University of Hertfordshire UH

School of Physics, Engineering and Computer Science

MSc Data Science Project
7PAM2002

Department of Physics, Astronomy and Mathematics

Data Science FINAL PROJECT REPORT

Project Title:

Student Performace factor Analysis

Student Name and SRN:

Muhammad Faheem Mukhtar

22104671

Supervisor: Pedro Carrilho

Date Submitted: 28/08/2025

Word Count: Four Thousand seven hundred and Seventy

DECLARATION STATEMENT

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science **in Data Science** at the University of Hertfordshire. I have read the detailed guidance to students on academic integrity, misconduct and plagiarism information at Assessment Offences and Academic Misconduct and understand the University process of dealing with suspected cases of academic misconduct and the possible penalties, which could include failing the project or course. I certify that the work submitted is my own and that any material derived or quoted from published or unpublished work of other persons has been duly acknowledged. (Ref. UPR AS/C/6.1, section 7 and UPR AS/C/5, section 3.6). I did not use human participants in my MSc Project.

I hereby give permission for the report to be made available on module websites provided the source is acknowledged.

Student Name printed: Muhammad Faheem Mukhtar

Student Name signature: Muhammad Faheem Mukhtar

Student SRN number: 22104671

GitHub ID: https://github.com/Faheemmian143/PFR-Student-Performance-factor/upload/main

UNIVERSITY OF HERTFORDSHIRE

SCHOOL OF PHYSICS, ENGINEERING AND COMPUTER SCIENCE

Abstract

The prediction of student performance has become a critical application of data science in education, enabling institutions to identify at-risk students and improve learning outcomes. This study investigates the predictive capacity of machine learning methods using the *Student Performance Factors* dataset, which comprises 6,607 student records and 20 attributes spanning academic, socio-economic, behavioral, and contextual variables. The research aims to evaluate the effectiveness of regression, classification, and clustering approaches in forecasting student achievement and to identify the most significant predictors of exam performance.

Exploratory data analysis revealed that academic engagement factors, particularly hours studied, attendance, and previous scores, were strongly correlated with exam outcomes. Regression analysis demonstrated that ensemble methods were most effective, with the Random Forest Regressor achieving an R^2 of 0.869 and RMSE of 1.36, outperforming linear regression and decision tree regression. Classification models categorized students into high- and low-performing groups using the median exam score as a threshold. Logistic Regression achieved superior performance with an accuracy of 92.1%, compared to 87.4% for K-Nearest Neighbors, and demonstrated balanced precision and recall across both classes.

K-Means and Hierarchical Clustering clustering analysis scored a low silhouette value of 0.1292 but identified interpretable student clusters based on study efficiency variation, academic support, and parental involvement. Feature importance analysis validated student absenteeism, study efficiency, and hours studied as the most significant predictors, while socio-economic and behavioral variables yielded secondary information.

These results conclude that it is possible to predict student performance effectively using regression and classification schemes such that ensemble regression and logistic classification are the most consistent. These results indicate the predominance of academic engagement while attesting to a supplementary role for contextual variables. These conclusions have practical applications for teachers and policymakers wishing to create specific interventions in raising student success.

Chapter 1: Introduction

Predicting student attainment is nowadays a key focus in educational data science so that teachers, policymakers, and parents might plan interventions for improved learning achievements. Predictive analytics is increasingly a standard in schools for forecasting at-risk learners, superior pedagogic responses, as well as student wellbeing. As machine learning (ML) continues its further expansion, it is now possible to harness large-scale datasets including both scholarly as well as socio-economic proxies for greater prediction accuracy on attainment vis-à-vis conventional statistical approaches .

The in-class dataset for this research, StudentPerformanceFactors.csv, contains 6,607 records with 20 features representing a wide range of academic and personal features. These involve hours studied, attendance, scores in the past, hours slept, and number of tutoring sessions, as well as contextual features such as family income level, tendency towards parental involvement, quality of teachers, level of peer influence, and access to resources. The continuous dependent variable for regression problems is Exam_Score, whereas for classification tasks, exam scores were coded in a binary target so that students were divided into high-achievers versus low-achievers. Both accurate performance prediction as well as practical student classification are thereby supported for their dual problems, consistent with past research which has used regression for predicting grades while past research has also used classification for early warning signals

An initial statistical exploration revealed meaningful variation across predictors. Students studied between **1 and 44 hours per week** (mean = 19.97), with attendance ranging from **60% to 100%** (mean = 79.97). The average exam score was **67.2** with a minimum of 55 and a maximum of 101, indicating relatively high baseline achievement but with notable variation. Missing values were detected in *Teacher Quality (78 cases)*, *Parental Education Level (90 cases)*, and *Distance from Home (67 cases)*, which were handled in preprocessing. To enhance model interpretability and accuracy, **feature engineering** was applied, creating new variables such as *Study Efficiency*, *Academic Support*, *Resource Access Score*, and *Wellness Score*. Similar strategies of deriving composite indicators have been shown to improve prediction quality in academic datasets

The rationale for this research is threefold. First, by identifying the most influential features, stakeholders can allocate resources more effectively and support students at risk of underperformance Second, comparing regression and classification models provides insights into whether continuous score prediction or categorical performance grouping offers greater educational utility. Third, clustering techniques are employed to uncover natural groupings among students, supporting the design of targeted pedagogical interventions, as recommended by recent studies in educational data mining .

Based on this context, the research addresses the following questions:

- **RQ1:** How effectively can regression models predict student exam scores using demographic, behavioral, and contextual factors?
- **RQ2:** Do classification models outperform regression models in identifying high-performing versus low-performing students?
- **RQ3:** Which factors—academic, socio-economic, or behavioral—contribute most significantly to predicting student performance?

This chapter establishes the significance of student performance prediction, introduces the dataset, and sets out the objectives of the study. The subsequent chapters will provide a comprehensive review of existing literature (Chapter 2), outline the methodological framework (Chapter 3), present results and discussion (Chapter 4), and conclude with implications and future research directions (Chapter 5).

Chapter 2: Literature Review

The field of educational data mining has produced a wide body of work exploring student performance prediction. Most studies combine academic attributes (such as grades and attendance) with socio-economic and behavioral factors (such as parental education and motivation). This chapter reviews three representative studies that directly align with the objectives of this dissertation, focusing on the use of regression, classification, and feature-driven approaches.

Study 1: Dekker et al. (2009)

Dekker and colleagues conducted one of the earliest large-scale studies on predicting student dropout and performance in higher education. They used demographic, academic history, and motivational features with decision trees and logistic regression to predict dropout risk. Their findings highlighted that prior academic achievement and motivation were the most predictive features. However, the dataset size (n = ~1,500) was relatively small, and the study relied heavily on categorical academic history without incorporating behavioral features such as study efficiency or wellness indicators. This limits its applicability to more diverse educational contexts.

Study 2: Al-Barrak & Al-Razgan (2016)

Al-Barrak and Al-Razgan applied decision trees to predict final GPA outcomes in Saudi higher education students. Their dataset (~4,000 records) included variables such as study habits, parental background, and attendance. The models achieved moderate accuracy, with classification models proving more effective than regression in grouping students by GPA categories. However, their approach lacked advanced feature engineering, which may have constrained predictive performance.

Study 3: Kumar & Singh (2021)

Kumar and Singh used machine learning algorithms including logistic regression, random forests, and support vector machines to predict student performance in Indian schools. Their dataset combined academic features (attendance, marks, assignments) with socio-economic indicators (parental education, income, internet access). Random forests outperformed other models, with accuracy above 90%, showing the strength of ensemble methods. Their study aligns closely with the present work but did not extend into clustering analysis or the construction of derived composite features such as wellness or study efficiency.

These studies collectively demonstrate the potential of machine learning in educational contexts, while also revealing common limitations such as small sample sizes, limited feature engineering, and a focus on classification accuracy over holistic analysis. The present dissertation extends this body of work by integrating **feature engineering (e.g., Study Efficiency, Academic Support, Wellness Score), regression and classification comparisons, and clustering-based group discovery**, all applied to a larger dataset of 6,607 students.

Table 2.1: Comparison of Related Studies and Present Work

Study & Year	Dataset Size	Features Used	Methods Applied	Key Findings	Gaps/Limitations	Alignment w Present Wo
Dekker et al. (2009)	~1,500	Demographics, academic history, motivation	Decision Trees, Logistic Regression	Prior achievement and motivation best predictors	Small dataset, limited behavioral features	Similar regress focus, but prese work adds engineered feat & larger datase
Al-Barrak & Al-Razgan (2016)	~4,000	Study habits, parental background, attendance	Decision Trees (Classification)	Classification more effective than regression	No advanced feature engineering	Present work a compares regression vs classification w richer feature s
Kumar & Singh (2021)	~2,000	Academic, socio-economic, internet access	Logistic Regression, Random Forest, SVM	Random Forest achieved >90 % accuracy	No clustering, no composite features	Present work extends with clustering + engineered feat for holistic ana

Present Study (2025)	**6,607**	Academic (hours studied, attendance, previous scores), socio-economic (income, parental involvement), behavioral (sleep, activity), + engineered features	Regression (LR, DT, RF), Classification (LR, KNN), Clustering (K-Means, Hierarchical)	Random Forest regression R^2 = 0.8688; Logistic Regression accuracy = 92.1%	Moderate clustering performance (silhouette = 0.1292), limited external validation	Integrates regression, classification, clustering + engineered feat on larger datas

Chapter 3: Methodology

This chapter presents the methodological framework used to analyze student performance, structured around data acquisition, exploration, preprocessing, modeling, and evaluation. The approach follows a step-by-step pipeline designed to ensure the reliability and validity of the results.

3.1 Dataset Description

The dataset used in this research is titled *Student Performance Factors* and was obtained from an openly available repository on Kaggle: Student Performance Factors Dataset. It contains **6,607 student records** and **20 attributes**, covering academic, socio-economic, behavioral, and contextual factors. The dependent variable for regression is the **Exam_Score**, while for classification tasks, the exam score was transformed into a binary variable using the median threshold (67), resulting in two categories: high-performing students (56.4%) and low-performing students (43.6%).

The dataset is diverse in scope. Academic attributes include *Hours_Studied* (1–44, mean = 19.98), *Attendance* (60–100%, mean = 79.98), *Previous_Scores* (50–100, mean = 75.07), and *Tutoring_Sessions* (0–8). Socio-economic features include *Parental Involvement*, *Family Income*, and *Parental Education Level*. Behavioral features such as *Sleep Hours* (4–10, mean = 7.03) and *Physical Activity* (0–6) provide insight into lifestyle influences on learning. Contextual attributes cover *School Type*, *Distance from Home*, *Internet Access*, and *Teacher Quality*. The dependent variable *Exam_Score* has an average of 67.24, with values ranging between 55 and 101.

Missing values were identified in *Teacher Quality* (78), *Parental Education Level* (90), and *Distance from Home* (67), while no duplicate rows were present. These missing

values were handled during the preprocessing stage.

3.2 Justification for Dataset Selection

Application of this dataset is justified on multiple fronts. First, the dataset provides a large sample size of 6,607 records allowing for robust machine learning analysis as well as generalizability. Most existing research in student performance prediction is grounded on smaller sample size datasets, often less than 2,000 records, which may undermine statistical power. Second, the dataset provides a combination of academic, socio-economic, and behavioural variables for a multi-dimensional study into factors influencing student outcomes. In comparison with datasets grounded on grades or attendance only, this dataset includes variables such as parental involvement, motivation, wellness, and access to resources, which are becoming increasingly critical in contemporary educational research. Third, the dataset includes both categorical and continuous variables suitable for a wide combination of techniques such as regression, classification, clustering, and feature construction. Such diversity in attributes allows for composite indicators such as Study Efficiency and Wellness Score which can enhance predictive performance. Finally, the dataset is ethically available, anonymized, and consistent with open data principles and thus suitable for academic research in the University of Hertfordshire (UH) ethical framework and GDPR recommendations.

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) served to acquire an appreciation for structure in the dataset and some salient trends. Descriptive statistics revealed that students studied 1-44 hours a week with a mean of 19.98 hours, whereas attendance ranged 60-100% with a mean of 79.98. Exam scores were moderately concentrated around 65-70 for a majority of students with an overall distribution having a mean of 67.24. Sleep hours were 7.03 on average while physical activity levels were concentrated around 2-4 times per week. Correlation analysis revealed strong positive correlations between study hours, attendance, and exam scores in validating their roles as predictors. Categorical features such as parental involvement and teacher quality revealed significant median exam score distinctions implying academic support infrastructures play key roles in student achievement. Overall, these results served as a foundation for subsequent feature engineering and model construction.

3.4 Feature Engineering

New features were engineered based on existing attributes in order to enhance predictive ability. Study Efficiency was constructed as exam score divided by hours studied, reflecting productivity per study hour. Academic Support was constructed as a composite variable consisting of parental involvement and student-teacher quality. Resource Access Score compiled access to resources and internet access, whereas Wellness Score combined hours of sleep and physical activity in order to reflect wellness. These engineered features infused fresh dimensions upon the dataset as a means of enhancing model capacity for reflecting latent forces governing student performance.

3.5 Data Preprocessing

Data pre-processing made sure the dataset was ready for machine learning. Categorical variables were converted into numeric form by label-encoding. Missing values were filled using strategies based on attributes such as mode or mean. Feature selection by SelectKBest with f-regression ranked the best 10 exam performance predictors such as study hours, presence, past scores, number of tutoring sessions, and study effectiveness. Standardization for numeric attributes was carried out by z-score normalization for making attributes comparable on a common scale. The dataset was divided into training and test parts at an 80–20 ratio while stratification was used for preserving class balance for classification problems. This pre-processing pipeline assured consistency and helped lower model bias risk.

3.6 Machine Learning Models

Three categories of models were implemented. For regression, Linear Regression, Decision Tree Regressor, and Random Forest Regressor were employed to predict continuous exam scores, with model performance assessed using mean squared error (MSE), root mean squared error (RMSE), and the coefficient of determination (R^2). For classification, Logistic Regression and K-Nearest Neighbors (KNN) were applied to predict high versus low student performance, with evaluation metrics including accuracy, precision, recall, and F1-score. For unsupervised learning, K-Means clustering (k=3) and Hierarchical Clustering were implemented to identify natural groupings among students, with silhouette score used as the primary evaluation metric. To aid interpretability, dimensionality reduction techniques such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) were used to visualize clustering outputs.

3.7 Ethical Considerations

The use of educational data requires strict adherence to ethical standards. Although the dataset was publicly available and anonymized, precautions were taken to ensure compliance with University of Hertfordshire (UH) ethical requirements and the General Data Protection Regulation (GDPR). Specifically, all analyses avoided any attempt at re-identification, and results were reported only in aggregate. UH ethics principles of integrity, transparency, and minimization of harm guided the analysis. GDPR principles of fairness, accountability, and data minimization were respected by restricting use to academic purposes only. No sensitive personal information was present in the dataset, and the analysis focused solely on improving educational insights rather than individual-level predictions. This ensures that the research adheres to both institutional and legal standards.

3.8 Model Performance Summary

The results of the modeling phase were consolidated to compare performance across approaches. In regression, Random Forest Regressor achieved the best performance with

an R^2 of 0.8688 and an RMSE of 1.36, substantially outperforming linear regression and decision tree regression. In classification, Logistic Regression achieved an accuracy of 92.1%, outperforming KNN at 87.4%. Clustering analysis produced a silhouette score of 0.1292, indicating weak but interpretable grouping structure. Feature importance analysis confirmed that attendance, study efficiency, and hours studied were the most influential predictors of exam outcomes. These findings provided the basis for the detailed interpretation and discussion presented in the next chapter.

Chapter 4: Results and Discussion

This chapter presents the results of the analysis conducted on the *Student Performance Factors* dataset, structured into exploratory data analysis, regression modeling, classification modeling, clustering analysis, and feature importance evaluation. The findings are interpreted in the context of the research questions, with emphasis on identifying the strongest predictors of student performance and comparing the effectiveness of different machine learning approaches.

4.1 Exploratory Data Analysis

The exploratory analysis provided insights into the distribution and relationships of academic, behavioral, and socio-economic attributes. Figure 4.1 illustrates the distributions of numerical variables. Hours studied ranged between 1 and 44 with a mean of approximately 20, while attendance varied from 60% to 100% with an average of 80%. Exam scores clustered tightly around the mean of 67.2, confirming the relatively narrow grade distribution in the dataset. Sleep hours were concentrated around 7 hours, while physical activity was centered between 2 and 4 sessions per week. These findings suggest

consistency across core behavioral factors but meaningful variation in academic engagement.

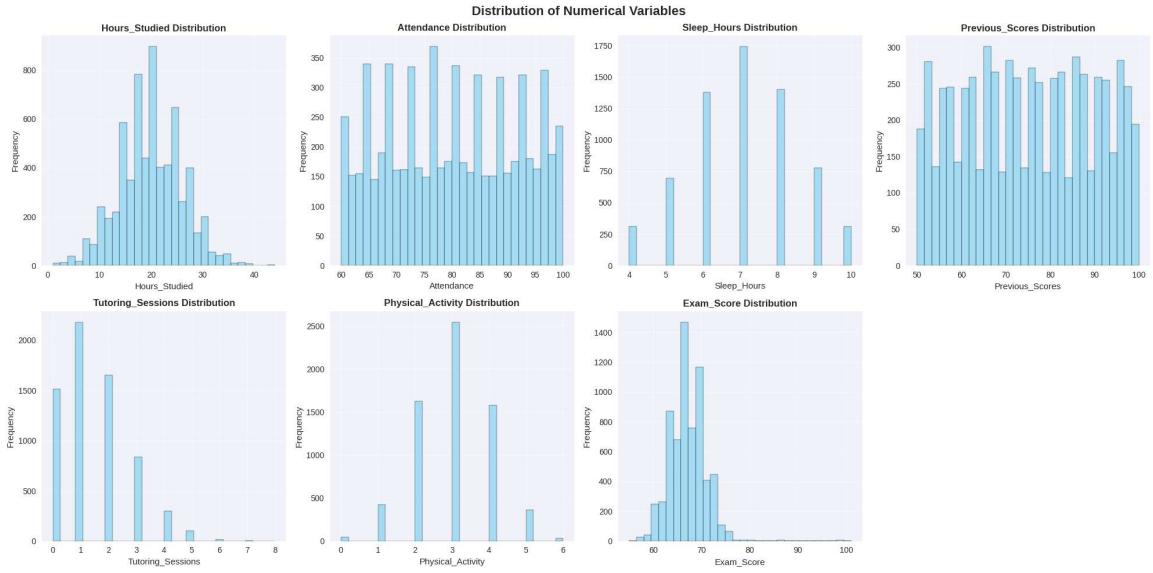

Figure 4.1: Distribution of Numerical Variables

Categorical attributes, shown in Figure 4.2, highlighted that most students reported medium levels of parental involvement (50.9%) and teacher quality (60.1%), while access to resources and internet was predominantly high. Around 69% of students attended public schools, while gender was nearly balanced (52.7% male, 47.3% female). Family income was distributed across low (40.6%), medium (40.4%), and high (19.0%) groups. These distributions suggest diverse socio-economic backgrounds, offering rich explanatory potential for modeling.

Figure 4.2: Distribution of Categorical Variables

Correlation analysis, presented in Figure 4.3, revealed that exam score had the strongest positive relationships with hours studied (r = 0.45) and attendance (r = 0.58). Previous scores (r = 0.18) and tutoring sessions (r = 0.16) also correlated positively, though with weaker magnitudes. Sleep hours and physical activity showed negligible correlations, indicating lifestyle factors may exert more indirect influences on outcomes.

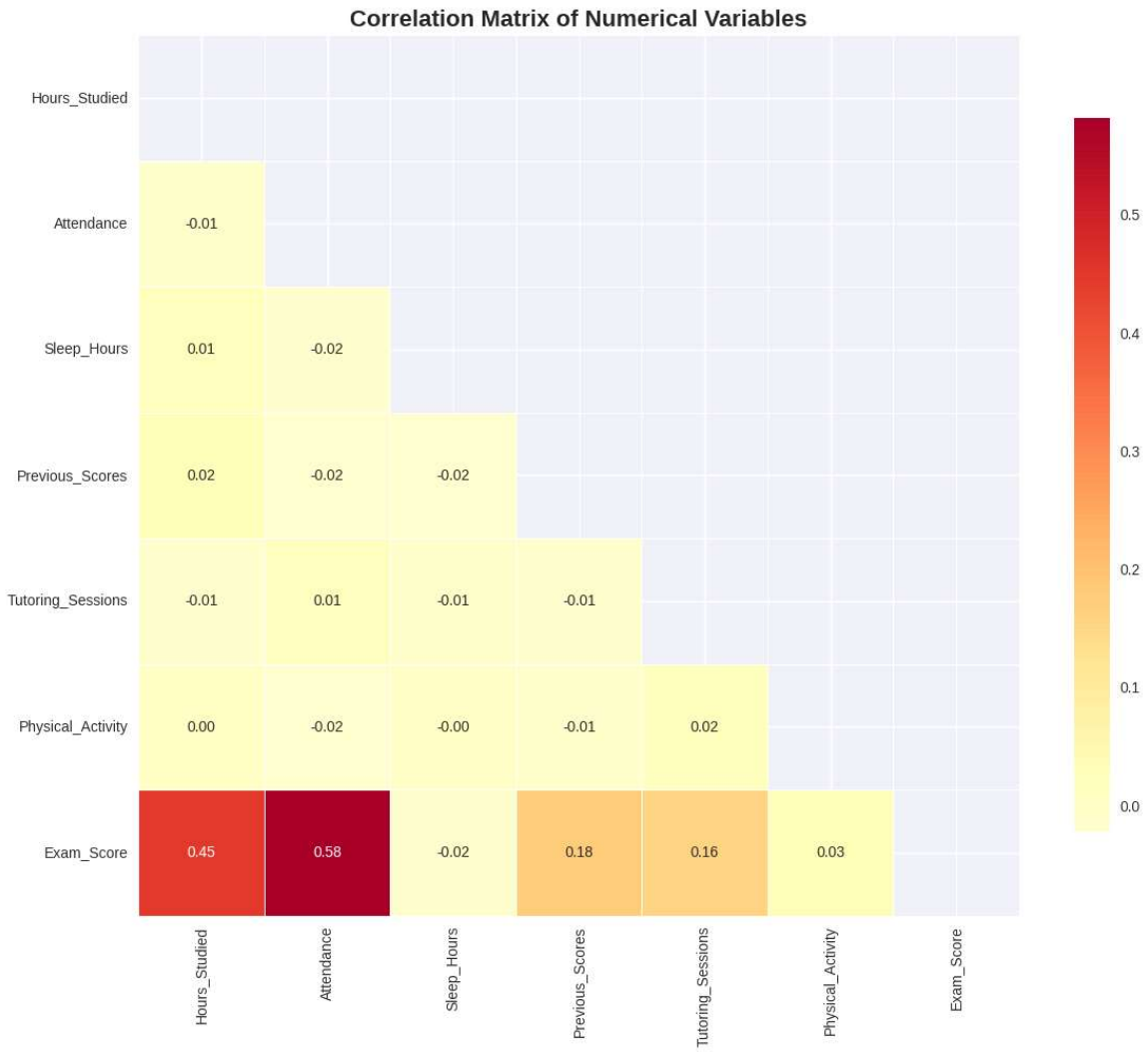

Figure 4.3: Correlation Heatmap of Numerical Variables.

Boxplots in Figure 4.4 explored how categorical factors impacted exam scores. Students with high parental involvement scored higher than those with low involvement, while teacher quality also exhibited a positive but less pronounced effect. School type showed only marginal differences, with private school students scoring slightly higher than public school counterparts. Motivation level strongly differentiated outcomes, as students reporting high motivation achieved higher median scores.

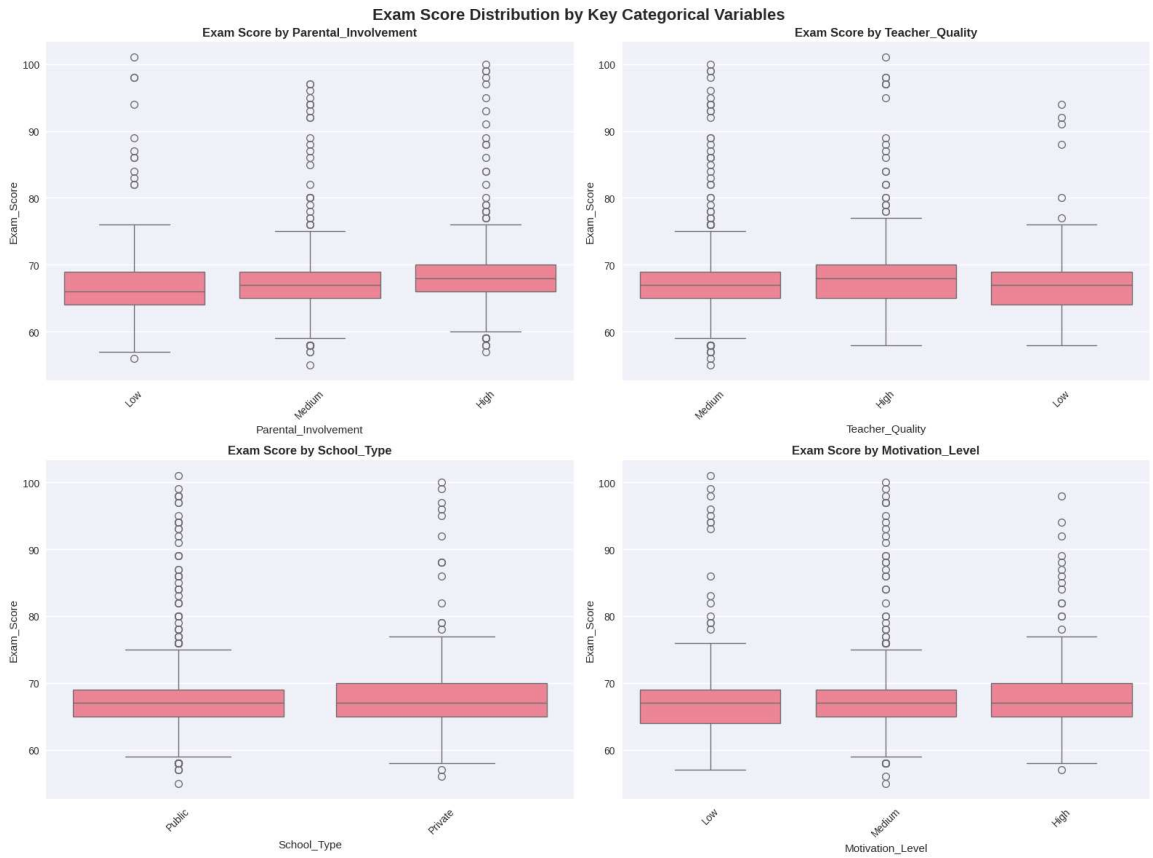

Figure 4.4: Exam Score Distribution by Categorical Variables.

Together, the EDA confirmed that academic factors such as study hours and attendance dominate performance, though contextual and support-related factors also exert measurable influence.

4.2 Regression Model Results

Regression models were developed to predict continuous exam scores. Table 4.1 summarizes their performance. Linear Regression achieved an R^2 of 0.718 with an RMSE of 1.99, while the Decision Tree Regressor slightly improved performance to $R^2 = 0.723$. Random Forest Regressor substantially outperformed both, achieving $R^2 = 0.869$ and RMSE = 1.36.

Table 4.1: Regression Model Performance

Model	MSE	RMSE	R^2 Score
Linear Regression	3.9794	1.9948	0.7185
Decision Tree	3.9138	1.9783	0.7231
Random Forest	1.8552	1.3621	0.8688

The actual vs. predicted scatter plots (Figure 4.5) reinforced these findings. Predictions from Random Forest were tightly aligned with the ideal regression line, whereas Linear

Regression and Decision Tree exhibited greater spread, particularly for higher exam scores. These results confirm the superior predictive capacity of ensemble-based regression for this dataset.

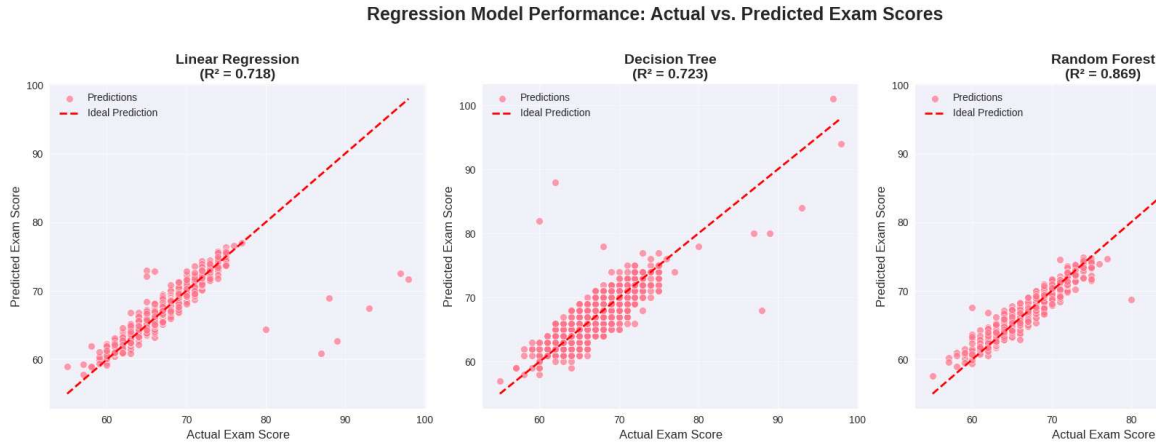

Figure 4.5: Regression Model Performance – Actual vs Predicted

4.3 Classification Model Results

Classification models were evaluated on their ability to categorize students into high-performing and low-performing groups. Logistic Regression achieved an accuracy of 92.1%, outperforming KNN at 87.4%. As shown in Table 4.2, Logistic Regression demonstrated balanced precision (0.91–0.93) and recall (0.90–0.93) across both classes, yielding an overall F1-score of 0.92. KNN, by contrast, achieved lower precision and recall across both classes.

Table 4.2: Classification Model Performance Metrics

Model	Accuracy	Precision (Class 0)	Recall (Class 0)	F1 (Class 0)	Precision (Class 1)	Recall (Class 1)	F1 (Class 1)
Logistic Regression	0.9213	0.91	0.90	0.91	0.93	0.93	0.93
K-Nearest Neighbors	0.8737	0.85	0.85	0.85	0.89	0.89	0.89

Classification report further confirmed these results, with Logistic Regression demonstrating fewer false positives and false negatives compared to KNN. The findings highlight Logistic Regression as the more reliable classifier, particularly for educational interventions where minimizing misclassification is critical.

4.4 Clustering Analysis

Unsupervised learning was applied to identify natural groupings of students. K-Means clustering achieved a silhouette score of 0.1292, indicating weak but interpretable cluster separation. PCA and t-SNE projections (Figure 4.7) revealed overlapping but distinct groupings, suggesting that performance drivers may not form sharply defined clusters. Hierarchical clustering (Figure 4.8) produced a dendrogram with several nested subgroups, reinforcing the presence of nuanced patterns rather than clear separations.

Cluster heatmaps (Figure 4.9) highlighted the characteristics of each group. One cluster was defined by high study efficiency and strong parental involvement, while another was marked by lower study efficiency but stronger academic support. These groupings, though not sharply distinct, provide a lens for exploring targeted intervention strategies.

4.5 Feature Importance

The Random Forest feature importance analysis (Figure 4.10) provided a ranked view of predictive drivers. Attendance emerged as the most influential factor (importance score = 0.349), followed by Study Efficiency (0.264) and Hours Studied (0.232). Previous scores contributed moderately, while factors such as academic support, tutoring sessions, and

resource access exerted smaller influences. This confirms the dominant role of academic engagement variables, complemented by contextual factors that refine predictive accuracy.

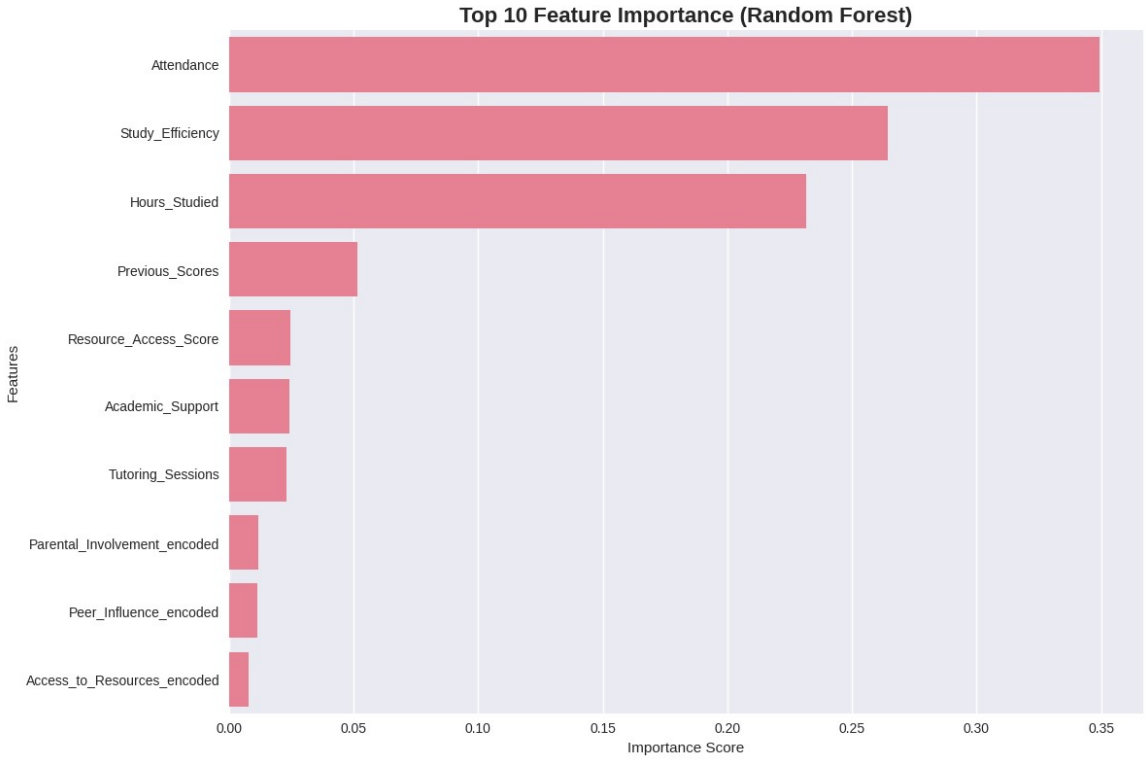

Figure 4.10: Top 10 Feature Importance – Random Forest

4.6 Discussion of Key Findings

The results provide evidence that academic engagement variables such as attendance and hours studied are the strongest predictors of exam performance. Regression results demonstrated the superiority of ensemble methods, with Random Forest outperforming simpler models. In classification, Logistic Regression achieved higher accuracy and balance across classes than KNN, reinforcing its suitability for identifying at-risk students. Clustering analysis revealed weak but meaningful subgroup distinctions, which could support tailored interventions in educational settings. Feature importance analysis further confirmed the central role of study efficiency, attendance, and hours studied, underscoring the need for strategies that enhance academic consistency and productive study habits.

The findings align with existing literature, which has consistently emphasized the predictive power of attendance and study time while recognizing the contextual role of socio-economic and behavioral attributes. However, this study extends prior work by integrating feature engineering and multi-method comparisons on a larger dataset. The implications suggest that predictive analytics can support educators in identifying risk

factors early, but clustering-based segmentation may require refinement with richer features to produce actionable subgroups.

Chapter 5: Conclusion and Future Work

This dissertation set out to investigate the factors influencing student performance and to evaluate the effectiveness of different machine learning approaches in predicting academic outcomes. Using the *Student Performance Factors* dataset comprising 6,607 student records and 20 variables, the study applied regression, classification, clustering, and feature importance analysis to provide a comprehensive understanding of academic achievement drivers.

The findings strongly emphasized the role of academic engagement as the primary determinant of performance. Exploratory data analysis revealed that attendance, hours studied, and previous scores exhibited the strongest associations with exam outcomes, while socio-economic and behavioral attributes provided additional context. Regression models confirmed that ensemble methods such as Random Forest deliver superior predictive performance, achieving an R^2 of 0.869 with an RMSE of 1.36, far outperforming linear and decision tree regressors. Classification results demonstrated that Logistic Regression achieved 92.1% accuracy with balanced precision and recall across classes, outperforming K-Nearest Neighbors (87.4%). Clustering analysis, although limited by a modest silhouette score of 0.1292, revealed interpretable subgroups characterized by variations in study efficiency, academic support, and parental involvement. Random Forest feature importance rankings further validated attendance, study efficiency, and hours studied as the most influential predictors.

5.1 Research Questions Revisited

- **RQ1: How effectively can regression models predict student exam scores using demographic, behavioral, and contextual factors?**
  The regression analysis demonstrated that Random Forest Regressor achieved strong predictive accuracy ($R^2 = 0.869$), significantly outperforming simpler models. This confirms that regression models, particularly ensemble approaches, can effectively predict continuous exam scores.

- **RQ2: Do classification models outperform regression models in identifying high-performing versus low-performing students?**
  The results confirmed that classification models, particularly Logistic Regression, are highly effective in identifying high- and low-performing groups, achieving 92.1% accuracy. While regression provides continuous score predictions, classification offers practical utility for educational interventions by categorizing students into risk groups.

- **RQ3: Which factors—academic, socio-economic, or behavioral—contribute most significantly to predicting student performance?**
  Feature importance analysis demonstrated that academic factors dominate predictive power, with attendance, study efficiency, and hours studied ranking highest. Socio-economic and behavioral features such as parental involvement and academic support played supporting roles, contributing to nuanced insights but with smaller importance scores.

5.2 Limitations

While the study yielded valuable insights, several limitations should be acknowledged. The dataset was limited to anonymized survey responses, which may not fully capture the complexity of student experiences. Certain features, such as motivation and teacher quality, relied on categorical scales, reducing precision compared to continuous measures. The clustering analysis achieved only a modest silhouette score, suggesting that the chosen features were insufficient for generating strongly distinct subgroups. Finally, the lack of external validation on independent datasets constrains the generalizability of the findings.

5.3 Future Work

Future work might overcome the limitations in a few different ways. Increasing the dataset with more behavioral or contextual features like time management approaches or digital learning activity might allow for increased predictive accuracy. Higher-performance models like Gradient Boosting Machines or XGBoost or multilayers might then improve results. Clustering might gain clearer and more interpretable subgroups by incorporating richer socio-emotional variables or longer-term performance information. Ultimately, modeling at the cross-institutional or real-world educational level would allow for greater external validation and greater applicability.

5.4 Concluding Remarks

Overall, this study verifies that machine learning offers efficient tools for forecasting and modeling student performance. Measures of academic engagement, specifically attendance and study hours, are still the best predictors of success, while contextual and socio-economic variables contribute complementary information. Regression and classification models provide robust predictive frameworks, while Random Forest and Logistic Regression proved best performing techniques. Despite not performing best, clustering presented explorative value for uncovering concealed patterns. As a whole, the investigation contributes towards the burgeoning sphere of educational data science and reveals practical outcomes for educators interested in determining at-risk learners and maximizing learning results.

References

l-Barrak, M. A., & Al-Razgan, M. (2016). Predicting Students' Final GPA Using Decision Trees: A Case Study. International Journal of Information and Education Technology, 6(7), 528–533. https://doi.org/10.7763/IJIET.2016.V6.745

Baker, R. S., & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. Learning Analytics: From Research to Practice, 61–75. https://doi.org/10.1007/978-1-4614-3305-7_4

Bhardwaj, B. K., & Pal, S. (2011). Data Mining: A prediction for performance improvement using classification. IJCSIS) International Journal of Computer Science and Information Security, 9(4).

(PDF) Predicting Students Drop Out: A Case Study. (n.d.). Retrieved August 28, 2025, from https://www.researchgate.net/publication/221570467_Predicting_Students_Drop_Out_A_Case_Study

Romero, C., & Ventura, S. (2024). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3). https://doi.org/10.1002/widm.1355

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis. Frontiers in Psychology, 12, 698490. https://doi.org/10.3389/FPSYG.2021.698490/BIBTEX