Tripoli College of Science and Technology
Journal of Tripoli College of Science and Technology
Peer_Reviwed Semi-annual Journal
Volume One - Issue Tow - January 2026
Legal deposit number for the Journal (2025/148) National House Benghazi - Libya

Exploring Student Performance Using Machine Learning: Effects of Attendance, Weekly Quizzes, and Midterm Exams

استكشاف أداء الطلاب باستخدام تعلم الآلة: تأثير الحضور والاختبارات الأسبوعية والاختبارات النصفية

Hanan Issa Ashtewi
Assistant Lecturer
Tripoli College of Science and Technology - Department of Computer
Ashtewi@tcst.edu.ly

حنان عيسى اشتيوي
مساعد محاضر
كلية طرابلس للعلوم والتقنية
قسم الحاسب الآلي

Rabab Mohammed Alztaf
Assistant Lecturer
Tripoli College of Science and Technology - Department of Computer
Alztaf@tcst.edu.ly

رباب محمد الزطاف
مساعد محاضر
كلية طرابلس للعلوم والتقنية
قسم الحاسب الآلي

Abstract

This research aims to explore the most influential factor of Absence, Weekly-quizzes and Midterm-Exam on Final-Exam score of 164 female student enrolled in first semester in fundamental information technology course. Using descriptive statistics and Pearson correlation analysis, K-Means clustering and regression models (Decision Tree, Random Forest, Gradient Boosting).

The Pearson correlation results show the strongest correlation was between Midterm-Exam and Final-Exam (r=0.667), followed by

Weekly-Quizzes (r=0.612), while Absence had low negative correlation with (r== –0.359).

K-Means clustering revealed variation in students' performance and were classified into three groups high, average and low performance, while Random Forest achieved the most accurate prediction with (R^2=0.62) and least prediction errors.

SHAP analysis confirmed that Midterm-Exam was most influential factor on Final-Exam score, followed with Weekly-Quizzes, while Absence had indirect effect on Final-Exam score. In conclusion, this study confirms the importance of continuous assessment to predict students at risk and providing appropriate support in the right time to enhance the overall education results.

Key words: students' performance, educational data mining, K-Means clustering, Random Forest, SHAP analysis.

الملخـص

تهدف هذه الدراسة إلى استكشاف العوامل الأكثر تأثيرًا الغياب والاختبارات القصيرة والامتحان النصفي، على درجة الامتحان النهائي لـ 164 طالبة مسجلة في الفصل الدراسي الأول في مادة أساسيات تقنية المعلومات. في هذه الدراسة تم استخدام الإحصاءات الوصفية وتحليل ارتباط بيرسون، بالإضافة إلى خوارزمية التجميع K-Means ونماذج الانحدار (شجرة القرار، والغابة العشوائية، والتعزيز التدريجي).

أظهرت نتائج ارتباط بيرسون أن أقوى ارتباط كان بين الاختبار النصفي والاختبار النهائي (r=0.667)، يليه الاختبارات القصيرة (r=0.612)، بينما كان للغياب ارتباط سلبي ضعيف (r=-0.359)، وكشف تحليل التجميع باستخدام خوارزمية K-Means عن تباين في أداء الطلاب، وتم تصنيفهم إلى ثلاث مجموعات: أداء عال، ومتوسط، ومنخفض. في حين حققت خوارزمية الغابة العشوائية أدق تنبؤ (R^2=0.62) وأقل نسبة خطأ.

أكد تحليل SHAP أن الامتحان النصفي كان العامل الأكثر تأثيرا على درجة الامتحان النهائي، يليه الاختبارات الأسبوعية، بينما كان للغياب تأثير غير مباشر على درجة الامتحان النهائي. في الختام، تؤكد هذه الدراسة أهمية التقييم المستمر للتنبؤ بالطلاب المعرضين للخطر، وتقديم الدعم المناسب في الوقت المناسب لتحسين نتائج التعليم بشكل عام.

1. Introduction

Tripoli College of Science and Technology, one of the higher educational institutions in Libya, seeks to elevate the quality of the educational process through monitoring students' performance. In higher education, high student performance in fundamental courses is considered evidence of the quality of education and students' ambition to study, while low performance indicates difficulties in educational methods or in providing the required support to students (Romero & Ventura, 2020). Moreover, students' performance is considered one of the measurement criteria of education quality in higher education institutions and contributes to the progression of curricula, teaching methods, and framing proper educational policies (Baker & Inventado, 2014).

The digital transformation in higher education institutions has led to the availability of large amounts of student digital data, including grades, attendance records, and other data, allowing the use of prediction models to predict and analyze academic performance. For example, Hasan et al. (2025) showed that combining decision trees and K-Means clustering provides accuracy in predicting academic performance in addition to distinguishing learning groups into active learners and at-risk

students. Similarly, Jin et al. (2025) used clustering techniques in vocational education to detect variation between student groups, which supported the implementation of customized learning strategies.

The integration of more than one predictive model, such as classification and clustering, helps to improve prediction accuracy. Shovon and Haque (2025) used Decision Trees and the K-Means algorithm to analyze the results of quizzes, assignments, and laboratories. Decision Trees achieved accuracy in predicting students' cumulative GPA, and K-Means helped identify a group of students at risk and provided early information to instructors before final examinations. The results confirmed that continuous assessment results (quizzes, assignments, and laboratories) are among the most prominent indicators of academic success and have a strong impact on final results and cumulative GPA.

With increasing development, predictive models have become more accurate in prediction, and it has become important that prediction results be interpretable alongside accuracy. In a study conducted by Guevara-Reyes et al. (2025), interpretability methods such as SHAP were integrated with predictive models, which contributed to understanding how social, economic, and institutional factors affect academic performance and helped support appropriate educational decisions.

Although Tripoli College of Science and Technology is concerned with the quality of education and monitoring students' performance, failure and withdrawal rates among first-semester female students in the Computer Technology Department are high. Student performance data such as attendance, quizzes, and

midterm exam scores, which are routinely collected, are not well utilized to predict final exam results, with limited use of advanced analytical methods for early prediction and intervention. There has become an urgent need to use advanced predictive techniques and to integrate correlation analysis, clustering, predictive modelling, and interpretable artificial intelligence techniques to identify factors affecting academic performance, identify female students at risk of failure and withdrawal, and intervene in a timely manner. In addition, despite many studies investigating machine learning algorithms in predicting students' performance, there are insufficient studies within Libyan higher education settings, and few studies examined the influence of quizzes, attendance, and midterm exams on final exam performance within foundational college courses. This study focuses on female students because the selected department had only female enrollment during the study period. Focusing on female students allows for a more consistent analysis of academic performance and reduces variability that could be introduced by mixed-gender data.

Based on the above research gaps, the research objectives are:

1. Studying the relationship between attendance, weekly quizzes, midterm exam scores, and final exam performance.
2. Classifying female students based on their academic performance using clustering techniques.
3. Comparing the results of predictive models for final exam scores.
4. Determining the importance of each variable using feature importance analysis and SHAP interpretation.

5. Identifying female students at risk of withdrawal and providing early support.

To achieve the above objectives, the following research questions must be answered:

1. What is the relationship between student attendance, weekly quizzes, midterm exam scores, and final exam performance?
2. Can students be classified into groups based on their results?
3. Which predictive model is the most accurate and has the lowest error in predicting final exam scores?
4. Which factor is the most influential on final exam results?

This study contributes to increasing the understanding of educational data mining techniques by integrating correlation analysis, clustering, predictive modelling, and interpretable artificial intelligence techniques within a single analytical framework. This study also demonstrates the possibility of benefiting from routinely collected data to identify female students at risk of failure and withdrawal.

The remainder of this paper is organized as follows. Section 2 presents previous research articles on clustering and student performance. Section 3 describes data pre-processing and the development of predictive and clustering models. Results and discussion are discussed in Section 4, while the conclusion and future works are outlined in Section 5.

2. Related Work

2.1 Machine Learning for Student Performance Prediction

Machine learning has become an effective tool for predicting student academic performance. In a study by Baker & Inventado (2014), the importance of educational data mining (EDM) in predicting students' performance and its role in supporting educational decision-making that leads to education enhancement was highlighted. Moreover, Romero & Ventura (2020) conducted a comprehensive literary review in EDM applications and confirmed that using hybrid methods and techniques to cover all educational data helps early prediction of students' performance, reduces student dropouts, and enhances student outcomes. The accuracy of prediction results was confirmed in Liu et al. (2022); they used a university education system to build an accurate predictive model to predict students' cumulative average, which led to providing early support to students and enhancing academic performance.

2.2 Clustering Techniques and Learning Behaviour Analysis

Clustering algorithms are widely used to group students based on similar characteristics. Breiman et al. (1984) provided fundamentals of decision tree classification and regression that facilitate complicated relational models among different academic performance variables such as test scores, homework, and learning behaviours. Breiman (2001) revealed the effectiveness of Random Forest in producing accurate predictions and extensive learning resistance, which makes it suitable for student performance prediction applications. Hasan et al. (2025), in Bangladesh, applied Decision Tree and K-Means to classify students into

learning groups (active, passive, at-risk), and the prediction results were high, reaching 0.99, confirming the effectiveness of this algorithm in early prediction of academic performance.

Shovon & Haque (2025) focused on first-year university students in science and engineering colleges, predicting their cumulative average using Decision Tree and K-Means, providing early warnings to teachers before the final exam, which helps reduce academic failure. Yağcı (2022) and Khairy et al. (2024) found that Random Forest, Decision Tree, and SVM algorithms provided the most accurate prediction of student outcomes and supported early warning systems in educational institutions.

2.3 Hybrid and Multi-Model Approaches

Several studies confirmed that combining multi-models can enhance prediction accuracy. Zou et al. (2025) combined multi-model fusion (K-Means, BIRCH, DBSCAN) with classification and clustering models to provide accurate prediction and enhance student digital management. Bhogan et al. (2025) presented a hybrid approach using an enhanced K-strange points clustering algorithm, Naïve Bayes classification, and multiple linear regression, confirming that combining multi-models reduces prediction errors and enhances early educational intervention. Guevara-Reyes et al. (2025) combined XGBoost and Random Forest with interpretable SHAP to help educational decision-making, allowing identification of the most influential factors on academic performance and providing accurate recommendations to decision-makers. (Romero & Ventura, 2020; Zou et al., 2025).

Overall, previous studies confirmed the importance of combining clustering and prediction techniques with interpretable SHAP in

providing accurate and interpretable results that boost educational decision-making.

3. Methodology

This research design is a quantitative correlational-predictive approach to explore the most influential factor of Absence, Weekly-Quizzes, and Midterm-Exam on Final-Exam. All analyses were conducted using Python and standard data analysis and machine learning libraries. The general architecture of the methodology is presented in Figure 1.

The dataset consists of 164 female students enrolled in the first semester in a fundamental information technology course in the Computer Science Department at Tripoli College of Science and Technology. The dataset included four variables: number of absences, total weekly quiz scores, midterm exam scores, and final exam scores. Continuous assessment results (quizzes, assignments, and laboratories) are among the most prominent indicators of academic success (Shovon & Haque, 2025).

To protect student privacy, all personal information that reveals students' identity was removed and replaced with random codes. A data quality check confirmed that there was no missing data; therefore, there was no need for data imputation or record removal. Continuous variables were scaled to ensure equivalence across features.

Descriptive statistics were calculated to summarize students' performance and attendance behavior; boxplots were used to present feature distributions; outliers were reserved and corrected only if they were data entry errors. Frequency distribution of scores was displayed through histograms. Pearson's correlation was computed to examine the relationship between research variables at a significance level of $\alpha = 0.05$ (Pearson, 1895)

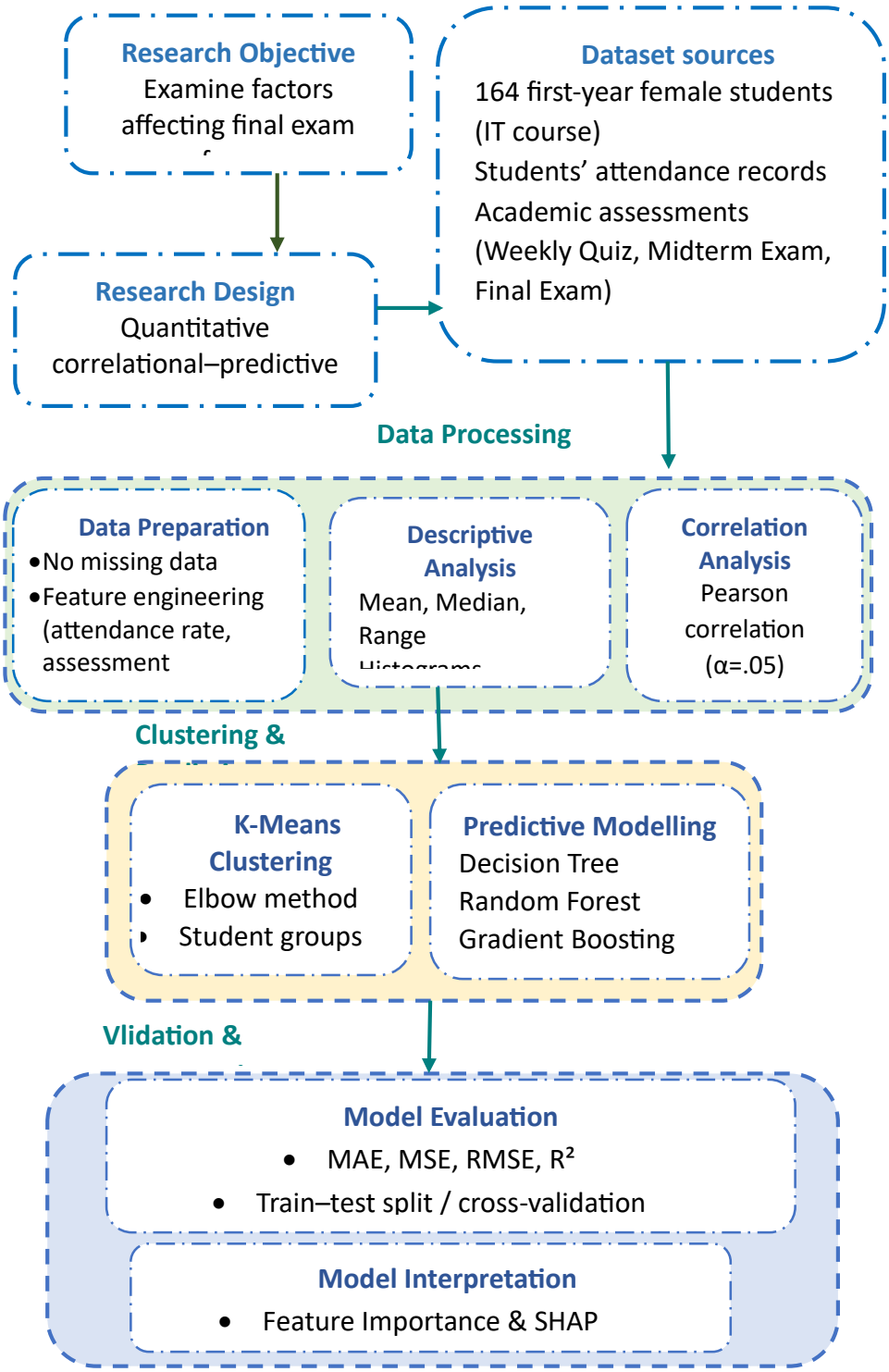
Research Objective
Examine factors affecting final exam
Dataset sources
164 first-year female students (IT course)
Students' attendance records
Academic assessments (Weekly Quiz, Midterm Exam, Final Exam)
Research Design
Quantitative correlational–predictive
Data Processing
Data Preparation
•No missing data
•Feature engineering (attendance rate, assessment
Descriptive Analysis
Mean, Median, Range
Correlation Analysis
Pearson correlation (α=.05)
Clustering &
K-Means Clustering
• Elbow method
• Student groups
Predictive Modelling
Decision Tree
Random Forest
Gradient Boosting
Vlidation &
Model Evaluation
• MAE, MSE, RMSE, R²
• Train–test split / cross-validation
Model Interpretation
• Feature Importance & SHAP

Figure 1-Workflow of the proposed methodology.

Furthermore, students were categorized into groups based on their results using K-Means clustering, and the Elbow method identified the optimal number of groups. The Elbow method indicated that three clusters provide a

good balance between within-cluster similarity and between-cluster differences, aligning with common grouping of student performance into high, average, and low achievement.

Three regression algorithms Decision Tree, Random Forest, and Gradient Boosting were applied to predict final exam scores. The dataset was divided into two sets: 80% for training and 20% for model performance evaluation and generalizability. Additionally, k-fold cross-validation was applied during model training, and the average performance metrics were reported to ensure stability and reduce the risk of overfitting. Model performance was evaluated using standard error metrics, including Mean Absolute Error (MAE) and Mean Squared Error (MSE). Feature importance and SHAP values were used to enhance predictive model explainability.

This study followed ethical standards of the institution. All student data were anonymized, and no personal information was used. The dataset was used only for research purposes, and results do not affect students' grades or progression. Ethical approval was obtained, or the study was considered exempt because it involved analysis of anonymized educational data.

4. Results and Discussions

4.1 Results

4.1.1 Descriptive Statistics

Descriptive statistics were calculated to summarize students' performance and attendance behavior. Table 1 presents the measures of all study variables.

Table 1: Descriptive Statistics of Study Variables

Variable	Mean	Median	Min	Max	25th Percentile	75th Percentile
ABSENCES_NO	4.00	5.00	0.00	10.0	3.00	5.00
WEEKLY-QUIZZES	13.41	16.00	0.00	20.0	8.75	20.00
MIDTERM-EXAM	11.47	11.50	10.0	20.0	7.00	16.00
FINAL_EXAM	37.94	39.50	15.0	50.0	31.00	45.25

Descriptive statistics show that students had regular attendance with an average of 4 absences, while some had more absences. Weekly quizzes exposed medium variation in grades from 0 to 20, reflecting differences in students' engagement and continuous performance. The midterm exam showed low variation in grades from 10 to 20, while final exam results showed greater variation.

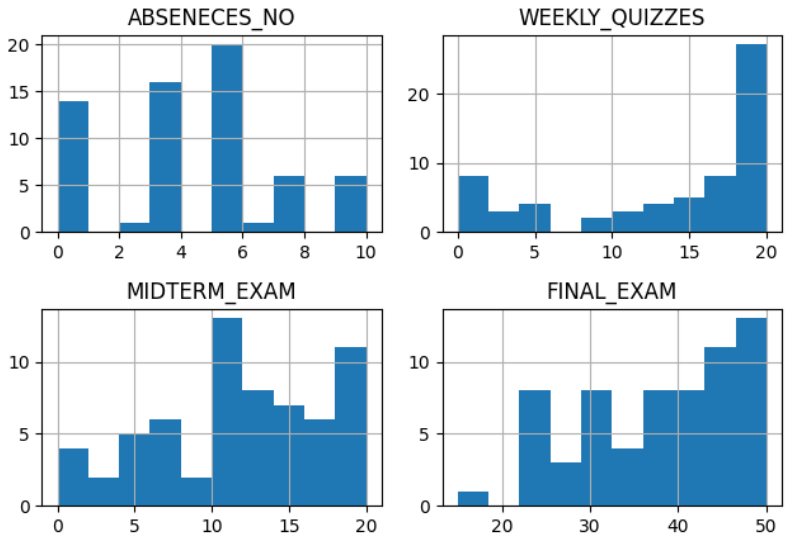

Figure 2 – Histograms of Student Performance Variables

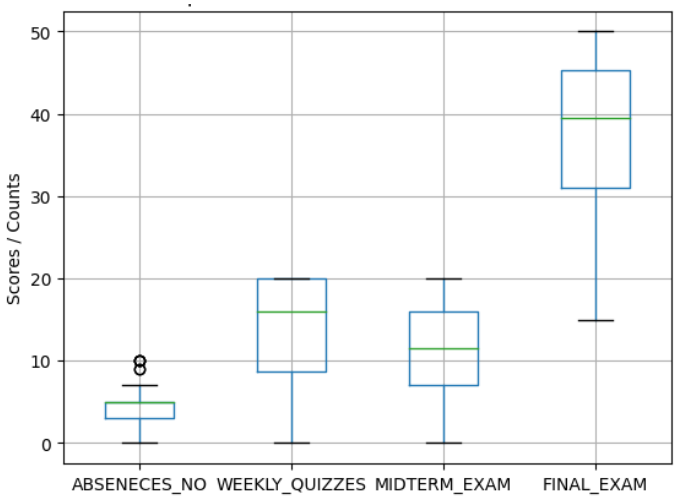

Figure 3 – Boxplots of Student Performance Variables

Frequency distribution of scores was displayed through histograms in Figure 2. Boxplots presented in Figure 3 confirmed a normal distribution of values with some outliers.

4.1.2 Correlation Analysis

The relationship between student attendance, weekly quizzes, midterm exam scores, and final exam performance was examined using Pearson's correlation analysis. The correlation matrix is presented in Table 2.

Table 2: Pearson Correlation Matrix

Variable	ABSENCES_NO	WEEKLY-QUIZZES	MIDTERM-EXAM	FINAL_EXAM
ABSENCES_NO	1.000	–0.731	–0.252	–0.359
WEEKLY-QUIZZES	–0.731	1.000	0.372	0.612
MIDTERM-EXAM	–0.252	0.372	1.000	0.667
FINAL_EXAM	–0.359	0.612	0.667	1.000

As shown in Table 2, there is a strong negative connection between weekly quiz scores and absences (r = –0.73), indicating that absences negatively affect continuous assessment, as students with more absences performed worse in quizzes. Absence had a low negative effect on both midterm (r = –0.25) and final examination scores (r = –0.36). Weekly quizzes had a moderate effect on midterm exam scores (r = 0.37) while having a strong positive influence on final exam scores (r = 0.61).

To assess the significance of relationships between student performance metrics Pearson correlation tests were performed as shown in Table 3.

Table 3: Pearson Correlation Coefficients with p-values and 95% Confidence Intervals (n = 164)

Variable Pair	R	p-value	95% CI
ABSENCES_NO – WEEKLY_QUIZZES	–0.731	<0.0001	–0.79 – –0.66
ABSENCES_NO – MIDTERM_EXAM	–0.252	0.0446	–0.49 – –0.01
ABSENCES_NO – FINAL_EXAM	–0.359	0.0035	–0.57 – –0.13
WEEKLY_QUIZZES – MIDTERM_EXAM	0.372	0.0025	0.14 – 0.57
WEEKLY_QUIZZES – FINAL_EXAM	0.612	<0.0001	0.44 – 0.75
MIDTERM_EXAM – FINAL_EXAM	0.667	<0.0001	0.50 – 0.77

The strongest positive correlation was between midterm and final examination scores (r = 0.67), revealing that the midterm exam is the most influential factor on final exam outcomes. A heatmap presenting the correlations is shown in Figure 4.

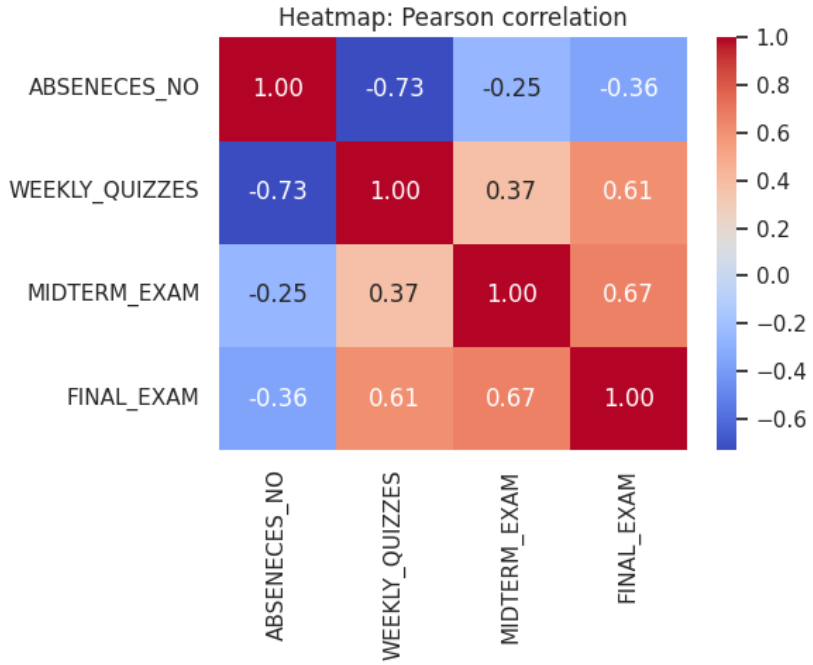

Figure 4 – Heatmap of Pearson Correlations

4.1.3 Clustering Analysis

A K-Means clustering algorithm was conducted to identify the best number of clusters. The Elbow method was applied based on midterm and final exam scores (Figure 5a).

Figure 5b shows the clustering results, which revealed variation in students' performance, classified into three groups: high, average, and low performance. This division shows differences in student achievement levels that help teachers provide the necessary support to students at risk.

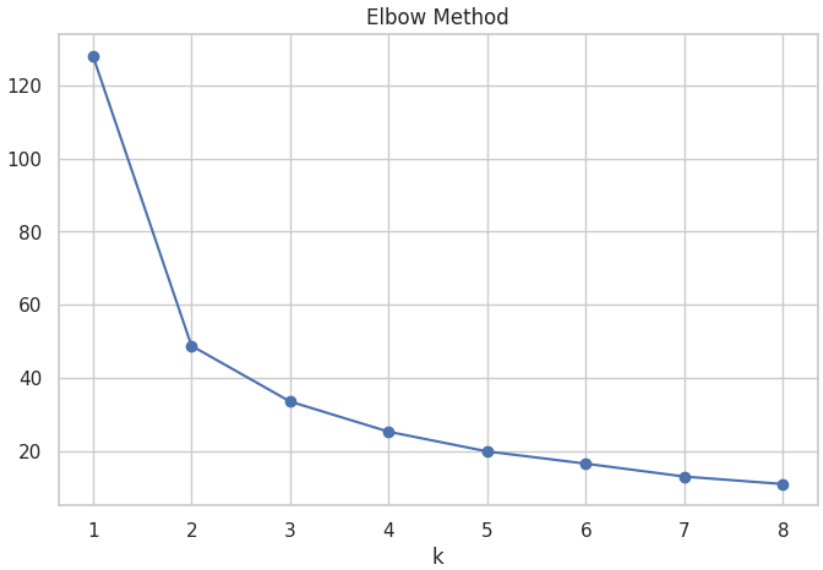

Figure 5a - Elbow Method

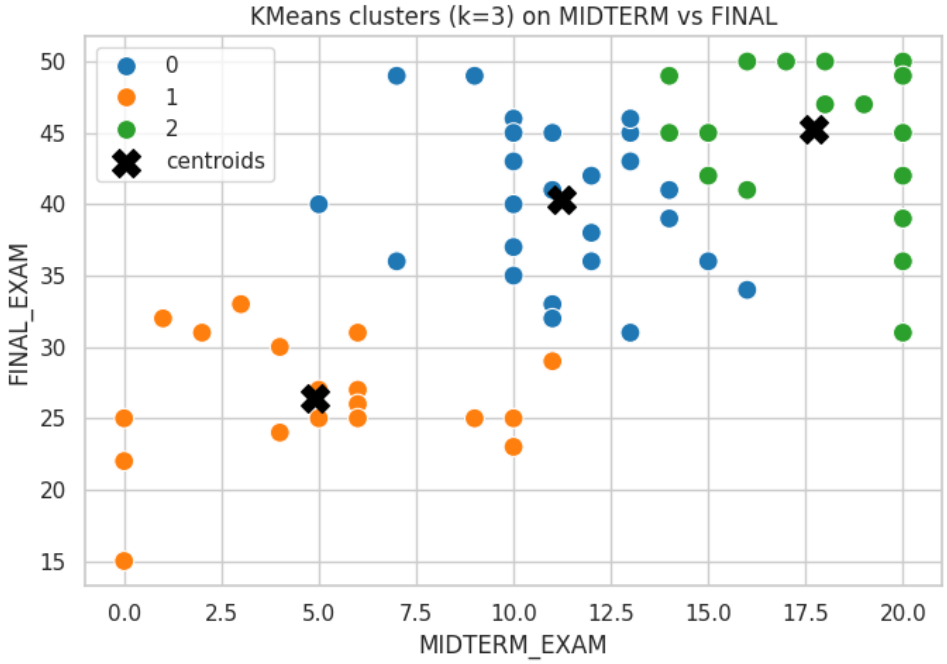

Figure 5b – Cluster Scatter Plot

Figure 5 - Clustering Analysis of Student Performance

The Elbow method (Figure 5a) shows the within-cluster sum of squares (WCSS) for k = 1–8, confirming that the ideal number is k = 3. Figure 5b shows a scatter plot where three groups based on midterm and final exam scores are clearly visible.

4.1.4 Predictive Modelling

Three regression algorithms Decision Tree, Random Forest, and Gradient Boosting were used to predict final exam scores. Each student's data, including the number of lecture absences, midterm exam grade, and total weekly quiz grades, were entered to predict the final exam result. Table 4 presents model performance metrics evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and cross-validation metrics.

Table 4: Predictive Model Performance Metrics

Model	MAE	MSE	RMSE	R^2	CV R^2 (Mean)	CV MAE (Mean)
Decision Tree	6.00	59.20	7.69	0.43	–0.13	6.90
Random Forest	5.17	39.74	6.30	0.62	0.21	5.62
Gradient Boosting	6.29	60.90	7.80	0.42	–0.12	6.54

Negative cross-validation scores in Decision Tree and Gradient Boosting suggest that these models were sensitive to data splitting and possibly overfitted the training data. Random Forest, however, showed more stable cross-validation results, confirming it as the most suitable predictive model for this dataset.

As revealed in Table 4, Random Forest achieved the most accurate prediction with ($R^2 = 0.62$) and the least prediction errors. Positive cross-validation results indicate better generalizability to unseen data. Decision Tree and Gradient Boosting showed lower predictive accuracy and negative cross-validation, indicating possible overfitting. Figure 6 presents a visualization comparing model performance metrics, showing Random Forest as the leading model. MAE represents the amount of error in scores when predicting, i.e., the difference between actual and predicted scores. R^2 represents the proportion of variance explained in predicting final exam scores.

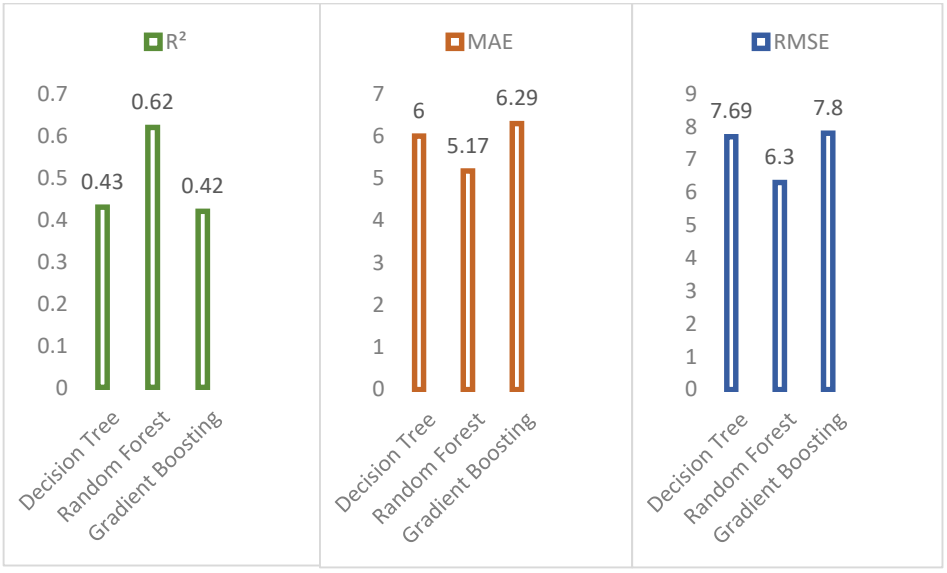

Figure 6 – Comparison of Model Performance Metrics (R^2, MAE, RMSE)

4.1.5 Feature Importance and SHAP Analysis

Feature importance was derived from the Random Forest model to assess the impact of each feature midterm exam, weekly quizzes,

and number of absences on final exam results. Table 5 presents the results.

Table 5: Feature Importance Based on Random Forest Model

Feature	Importance
MIDTERM-EXAM	0.539
WEEKLY-QUIZZES	0.322
ABSENCES_NO	0.139

As shown in Table 4, the midterm exam is the most correlated and predictive of final exam results (0.539), highlighting the importance of early assessments. Weekly quizzes had moderate importance (0.322), while absence had a minor effect (0.139).

Figure 7a displays ranked feature importance scores. Figure 7b shows the SHAP summary plot, providing an overview of each factor's effect. Figure 7c presents the SHAP dependence plot for MIDTERM-EXAM, showing higher midterm scores strongly predict higher final exam scores.

Overall, midterm exams are a strong indicator of predicting final exam outcomes. Continuous assessment also had a significant role and was the second strongest indicator.

4.2 Discussion

This study explored the relationship between absence, weekly quizzes, midterm exam, and final exam using Pearson correlation analysis; grouped students based on their performance using K-Means clustering; and identified the most predictive factor on final exam among absence, weekly quizzes, and midterm exam. The findings explain how these factors affect overall student

performance (Romero & Ventura, 2020; Baker & Inventado, 2014).

4.2.1 Interpretation of Descriptive Statistics

Descriptive statistics showed that most students had regular attendance with a mean of four absences, although some students had more absences, which affected their achievement in continuous assessment. The study confirms that regular attendance is the basis of academic success, especially in regular educational environments (Credé, Roch, & Kieszczynka, 2010; Gottfried, 2014).

Weekly quizzes and midterm examination scores presented medium variation, while final exam showed larger differences, showing that the gap between student academic performance increases with the end of semester, which is evidence of the role of early and continuous assessment on shaping final academic outcomes (Shovon & Haque, 2025; Roediger & Karpicke, 2006).

4.2.2 Correlation Analysis of Variables

Correlation between absence and weekly quizzes was the strongest negative correlation, indicating that students who missed classes also missed quizzes, affecting understanding and performance. This is consistent with studies showing absence reduces the opportunity to receive educational information (Gottfried, 2014). Absence had a low negative correlation on midterm and final exam, with an indirect effect through reduced quizzes (Shovon & Haque, 2025).

Weekly quizzes had a strong positive correlation with midterm and final exam, especially final exam. Frequent continuous assessment

helps in gaining knowledge, known as the "testing effect" (Roediger & Karpicke, 2006). Midterm exam had the strongest correlation with final exam, meaning final exam outcomes can be expected through midterm exam, aligning with other studies that confirm early assessments can predict final exam results and identify at-risk students (Credé, Roch, & Kieszczynka, 2010; Hasan et al., 2025).

4.2.3 Clustering of Student Performance Profiles

Three clusters were identified using midterm and final exam scores: high, average, and low performance. This presents variation in students' performances that can be recognized easily. Clustering divides students based on performance to understand learning patterns and form adaptable learning strategies (Romero & Ventura, 2020; Jin, 2025).

Recognizing these clusters helps teachers offer the low-performance group supplement classes and provide the high-performance group opportunities to advance knowledge, supporting educational decision making (Hasan et al., 2025; Shovon & Haque, 2025).

4.2.4 Predictive Modelling and Model Performance

Among three predictive models, Random Forest achieved the most accurate prediction ($R^2 = 0.62$) and the least prediction errors, with positive cross-validation. Decision Tree and Gradient Boosting gave lower predictive accuracy and negative cross-validation, indicating possible overfitting. This is consistent with studies showing Random Forest is suitable for educational prediction due to capturing complex, non-linear relationships (Breiman, 2001; Yağcı, 2022; Khairy et al., 2024).

Decision Tree is easier to understand but has low prediction accuracy. Random Forest is more accurate. Choosing a model requires balancing accuracy and interpretability; Random Forest may be more suitable when prediction accuracy is prioritized (Bhogan et al., 2025; Guevara-Reyes et al., 2025).

4.2.5 Feature Importance and Suggestions for Early Intervention

Feature importance from Random Forest confirmed that midterm exam is the most correlative and predictive factor, underscoring the importance of early assessments (Hasan et al., 2025; Shovon & Haque, 2025). Weekly quizzes had a moderate effect. Absence had an indirect effect on final exam performance.

These results support the importance of early monitoring of student performance and providing suitable interventions to improve educational outcomes (Arnold & Pistilli, 2012; Guevara-Reyes et al., 2025).

4.2.6 Summary of Key Significances

The study presents the connection between attendance, continuous assessment, and summative performance. Midterm exam is the strongest influence on final exam, while weekly quizzes play a big role in learning and retaining. Attendance has a low effect but is considered important for success. The results confirm the importance of continuous monitoring and educational practices based on academic data (Romero & Ventura, 2020; Zou et al., 2025).

5. Conclusion

This study examined attendance, weekly quizzes, midterm exam, and final exam scores, identifying the most predictive factor. Using descriptive statistics, Pearson correlation, K-Means clustering, and regression models (Decision Tree, Random Forest, Gradient Boosting), the study provided evidence that midterm exams are the most predictive, while attendance and weekly quizzes have supportive roles.

Descriptive statistics showed regular attendance, variation in continuous assessments, and larger differences in final exams. Correlation analysis indicated that absence negatively affects quizzes, midterm exams strongly predict final exams, and continuous assessment is important.

Clustering revealed three groups high, average, and low performance. Identifying at-risk students supports early intervention. Monitoring midterm and weekly quizzes helps instructors provide timely support. Future work could include behavioral and demographic variables and larger datasets to study long-term intervention impact.

1. References

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 267–270. https://doi.org/10.1145/2330601.2330666

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning Analytics* (pp. 61–75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4

Bhogan, S., Sawant, K., Naik, P., Shaikh, R., Diukar, O., & Dessai, S. (2025). Predicting student performance based on clustering and classification. *IOSR Journal of Computer Engineering (IOSR-JCE) AITD, Goa University, India*. https://doi.org/10.9790/0661-1903054952

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315139470

Credé, M., Roch, S. G., & Kieszczynka, U. M. (2010). Class attendance in college: A meta-analytic review. *Review of Educational Research, 80*(2), 272–295. https://doi.org/10.3102/0034654310362998

Gottfried, M. A. (2014). Chronic absenteeism and its effects on students' academic and socioemotional outcomes. *Journal of Education for Students Placed at Risk, 19*(2), 53–75. https://doi.org/10.1080/10824669.2014.962696

Guevara-Reyes, R., Ortiz-Garcés, I., Andrade, R., Cox-Riquetti, F., & Villegas-Ch, W. (2025). Machine learning models for academic performance prediction: Interpretability and application in educational decision-making. *Frontiers in Education, 10*, 1632315. https://doi.org/10.3389/feduc.2025.1632315

Hasan, M. M., Islam, M. N., Nirjon, M. I. H., Uddin, M. S., Mamun, M., Munna, Z. A., & Rumman, A. M. (2025). Predicting student performance and identifying learning behaviors using decision trees and K-means clustering. *International Journal of Evaluation and Research in Education (IJERE), 14*(5), 3872–3881. https://doi.org/10.11591/ijere.v14i5.33815

Islam Shovon, M. H., & Haque, M. (2025). An approach of improving student's academic performance by using K-means clustering algorithm and decision tree. *Rajshahi University of Engineering & Technology, Bangladesh*.

Jin, J. (2025). Student behavior patterns in vocational education big data based on clustering algorithm. *Discover Artificial Intelligence, 5*, 197. https://doi.org/10.1007/s44163-025-00433-3

Khairy, D., Alharbi, N., Amasha, M. A., et al. (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

Liu, C., Wang, H., & Yuan, Z. (2022). A method for predicting the academic performances of college students based on education system data. *Mathematics, 10*(20), 3737. https://doi.org/10.3390/math10203737

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London, 58*(347–352), 240–242. https://doi.org/10.1098/rspl.1895.0041

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355. https://doi.org/10.1002/widm.1355

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*, Article 11. https://doi.org/10.1186/s40561-022-00192-z

Zou, W., Zhong, W., Du, J., & Yuan, L. (2025). Prediction of student academic performance utilizing a multi-model fusion approach in the realm of machine learning. *Applied Sciences, 15*, 3550. https://doi.org/10.3390/app15073550