their crucial role in determining academic performance.

To better understand how these metrics can predict student performance, data mining techniques were applied. The confusion matrices presented in Figure 6 and the data evaluation metrics summarized in Table 1 provide a comparative analysis of the performance of four different classification algorithms, Decision Tree, Random Forest, Multinomial Naïve Bayes, and Gaussian Naïve Bayes, on student data from the 2020 cohort. These matrices help understand the accuracy and error rates of each model in predicting whether a student's GPA is above or below 3.0.

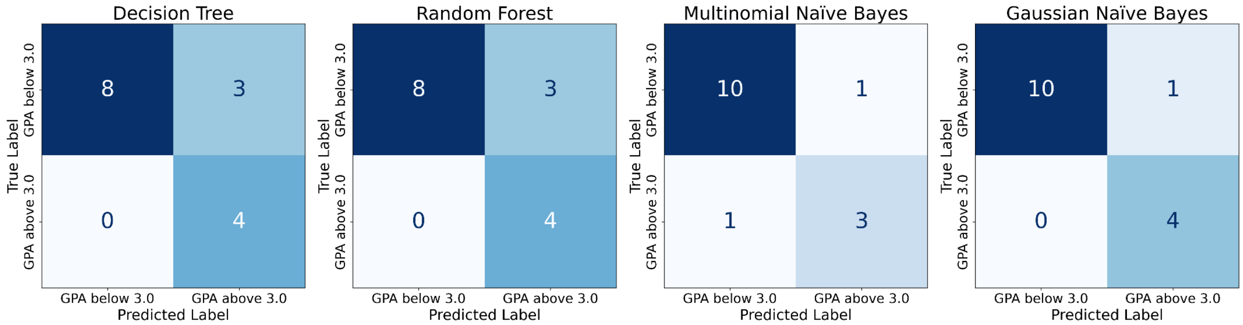


Figure 6. Confusion matrices for the 2020 cohort

Table 1. Data Evaluation Metrics for the 2020 Cohort

Models	Accuracy	Target	Precision	Recall	F1-Score	Support
Decision Tree	0.80	GPA below 3.0	1.00	0.73	0.84	11
		GPA above 3.0	0.57	1.00	0.73	4
Random Forest	0.80	GPA below 3.0	1.00	0.73	0.84	11
		GPA above 3.0	0.57	1.00	0.73	4
Multinomial Naïve Bayes	0.87	GPA below 3.0	0.91	0.91	0.91	11
		GPA above 3.0	0.75	0.75	0.75	4
Gaussian Naïve Bayes	0.93	GPA below 3.0	1.00	0.91	0.95	11
		GPA above 3.0	0.80	1.00	0.89	4

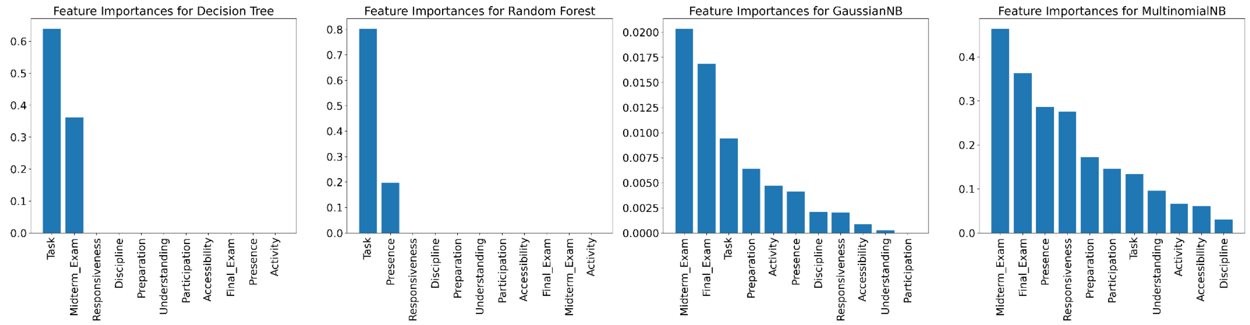


Figure 7. Feature importance for the 2020 cohort.

The Decision Tree model shows optimal parameters with a maximum depth of 2, a minimum leaf sample of 1, and a minimum split sample of 2. The training accuracy score reached 0.971, while the testing accuracy score reached 0.8. The model's precision is 0.886, with a recall of 0.8 and an F1 score of 0.811. The classification report shows that this model has a precision of 1.00 for the GPA category below 3.0 and a precision of 0.57 for the GPA category above 3.0.

The Random Forest model, with the same optimal parameters as the Decision Tree and an additional number of estimators of 2, shows a training accuracy score of 0.914 and a testing accuracy score of 0.8. The precision, recall, and F1 score for this model are the same as the Decision Tree, indicating similar performance in predicting student GPA categories.

The Multinomial Naïve Bayes model, with an optimal alpha parameter of 2.0, shows a training accuracy score of 0.771 and a testing accuracy score of 0.867. This model has a precision of 0.867, recall of 0.867, and an F1 score of 0.867. The classification report shows a precision of 0.91 for the GPA category below 3.0 and a precision of 0.75 for the GPA category above 3.0, indicating more consistent performance compared to the Decision Tree and Random Forest models.

The Gaussian Naïve Bayes model shows the best performance with an optimal var_smoothing parameter of 1e-09. The training accuracy score is 0.914, and the testing

ISSN 2089-8673 (Print) | ISSN 2548-4265 (Online)
Volume 13, Issue 3, December 2024

accuracy score is 0.933. This model has a precision of 0.947, a recall of 0.933, and an F1 score of 0.935. The classification report shows a precision of 1.00 for the GPA category below 3.0 and a precision of 0.80 for the GPA category above 3.0.

The conclusion from this analysis is that Gaussian Naïve Bayes is the most effective model for predicting student performance during the pandemic, with high accuracy and low error rates. The Multinomial Naïve Bayes also shows good predictive ability, while the Decision Tree and Random Forest, although effective, show higher false positive rates, indicating that these models may require further refinement for optimal performance.

The feature importance analysis provides further insights into the most influential features in each model. Figure 7 shows the feature importance for each model. For the Decision Tree and Random Forest, the most influential features are 'Task', 'Midterm Exam', and 'Presence'. Gaussian Naïve Bayes shows that 'Midterm Exam' and 'Final Exam' are the most influential features. Meanwhile, Multinomial Naïve Bayes highlights 'Midterm Exam', 'Final Exam', and 'Presence' as the main features.

The 'Task' feature indicates how well students complete assigned tasks, which is a direct indicator of their academic performance. 'Presence' indicates students' attendance in class, which is important for learning continuity. 'Midterm Exam' and 'Final Exam' are formal assessments that provide an overview of students' understanding of the material. This analysis shows that during the pandemic, direct metrics such as tasks and exams have a greater influence on predicting academic performance than students' perceptions of accessibility, participation, and other factors.

This confirms that in the 2020 cohort, student perceptions are not the primary determinants in the academic performance prediction model. Instead, direct metrics related to activity and academic outcomes play a more dominant role. This study shows that during the pandemic, focusing on tasks and formal assessments can provide a more accurate picture of students' academic performance.

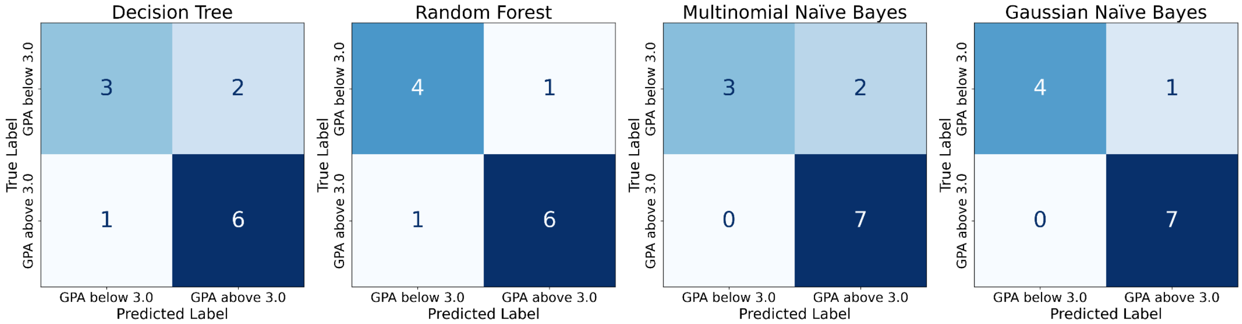


Figure 7. Confusion matrices for the 2023 cohort

Table 2. Data Evaluation Metrics for the 2023 Cohort

Models	Accuracy	Target	Precision	Recall	F1-Score	Support
Decision Tree	0.75	GPA below 3.0	0.75	0.60	0.67	5
		GPA above 3.0	0.75	0.86	0.80	7
Random Forest	0.83	GPA below 3.0	0.80	0.80	0.80	5
		GPA above 3.0	0.86	0.86	0.86	7
Multinomial Naïve Bayes	0.83	GPA below 3.0	1.00	0.60	0.75	5
		GPA above 3.0	0.78	1.00	0.88	7
Gaussian Naïve Bayes	0.92	GPA below 3.0	1.00	0.80	0.89	5
		GPA above 3.0	0.88	1.00	0.93	7

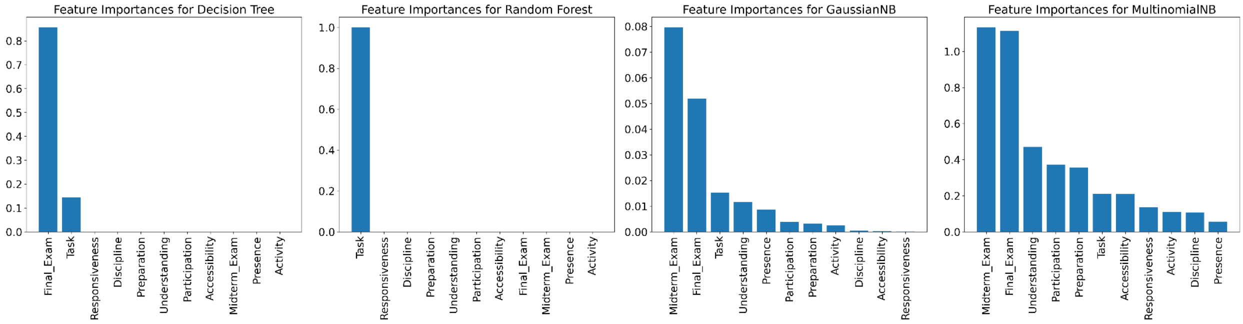


Figure 8. Feature importance for the 2023 cohort

The analysis for the 2023 cohort follows a similar pattern. The confusion matrices presented in Figure 7, the data evaluation metrics summarized in Table 2, and the feature importance graphs provide insights into the performance of the same four classification algorithms on the 2023 cohort.

The Decision Tree model shows optimal parameters with a maximum depth of 2, a minimum leaf sample of 2, and a minimum split sample of 2. The training accuracy score reached 0.964, while the testing accuracy score reached 0.75. The model's precision is 0.75, with a recall of 0.75 and an F1 score of 0.744. The classification report shows that this model has a precision of 0.75 for both GPA categories below and above 3.0.

The Random Forest model, with optimal parameters including a maximum depth of 1, a minimum leaf sample of 1, a minimum split sample of 2, and 1 estimator, shows a training accuracy score of 0.857 and a testing accuracy score of 0.833. The precision, recall, and F1 score for this model are 0.833, indicating consistent performance in predicting student GPA categories.

The Multinomial Naïve Bayes model, with an optimal alpha parameter of 0.01, shows a training accuracy score of 0.964 and a testing accuracy score of 0.833. This model has a precision of 0.870, recall of 0.833, and an F1 score of 0.823. The classification report shows a precision of 1.00 for the GPA category below 3.0 and a precision of 0.78 for the GPA category above 3.0.

The Gaussian Naïve Bayes model shows the best performance with an optimal var_smoothing parameter of 1e-09. The training accuracy score is 0.964, and the testing accuracy score is 0.917. This model has a precision of 0.927, a recall of 0.917, and an F1 score of 0.915. The classification report shows a precision of 1.00 for the GPA category below 3.0 and a precision of 0.88 for the GPA category above 3.0.

The conclusion from the 2023 cohort analysis is consistent with the 2020 cohort, with Gaussian Naïve Bayes being the most effective model for predicting student performance, followed by Multinomial Naïve Bayes. Decision Tree and Random Forest models also show good performance but require further refinement.

Figure 8 shows the feature importance for each model in the 2023 cohort. For the Decision Tree and Random Forest, 'Final Exam' and 'Task' are the most influential features. Gaussian Naïve Bayes shows that 'Midterm Exam' and 'Final Exam' are the most influential features. Multinomial Naïve Bayes highlights 'Midterm Exam' and 'Final Exam' as the main features.

Similar to the 2020 cohort, 'Final Exam', 'Midterm Exam', and 'Task' are direct indicators of academic performance. 'Midterm Exam' and 'Final Exam' provide a comprehensive assessment of students' understanding. The analysis confirms that direct metrics continue to play a dominant role in predicting academic performance, rather than student perceptions.

In conclusion, the analysis for both cohorts underscores the importance of direct academic metrics in predicting student performance. During and post-pandemic, focusing on tasks and formal assessments has been shown to provide a more accurate picture of students' academic performance.

CONCLUSION

The analysis conducted on the 2020 and 2023 student cohorts reveals significant insights into the factors that influence student performance. The study demonstrated that there are marked differences in academic metrics such as student activity, attendance, task completion, midterm exam scores, and final exam scores between the two cohorts. Notably, the GPA distribution showed an improvement in academic performance post-pandemic, with fewer students in the 2023 cohort having a GPA below 3.0 compared to the 2020 cohort.

The application of data mining techniques provided a deeper understanding of how these metrics predict student performance. The Gaussian Naïve Bayes model emerged as the most effective algorithm, achieving the highest accuracy of 0.93 and lowest error rates for both cohorts. This was followed closely by the Multinomial Naïve Bayes model with an accuracy of 0.87, which also showed strong predictive ability, while the Decision Tree and Random Forest models had similar accuracies 0f 0.80, for the 2020 cohort. In the 2023 cohort, the model with the highest accuracy was also the Gaussian Naïve Bayes model at 0.92, followed by the Multinomial Naïve Bayes and Random Forest models with similar accuracies of 0.83, while the Decision Tree model had an accuracy of only 0.75. This suggests that the model with the best accuracy across both cohorts was the Gaussian Naïve Bayes model.

The feature importance analysis revealed that direct academic metrics such as task completion, midterm exams, and final exams are the most influential factors in predicting GPA. This finding suggests that these metrics should be the primary focus in educational strategies to enhance academic performance. The study also highlights that student perceptions of accessibility, participation, understanding, preparation, discipline, and responsiveness were less significant in the predictive models, underscoring the dominance of direct performance indicators.

The improvement in academic performance in the 2023 cohort suggests that the return to in-person learning post-pandemic has had a positive impact. However, the dominance of direct academic metrics over student perceptions in predictive models suggests that educators should focus more on these direct indicators to gain a comprehensive understanding of student performance.

For future research, it is recommended to explore the impact of blended learning environments that combine the strengths of both online and offline learning methods. Additionally, further studies could investigate the long-term effects of the pandemic on student performance across different educational levels and subjects. By continuing to refine predictive models and incorporating a broader range of variables, educators can develop more effective strategies to support student success in various learning environments.

ACKNOWLEDGMENT

This research was funded by the Institute for Research and Community Service of Universitas Sanata Dharma through the Master Doctoral Research grant program. We express our deepest gratitude for the support and funding provided. Without this support, the research would not have been successfully conducted. We also extend our thanks to all parties who have contributed and supported this research, both directly and indirectly.

REFERENCES

[1] D. Y. Irawati and J. Jonatan, “Evaluasi Kualitas Pembelajaran Online Selama Pandemi Covid-19: Studi Kasus di Fakultas Teknik, Universitas Katolik Darma Cendika,” *Jurnal Rekayasa Sistem Industri*, vol. 9, no. 2, 2020, doi: 10.26593/jrsi.v9i2.4014.135-144.

[2] S. L. Nasution, F. Windari, S. Z. Harahap, and E. Elvina, “Pengaruh Media Pembelajaran Online Dalam Pemahaman Dan Minat Belajar Mahasiswa Pada Bidang Studi Akutansi Di Feb Universitas Labuhanbatu,” *ECOBISMA (Jurnal Ekonomi, Bisnis Dan Manajemen)*, vol. 8, no. 1, 2021, doi: 10.36987/ecobi.v8i1.2068.

[3] Y. S. Nuraeni and D. Irawati, “Growing Learning Motivation Through the Use of E- Learning Lms (Learning Management System) and Lecturer Competence,” *Procuratio : Jurnal Ilmiah Manajemen*, vol. 10, no. 3, 2022.

[4] A. M. Al-Smadi, A. Abugabah, and A. Al Smadi, “Evaluation of E-learning Experience in the Light of the Covid-19 in Higher Education,” in *Procedia Computer Science*, Elsevier, Jan. 2022, pp. 383–389.

[5] L. Vicent and M. Segarra, “Learning management system,” *Multimedia in Education*, pp. 21–48, 2010.

[6] S. T. Sianturi and U. L. Yuhana, “Student Behaviour Analysis To Detect Learning Styles Using Decision Tree, Naïve Bayes, And K-Nearest Neighbor Method In Moodle Learning Management System,” *IPTEK The Journal for Technology and Science*, vol. 33, no. 2, pp. 94–104, Aug. 2022.

[7] W. Widiyawati and Y. Anistyasari, “Studi Literatur Evaluasi Dan Pemeriksaan Fitur Alat Kuis Pada Learning Management System Berbasis Open Source,” *IT-Edu : Jurnal Information Technology and*

Education, vol. 5, no. 01, pp. 512–519, 2020.

[8] E. O. Istiqomah, A. Atiqoh, and Suhari, “Pengembangan Elkid Sebagai LMS Dengan Model Blended Learning di Era New Normal,” *Jurnal Teknologi Pendidikan : Jurnal Penelitian dan Pengembangan Pembelajaran*, vol. 8, no. 1, pp. 100–110, Jan. 2023, Accessed: Jan. 29, 2024. [Online]. Available: https://e-journal.undikma.ac.id/index.php/jtp/article/view/6219

[9] H. Dhika, F. Destiawati, M. Sonny, and Surajiyo, “Ease Evaluation Using the Best Moodle Learning Management System with Data Mining Concepts,” in *3rd International Conference on Learning Innovation and Quality Education*, Atlantis Press, Feb. 2020, pp. 944–952.

[10] E. Ikhsan, “Penerapan K-Means Clustering dari Log Data Moodle untuk Menentukan Perilaku Peserta pada Pembelajaran Daring,” *Sistemasi: Jurnal Sistem Informasi*, vol. 10, no. 2, pp. 414–422, May 2021.

[11] K. Lavidas *et al.*, “Predicting the Behavioral Intention of Greek University Faculty Members to Use Moodle,” *Sustainability*, vol. 15, no. 7, p. 6290, 2023.

[12] C. J. Chen, H. J. Tsai, M. Y. Lee, Y. C. Chen, and S. M. Huang, “Effects of a Moodle-based E-learning environment on E-collaborative learning, perceived satisfaction, and study achievement among nursing students: A cross-sectional study,” *Nurse Educ Today*, vol. 130, 2023, doi: 10.1016/j.nedt.2023.105921.

[13] E. Romero, L. García, and J. Ceamanos, “Moodle and Socrative quizzes as formative aids on theory teaching in a chemical engineering subject,” *Education for Chemical Engineers*, vol. 36, 2021, doi: 10.1016/j.ece.2021.03.001.

[14] H. Suparwito, “Student Perceptions and Achievements of Online Learning: Machine Learning Approaches,” in *AIP Conference Proceedings*, 2022. doi: 10.1063/5.0103688.

[15] G. Akçapinar and A. Bayazit, “Moodleminer: Data mining analysis tool for moodle learning management system,” *Elementary Education Online*, vol. 18, no. 1, 2019, doi: 10.17051/ilkonline.2019.527645.

[16] J. Calderon-Valenzuela, K. Payihuanca-Mamani, and N. Bedregal-Alpaca, “Educational Data Mining to Identify the Patterns of Use made by the University Professors of the Moodle Platform,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022, doi: 10.14569/IJACSA.2022.0130140.

[17] S. M. Dol and P. M. Jawandhiya, “Classification Technique and its Combination with Clustering and Association Rule Mining in Educational Data Mining — A survey,” 2023. doi: 10.1016/j.engappai.2023.106071.

[18] M. M. Tamada, R. Giusti, and J. F. de M. Netto, “Predicting Students at Risk of Dropout in Technical Course Using LMS Logs,” *Electronics (Switzerland)*, vol. 11, no. 3, 2022, doi: 10.3390/electronics11030468.

[19] A. Kika, L. Leka, S. Maxhelaku, and A. Ktona, “Using Data Mining Techniques on Moodle Data for Classification of Student’s Learning Styles,” 2019. doi: 10.20472/iac.2019.047.010.

[20] C. N. Egwim, H. Alaka, L. O. Toriola-Coker, H. Balogun, and F. Sunmola, “Applied artificial intelligence for predicting construction projects delay,” *Machine Learning with Applications*, vol. 6, p. 100166, Dec. 2021, doi: 10.1016/j.mlwa.2021.100166.

[21] R. Malhotra and M. Cherukuri, “Software Defect Categorization based on Maintenance Effort and Change Impact using Multinomial Naïve Bayes Algorithm,” in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2020, pp. 1068–1073. doi: 10.1109/ICRITO48877.2020.9198037.

[22] D. M. W. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation,” *arXiv preprint arXiv:2010.16061*, 2020.

[23] S. Shrestha and M. Pokharel, “Educational data mining in moodle data,” *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 10, no. 1, 2021, doi: 10.11591/ijict.v10i1.pp9-18.

[24] “RobustScaler — scikit-learn 1.5.1 documentation.” Accessed: Aug. 06, 2024. [Online]. Available: https://scikit-

learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html

[25] "MinMaxScaler — scikit-learn 1.5.1 documentation." Accessed: Aug. 06, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

[26] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," 2002.

[27] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from imbalanced data sets*, vol. 10, no. 2018. Springer, 2018.

[28] H. Marlisa, N. Satyahadewi, N. Imro'ah, and N. N. Debataraja, "Application of ADASYN Oversampling Technique on K-Nearest Neighbor Algorithm," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 18, no. 3, pp. 1829–1838, Jul. 2024, doi: 10.30598/BAREKENGVOL18ISS3PP1829-1838.

[29] J. Guo, H. Wu, X. Chen, and W. Lin, "Adaptive SV-Borderline SMOTE-SVM algorithm for imbalanced data classification," *Applied Soft Computing Journal*, vol. 150, 2024, doi: 10.24433/CO.4255147.v1.

[30] Z. Shao and M. J. Er, "Efficient Leave-One-Out Cross-Validation-based Regularized Extreme Learning Machine," *Neurocomputing*, vol. 194, pp. 260–270, Jun. 2016, doi: 10.1016/j.neucom.2016.02.058.

[31] J. H. Fan, X. Y. Li, and S. H. Teng, "Research on spam message recognition algorithm based on improved naive Bayes," in *Proceedings - 2022 International Conference on Intelligent Transportation, Big Data and Smart City, ICITBS 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 241–244. doi: 10.1109/ICITBS55627.2022.00059.

[32] M. L. Fotteler *et al.*, "Use and benefit of information, communication, and assistive technology among community-dwelling older adults–a cross-sectional study," *BMC Public Health*, vol. 23, no. 1, p. 2004, 2023.