Journal of Data Intelligence and Applied Machine Learning
Volume 1, Number 1, 2026, pp. 15-24
Journal Homepage: https://journal.makwafoundation.org/index.php/datamind

Classification of Students' Academic Performance Using Random Forest Based on Behavioural and Environmental Factors

Elvi Yanti Tanjung[1*], Rais Amar Budiman[2], Salman Al Zikri[3]
[1,2,3] Faculty of Science and Technology, UIN Sjech M. Djamil Djambek, Bukittinggi, Indonesia
*Corresponding author: elpiyantitanjung@gmail.com

Received 17 March 2026; revised 02 May 2026; accepted 02 June 2026; published 07 July 2026

Abstract. Student academic performance is an important indicator of educational quality. This study implements the Random Forest algorithm to classify student academic performance based on behavioural and environmental support factors, whilst explicitly examining the effect of class imbalance commonly found in educational data. The dataset used is the Student Performance Dataset from Kaggle, comprising 2,392 student records. The research stages include data understanding, data cleaning, outlier handling using the Interquartile Range (IQR) method and winsorisation, exploratory data analysis (EDA), feature selection, data standardisation, an 80:20 data split, model training, evaluation, 5-fold cross-validation, and hyperparameter tuning using GridSearchCV. The results show that the Random Forest model achieved 69.9% accuracy on the test data, with a 5-fold cross-validation mean accuracy of 69.9% (standard deviation 0.054). The model performs considerably better on the majority class (precision 0.90; recall 0.91) than on minority classes, particularly the lowest-performing class (recall only 0.14); therefore, claims regarding the model's generalisation ability must be interpreted with caution. Feature importance analysis identifies 'Absences' (45.3 per cent) and 'StudyTimeWeekly' (18.5 per cent) as the two most influential factors, followed by 'ParentalSupport' and 'ParentalEducation' as environmental factors. These findings support the development of an attendance- and study-time-based early warning system, with the caveat that techniques for handling class imbalance, such as SMOTE, are recommended for future work.

Keywords: Random Forest; Classification; Academic Performance; Class Imbalance; Feature Importance

1. Introduction

Pupils' academic performance is a reflection of various interacting factors, ranging from study habits and school attendance to family support. In the era of educational digitalisation, the use of machine learning techniques to analyse and predict students' academic performance is growing rapidly. Batool et al. [1] in a systematic review emphasise that the ability to identify at-risk students at an early stage enables educational institutions to carry out targeted and efficient interventions.

Various studies have been conducted to predict students' academic performance using a range of classification algorithms. Ahmed [2] demonstrated that machine learning algorithms are capable of predicting student performance with competitive accuracy. Meanwhile, Jawad et al. [3] and Khosravi and Azarnik [4] demonstrated the superiority of ensemble learning algorithms, particularly Random Forest and XGBoost, in handling educational data containing a mix of numerical and categorical variables.

Random Forest is a decision-tree-based ensemble algorithm that works by building multiple decision trees in parallel and combining the results through majority voting [16]. Nachouki et al. [5] demonstrated that Random Forest excels at identifying key predictors in student academic performance data. Chen et al. [6] also confirmed that Random Forest consistently achieves high accuracy in predicting academic performance using educational data mining.

One challenge that is often overlooked in research on academic performance prediction is class imbalance in the target variable. Haixiang et al. [18] demonstrated that imbalanced data tends to bias the model towards the

majority class and reduces the model's ability to recognise the minority class, whilst Chawla et al. [17] introduced the synthetic oversampling (SMOTE) technique as a common solution to this problem. The Student Performance dataset used in this study also exhibits this imbalance; therefore, this issue needs to be discussed explicitly—rather than merely mentioned in passing—to ensure that the interpretation of the model's results is not misleading.

This study aims to implement the Random Forest algorithm to classify students' academic performance based on the Student Performance Dataset from Kaggle, which contains 2,392 student records with 15 attributes. It is hoped that this study will identify the behavioural and environmental factors that most significantly influence academic performance, as well as the impact of class imbalance on model reliability, so that the research findings can serve as a basis for more realistic, data-driven educational policy-making.

The novelty of this study lies in the combination of feature importance analysis with comprehensive hyperparameter tuning, accompanied by an explicit evaluation of the consequences of class imbalance on model performance by category, as well as the simultaneous interpretation of behavioural and environmental factors. This study addresses the gap in the literature caused by the lack of studies that explicitly compare the relative contributions of behavioural factors (absenteeism, study time) and environmental factors (parental support, parental education) whilst also reporting the limitations of the model for minority groups, as identified by Beckham et al. [7] in their analysis of the determinants of student performance using various machine learning methods.

2. Research Methodology

This study employs a quantitative approach using computational experimental methods. The entire research process was conducted using the Python programming language, utilising the pandas, scikit-learn, matplotlib and seaborn libraries. The research stages included data collection, data understanding, data cleaning, outlier handling, exploratory data analysis (EDA), feature selection and pre-processing, data splitting, model training, model evaluation, hyperparameter tuning, interpretation of results, and drawing conclusions, as outlined in the following sub-sections.

2.1 Dataset

The dataset used is the Student Performance Dataset, obtained from the Kaggle platform. This dataset contains 2,392 rows of student data with 15 attributes covering demographic information, learning behaviour, parental involvement, extracurricular activities, and academic performance. A description of the dataset variables is presented in Table 1.

Table 1. Description of Variables in the Student Performance Dataset

No	Variable	Data Type	Description
1	StudentID	Integer	Unique student ID (1001-3392)
2	Age	Integer	Student age (15-18 years)
3	Gender	Binary (0/1)	Student sex
4	Ethnicity	Categorical (0-3)	Student ethnicity category
5	ParentalEducation	Categorical (0-4)	Parents' educational attainment
6	StudyTimeWeekly	Float	Weekly study time in hours
7	Absences	Integer	Number of days absent (0-29)
8	Tutoring	Binary (0/1)	Participation in private tutoring
9	ParentalSupport	Categorical (0-4)	Level of parental learning support
10	Extracurricular	Binary (0/1)	Participation in extracurricular activities
11	Sports	Binary (0/1)	Participation in sports activities
12	Music	Binary (0/1)	Participation in music activities
13	Volunteering	Binary (0/1)	Participation in volunteering activities
14	GPA	Float	Grade point average (0-4.0)
15	GradeClass	Categorical target (0-4)	Academic performance class used as target variable

Source: Kaggle Student Performance Dataset.

It should be emphasised from the outset that the distribution of target classes (GradeClass) in this dataset is unbalanced: class 4 (F) dominates with over 1,200 data points, whilst class 0 (A) comprises only around 100 data points (see Figure 2 in subsection 2.5). This imbalance is a common challenge in educational data mining [17], [18] because classification models tend to be biased towards predicting the majority class, meaning that overall accuracy metrics may mask the model's weakness in identifying students in the minority class. This issue is an important consideration when interpreting all results in Section 3.

2.2 Data Understanding

The data understanding stage was carried out by analysing the dataset structure using the info() and describe() functions from the pandas library. This analysis revealed that the dataset comprises 15 columns with integer and float data types; there are no missing values and no duplicate data. Descriptive statistics show that the Absences variable ranges from 0 to 29 days and StudyTimeWeekly ranges from 0 to nearly 20 hours per week.

2.3 Data Cleaning

The data cleaning process involved checking for missing values, identifying duplicate data, and validating the value ranges of each numeric variable. The results of the checks showed that the dataset was clean, with no missing values or duplicate data. Data validation was carried out by verifying the minimum and maximum values of each numeric column to ensure data consistency; for example, 'Age' falls within the range of 15–18 years and 'GPA' within the range of 0–4.0, in accordance with the variable definitions.

2.4 Outlier Handling

Outliers were handled using the Interquartile Range (IQR) method on all numeric features, except for StudentID and GradeClass. Values outside the lower bound and upper bound were replaced with the corresponding boundary values using the winsorisation method, so that the total number of data points was maintained whilst extreme values were adjusted. The results of outlier detection for each variable are presented in Table 2.

Table 2. Outlier Detection and Handling Results

Variable	Number of Outliers	Action
Age	0	No action
Gender	0	No action
Ethnicity	0	No action
Parental Education	120	Winsorisation using the IQR limits
Weekly Study Time	0	No action
Absences	0	No action
Tutoring	0	No action
Parental Support	0	No action
Extracurricular	0	No action
Sports	0	No action
Music	471	Winsorisation using the IQR bounds
Volunteering	376	Winsorisation using the IQR limits
GPA	0	No action

Table 2 shows that the majority of numerical variables do not contain significant outliers. The highest number of outliers was found in the Music variable (471 data points) and the Volunteering variable (376 data points) because both variables are binary, with the proportion of class 1 being significantly smaller than that of class 0; consequently, minority values were detected as outliers by the IQR method, even though they are, in substance, valid data. The 'ParentalEducation' variable has 120 outliers due to the uneven distribution of parental education categories. The detection of outliers in this binary variable is a mathematical consequence of the IQR method applied to an unbalanced data distribution; it does not indicate any data recording errors.

As all outliers were handled using winsorisation rather than row deletion, the final data set remains at 2,392 rows.

2.5 Exploratory Data Analysis (EDA)

EDA was carried out to understand the distribution and relationships between variables in the dataset. The analysis included visualisations of the target class distribution (GradeClass), histograms of all variables, and a correlation heatmap.

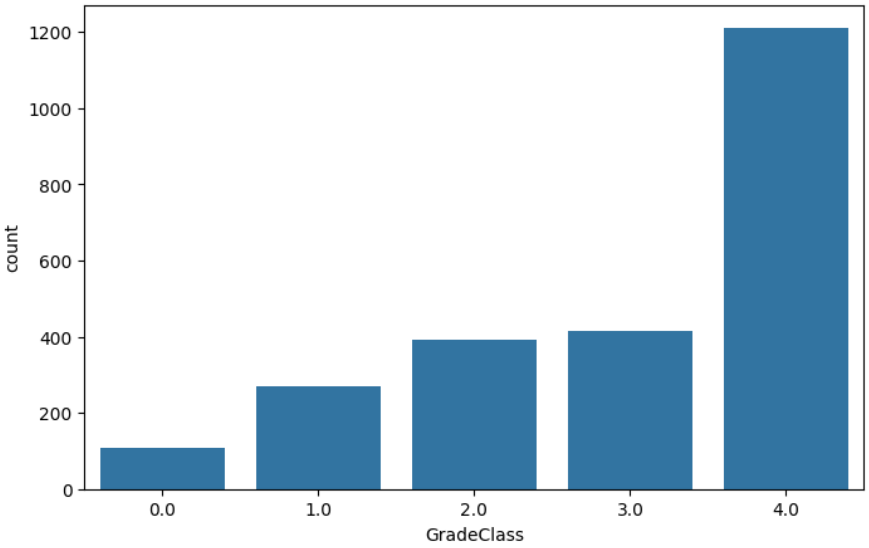

Figure 1. GradeClass distribution

Based on Figure 1, the distribution of the target class is skewed, with class 4 (F) dominating with over 1,200 data points, whilst class 0 (A) has only around 100–120 data points. This situation forms the basis for the warning given in sub-section 2.1 and is a key consideration in the interpretation of the model evaluation results in Section 3.

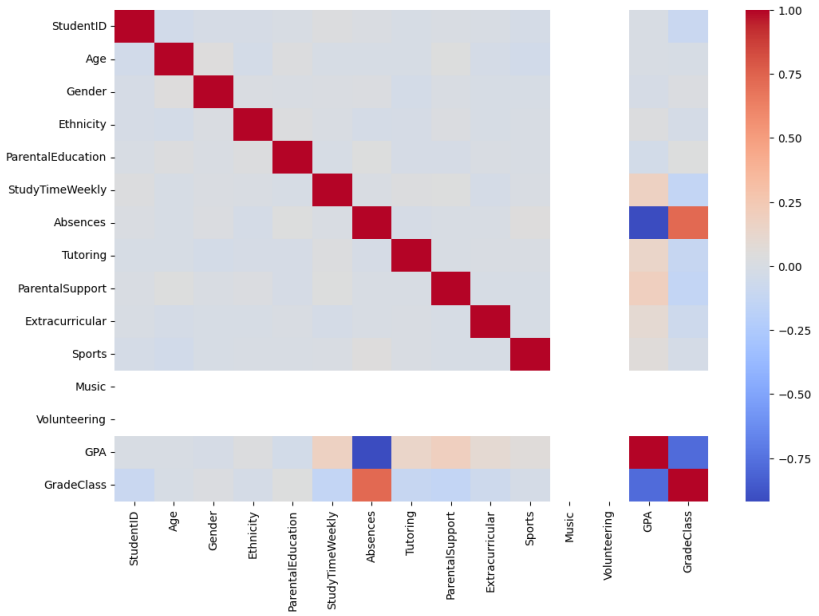

Figure 3. Heatmap of Correlations Between Variables

The correlation heatmap in Figure 3 shows that GPA has a very high correlation with GradeClass, confirming the need to exclude GPA as a precaution against data leakage, since GPA essentially forms the basis for determining the GradeClass category itself. The Absences variable appears to have a relatively strong negative correlation with both GPA and GradeClass compared to other variables, followed by StudyTimeWeekly.

When considered within the context of Indonesia's educational demographics, the finding that Absences and StudyTimeWeekly are dominant factors aligns with various international studies on student absenteeism. Klein et al. [21] found that both unexcused absences and absences due to illness have a negative impact on academic achievement, whilst Swiderski et al. [9] confirmed that this relationship remained consistent both before and after the pandemic. In the context of Indonesian education, factors such as the distance from home to school, limited access to transport, and the need for some pupils to help support their families' livelihoods are common causes of absenteeism and low levels of independent study time outside school hours.

Consequently, these two behavioural factors are relevant as key indicators for an early warning system within the school environment, without overlooking the influence of parental support, which also contributes, as discussed in sub-section 3.4.

2.6 Feature Selection and Preprocessing

Feature selection was carried out by removing the StudentID column (which is unique but irrelevant) and the GPA column (which has the potential to cause data leakage due to its very high correlation with the target variable, GradeClass). The features used for modelling are Age, Gender, Ethnicity, ParentalEducation, StudyTimeWeekly, Absences, Tutoring, ParentalSupport, Extracurricular, Sports, Music, and Volunteering.

Data standardisation was carried out using the StandardScaler from scikit-learn so that all features have a uniform scale with a mean of 0 and a standard deviation of 1. Although the Random Forest algorithm is generally not sensitive to differences in feature scale, the standardisation process was still carried out to maintain consistency in the pre-processing stages in line with the research implementation, as well as to facilitate the reproduction and further development of the research using other algorithms.

2.7 Data Split

The dataset was split into training data (80%) and test data (20%) using the train_test_split function with the stratify=y parameter to maintain the same class proportions between the training and test data, given the uneven class distribution as described in subsection 2.1. The total training data comprises 1,913 rows and the test data comprises 479 rows.

2.8 Model Training

The model used is the Random Forest Classifier from scikit-learn with the parameter random_state=42 to ensure reproducibility. Random Forest works by constructing a number of decision trees independently using the bootstrap aggregation (bagging) technique and random feature selection at each node split, then combining the predictions of each tree via majority voting for classification [16]. Tang and Chen [13] confirm that this approach is effective in the context of educational data mining as it is capable of handling both categorical and numerical features simultaneously.

2.9 Model Evaluation

Model evaluation was carried out using the metrics of accuracy, precision, recall, F1-score and the confusion matrix. Accuracy measures the proportion of correct predictions overall; precision measures the accuracy of positive predictions; recall measures the completeness of positive class detection; and the F1-score is the harmonic mean of precision and recall. As the class distribution is imbalanced, the precision, recall and F1-score metrics for each class are reported separately so that the model's performance on the minority class is not masked by a high overall accuracy value. Five-fold cross-validation was performed to obtain a more robust estimate of performance, as recommended by Haron et al. [14].

2.10 Hyperparameter Tuning

Hyperparameter optimisation was carried out using GridSearchCV with a parameter grid of n_estimators ∈ {100, 200}, max_depth ∈ {10, 20, None}, and min_samples_split ∈ {2, 5}, using a 5-fold cross-validation scheme on the training data with the accuracy metric. The aim of this search was to find the combination of parameters that maximises the average cross-validation accuracy.

3. Results and Discussion

3.1 Model Evaluation Results

The Random Forest model trained on the training data produced predictions for 479 test data points, which were evaluated using several metrics. The confusion matrix of the prediction results is shown in Figure 4.

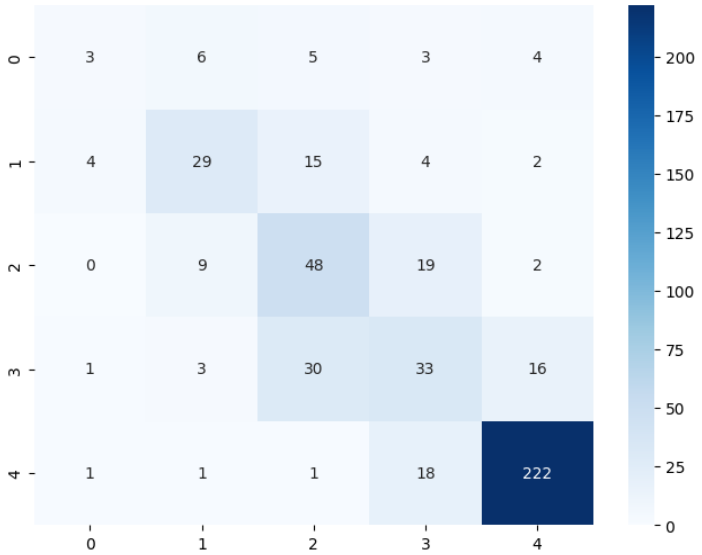

Figure 4. Confusion Matrix of the Random Forest Model's Predictions

Table 3. Evaluation Metrics for the Random Forest Model

Class	Precision	Recall	F1-Score	Support
0 (A)	0.33	0.14	0.20	21
1 (B)	0.60	0.54	0.57	54
2 (C)	0.48	0.62	0.54	78
3 (D)	0.43	0.40	0.41	83
4 (F)	0.90	0.91	0.91	243
Accuracy			0.699	479
Macro average	0.55	0.52	0.53	479
Weighted average	0.69	0.70	0.69	479

The model achieved an overall accuracy of 69.9% on the test data, with a weighted average F1-score of 0.69. As shown in Figure 4 and Table 3, the model performed best on class 4/F with a precision of 0.90 and a recall of 0.91, whilst its lowest performance was observed in class 0/A, with a precision of just 0.33 and a recall of 0.14. This means that, of the 21 pupils who were actually in class A, the model only correctly identified a small proportion, whilst the majority were misclassified into neighbouring classes (B or C).

These results confirm that the model tends to perform significantly better at recognising the majority class (GradeClass = F) compared to the minority classes (GradeClass = A, B, and D). The macro average F1-score (0.53), which is significantly lower than the weighted average F1-score (0.69), also highlights the performance gap between classes: the weighted average is higher because it is dominated by the large majority class, whilst the unweighted average reflects the model's actual, much more moderate performance across all categories.

The results of the 5-fold cross-validation yielded accuracy scores for each fold of 0.722; 0.727; 0.720; 0.734; and 0.592, with an average accuracy of 0.699 (69.9 per cent) and a standard deviation of 0.054. This average value indicates that the model performs relatively consistently across most folds. However, the drop in accuracy in the fifth fold indicates that the model's performance is still affected by class imbalance. Therefore, the cross-validation results show that the model is generally stable, but its generalisation ability for minority classes still needs to be improved through class imbalance handling techniques such as SMOTE or class weighting in future research.

The hyperparameter tuning process using GridSearchCV on the training data yielded the optimal parameter combination of max_depth=20, min_samples_split=5, and n_estimators=100, with an average cross-validation accuracy of 72.1 per cent. This value is slightly higher than the accuracy of the baseline model on the training data; however, as the tuned model was not re-evaluated separately on the test data (479 data points) in this study, all evaluation results in Table 3 and Figure 4 continue to use the Random Forest model with default parameters (random_state=42) to ensure consistency across evaluation stages. Further research is recommended to evaluate the tuned model directly on separate test data to ensure that a performance improvement has genuinely been achieved, rather than merely in the cross-validation score on the training data.

The imbalance in class distribution causes the model to learn patterns in the majority class more frequently than in the minority class. Consequently, the overall accuracy remains relatively high, but the model's ability to

recognise classes with few data points is low. This is evident from the recall value for class A, which is 0.14, whilst the recall for class F reaches 0.91.

3.2 Feature Importance Analysis

A feature importance analysis was conducted to identify the features that most influence the prediction of students' academic performance. Feature importance values in the Random Forest are calculated based on the mean decrease in Gini impurity contributed by each feature across all decision trees in the forest [16]. The greater a feature's contribution to cleanly separating the data during tree construction, the higher its importance value. The results of the analysis are shown in Figure 5 and Table 4.

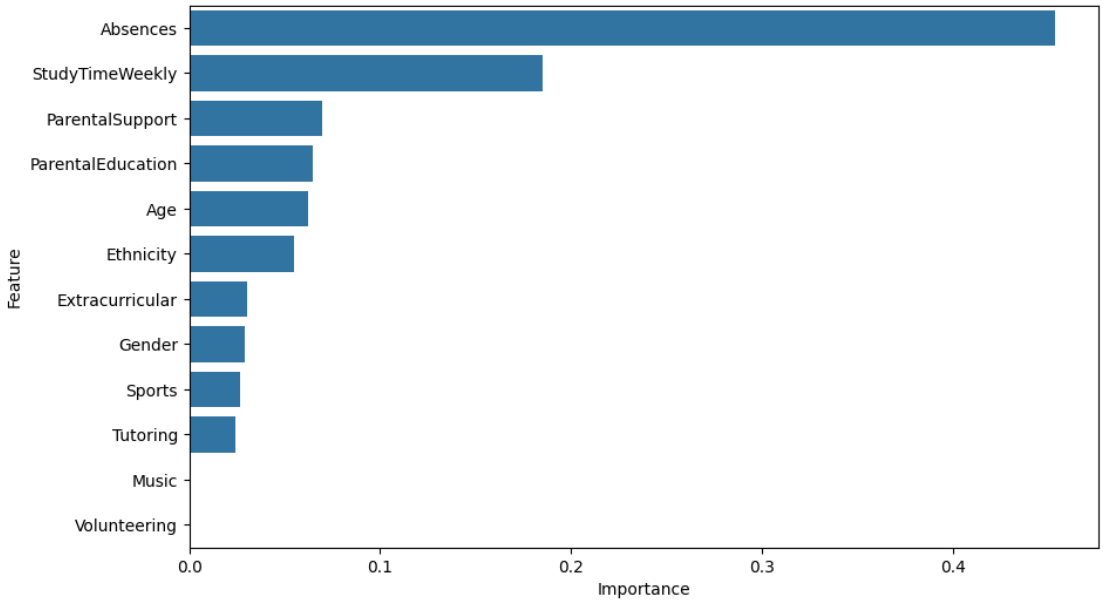

Figure 5. Visualisation of Random Forest Feature Importance

Table 4. Feature Importance Ranking

Ranking	Feature	Importance Score
1	Absences	0.453
2	Weekly Study Time	0.185
3	Parental Support	0.069
4	Parental Education	0.065
5	Age	0.062
6	Ethnicity	0.055
7	Extracurricular	0.030
8	Gender	0.029
9	Sports	0.027
10	Tutoring	0.024
11	Music	0.000
12	Volunteering	0.000

The feature importance values in Random Forest indicate the relative contribution of each feature to the decision tree construction process, based on the reduction in Gini impurity. These values should not be interpreted as indicating a cause-and-effect relationship, but rather as a measure of how much a feature helps the model to distinguish between target classes.

Based on Figure 5 and Table 4, the 'Absences' feature is the most dominant factor with an importance score of 0.453, or approximately 45.3 per cent of the total contribution of all features, indicating that student attendance is the strongest predictor of academic performance. This finding is consistent with Bowen et al. [8], who highlighted the fundamental role of absenteeism and its impact on learning outcomes, as well as Klein et al. [21], who confirmed the negative influence of absenteeism on academic achievement.

The second factor is 'StudyTimeWeekly' with an importance score of 0.185 (18.5 per cent), confirming the role of independent study time outside school hours, in line with the findings of Asselman et al. [10]. Conversely, the 'Music' and 'Volunteering' features have importance scores close to zero, indicating that these two variables make virtually no contribution to the decision-making process in the constructed Random Forest model.

3.3 Analysis of Behavioural Factors

The behavioural factors analysed include Absences, StudyTimeWeekly and Tutoring. Absences is the most influential factor with an importance of 0.453 (45.3 per cent), indicating that pupils' attendance levels have a very strong relationship with academic achievement: the higher the absence rate, the greater the likelihood of pupils experiencing a decline in performance. Apart from absences, StudyTimeWeekly was the second most influential behavioural factor with an importance of 0.185 (18.5%), indicating that students who consistently allocate time to studying tend to achieve better academic results.

Meanwhile, the Tutoring variable has an importance of only 0.024 (2.4 per cent), suggesting that private tuition is not a dominant factor in this dataset compared to independent study habits and discipline regarding school attendance. This finding is supported by Asselman et al. [10], who found that study time and student engagement consistently emerge as key predictors in machine learning-based models for predicting academic performance.

3.4 Analysis of Environmental Support Factors

The environmental support factors analysed consist of 'ParentalSupport' and 'ParentalEducation'. 'ParentalSupport' has an importance of 0.069 (6.9 per cent) and ranks third, indicating that parental involvement continues to contribute to pupils' academic performance through learning support, motivation and supervision of learning activities. ParentalEducation has an importance of 0.065 (6.5 per cent) and ranks fourth. When these two variables are combined, the contribution of environmental factors amounts to approximately 13.4 per cent of the model's total feature importance, which is significantly smaller than the combined contribution of Absences and StudyTimeWeekly, which reaches 63.8 per cent. These results indicate that, in this dataset, student behavioural factors are far more dominant than family environmental factors, although both remain statistically relevant.

This finding is consistent with Yeh et al. [11], who state that parental involvement is one of the predictors of academic success, as well as Affuso et al. [12], who show that parental support and monitoring have a positive effect on long-term academic achievement, although the magnitude of this effect is smaller than that of attendance and study time in the present study.

3.5 Limitations Related to Class Imbalance

The results in subsection 3.1 confirm that the model's performance is not uniform across all GradeClass categories. A recall of 0.14 in Class A indicates that the model is only able to identify a small proportion of high-achieving pupils, whilst moderate performance in Classes B and D (F1-scores of 0.57 and 0.41) suggests that the model still frequently misclassifies pupils in the middle categories. This finding is consistent with the literature on class imbalance in educational data [17], [18], [20], which states that an apparently good overall accuracy can mask the model's weakness in identifying minority classes.

Therefore, the assertion in previous research that the model possesses 'sufficiently good generalisation ability' needs to be revised: the model's generalisation can only be considered good for the majority class (F), whilst for minority classes (particularly A) the model is not yet practically reliable. Further treatment using the SMOTE technique [17] or class weighting, as applied by Kesgin et al. [20] and Jawad et al. [3] in similar contexts, is recommended as follow-up research to improve the balance of performance across classes.

This study utilised only one public dataset from Kaggle; therefore, the results still need to be validated on other datasets with different characteristics so that the model's generalisation ability can be evaluated more comprehensively.

3.6 Comparison with Previous Research

The accuracy of 69.9 per cent achieved in this study can be compared with similar research. Tang and Chen [13], in their AI-based educational data mining study, reported comparable accuracy in multi-class classification. Vimarsha et al. [15] used a co-evolutionary hybrid intelligence model and achieved an accuracy of up to 73 per cent on a student performance dataset, whilst Haron et al. [14], using an artificial intelligence approach, achieved an accuracy of 71 per cent. Chen and Jin [19] reported an improvement in Random Forest accuracy through the use of an additional optimiser in a similar case, whilst Kesgin et al. [20] found that, on a different student performance dataset, a linear model could actually rival Random Forest after class imbalance was addressed using SMOTE, underscoring the importance of class balancing strategies in enhancing model reliability.

Differences in performance across these studies are to be expected due to variations in dataset characteristics, the number of classes, and pre-processing methods. The Random Forest model, with an accuracy of 69.9% in this study, remains a competitive approach and is relatively easy to interpret via feature importance, although its reliability for minority classes still needs to be improved.

4 Conclusion

This study successfully implemented the Random Forest algorithm to classify pupils' academic performance based on behavioural factors and environmental support using the Student Performance Dataset comprising 2,392 pupil records. The developed model achieved an accuracy of 69.9% on the test data with a weighted average F1-score of 0.69; however, this performance was largely underpinned by the model's ability to recognise the majority class (GradeClass F) with precision and recall above 0.90, whilst performance on minority classes—particularly Class A, with a recall of just 0.14—remained significantly weaker; consequently, the model's generalisation capabilities must be interpreted proportionately and should not be over-generalised. 5-fold cross-validation yielded an average accuracy of 69.9 per cent with considerable variation between folds, indicating the model's sensitivity to class distributions in each data partition due to target class imbalance.

Feature importance analysis revealed that 'Absences' (45.3%) and 'StudyTimeWeekly' (18.5%) are the two most dominant factors influencing pupils' academic performance, followed by 'ParentalSupport' (6.9%) and 'ParentalEducation' (6.5%) as relevant environmental factors, although their contributions are much smaller. These findings reinforce the importance of programmes to improve student attendance and foster a culture of independent learning as priority interventions, without neglecting the role of parental support, in efforts to enhance the quality of education.

The results of this study provide practical recommendations for education stakeholders, namely the utilisation of the model as an early warning system based on attendance data and study time to identify students at risk of a decline in academic performance, with the caveat that the model's reliability in extreme performance categories remains limited. Further research is recommended to apply class imbalance handling techniques such as SMOTE or class weighting, to evaluate models resulting from hyperparameter tuning directly on separate test data, and to compare the performance of Random Forest with other ensemble algorithms such as XGBoost or LightGBM to obtain a more reliable model across all categories of student academic performance.

Thus, Random Forest is suitable for use as a baseline model in predicting students' academic performance; however, the application of class imbalance handling techniques is still required to ensure the model's performance across all student categories is more consistent.

5 Acknowledgements

The authors would like to thank Mr Ihdi Syahputra Ritonga, S.Kom., M.Kom. and Ms Agus Sundari, S.Kom., M.Kom. for their guidance, supervision and feedback provided throughout the preparation and conduct of this research. Thanks are also extended to my team-mates, Rais Amar Budiman and Salman Al Zikri, for their cooperation, discussions and contributions during the data processing, model development and preparation of this scientific article. Furthermore, the author would like to thank the Kaggle dataset provider for making the data openly available, thereby enabling this research to be carried out successfully.

References

[1] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H.-Y. Kwon, and A. Hussain, "Educational data mining to predict students' academic performance: A survey study," Educ. Inf. Technol., vol. 28, no. 1, pp. 905–971, 2023. doi: 10.1007/s10639-022-11152-y

[2] E. Ahmed, "Student performance prediction using machine learning algorithms," Appl. Comput. Intell. Soft Comput., vol. 2024, pp. 1–14, 2024. doi: 10.1155/2024/4067721

[3] K. Jawad, M. A. Shah, and M. Tahir, "Predicting students' academic performance and engagement in a virtual learning environment using random forest with data balancing," Sustainability, vol. 14, no. 22, p. 14795, 2022. doi: 10.3390/su142214795

[4] A. Khosravi and A. Azarnik, "Leveraging educational data mining: XGBoost and random forest for predicting student achievement," Int. J. Data Sci. Adv. Anal., vol. 6, no. 7, pp. 387–393, 2024. doi: 10.69511/ijdsaa.v6i7.229

[5] M. Nachouki, E. A. Mohamed, R. Mehdi, and M. Abou Naaj, "Predicting student course grades using the random forest algorithm: Analysis of predictor importance," Educ. Inf. Technol., vol. 28, pp. 12909–12931, 2023. doi: 10.1007/s10639-023-11624-3

[6] Z. Chen, G. Cen, Y. Wei, and Z. Li, "Student performance prediction approach based on educational data mining," IEEE Access, vol. 11, pp. 131260–131272, 2023. doi: 10.1109/ACCESS.2023.3335279

[7] N. R. Beckham, L. J. Akeh, G. N. P. Mitaart, and J. V. Moniaga, "Determining factors that affect student performance using various machine learning methods," Procedia Comput. Sci., vol. 216, pp. 597–603, 2023. doi: 10.1016/j.procs.2022.12.174

[8] F. Bowen, C. Gentle-Genitty, J. Siegler, and M. Jackson, "Revealing underlying factors of absenteeism: A machine learning approach," Front. Psychol., vol. 13, p. 958748, 2022. doi: 10.3389/fpsyg.2022.958748

[9] T. Swiderski, S. C. Fuller, and K. C. Bastian, "The relationship between student attendance and achievement, pre- and post-COVID," AERA Open, vol. 11, 2025. doi: 10.1177/23328584251371041

[10] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," Interact. Learn. Environ., vol. 31, no. 6, pp. 3360–3379, 2023. doi: 10.1080/10494820.2021.1928235

[11] Y.-C. Yeh, C.-H. Lin, and Y.-S. Hsu, "Predicting academic success: Machine learning analysis of student, parental, and school efforts," Asia Pac. Educ. Rev., vol. 27, pp. 19–34, 2024. doi: 10.1007/s12564-023-09915-4

[12] G. Affuso, R. Bacchini, and M. Miranda, "Effects of teacher support, parental monitoring, motivation, and self-efficacy on academic performance over time," Eur. J. Psychol. Educ., vol. 38, no. 1, pp. 1–23, 2023. doi: 10.1007/s10212-021-00594-6

[13] L. Tang and S. Chen, "Educational data mining for student performance prediction in an artificial intelligence environment," Mol. Cell. Biomech., vol. 22, no. 5, p. 692, 2025. doi: 10.62617/mcb692

[14] N. H. Haron, R. Mahmood, N. M. Amin, A. Ahmad, and S. R. Jantan, "An artificial intelligence approach to monitor and predict student academic performance," J. Adv. Res. Appl. Sci. Eng. Technol., vol. 44, no. 1, pp. 105–119, 2025. doi: 10.37934/araset.44.1.105119

[15] K. Vimarsha, S. S. Prakash, K. Krinkin, and Y. A. Shichkina, "Student performance prediction: A co-evolutionary hybrid intelligence model," Procedia Comput. Sci., vol. 235, pp. 436–446, 2024. doi: 10.1016/j.procs.2024.04.044

[16] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002. doi: 10.1613/jair.953

[18] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," Expert Syst. Appl., vol. 73, pp. 220–239, 2017. doi: 10.1016/j.eswa.2016.12.035

[19] Y. Chen and K. Jin, "Educational performance prediction with random forest and innovative optimisers: A data mining approach," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 3, 2024. doi: 10.14569/IJACSA.2024.0150308

[20] K. Kesgin, S. Kiraz, S. Kosunalp, and B. Stoycheva, "Beyond performance: Explaining and ensuring fairness in student academic performance prediction with machine learning," Appl. Sci., vol. 15, no. 15, p. 8409, 2025. doi: 10.3390/app15158409

[21] M. Klein, E. M. Sosu, and S. Dare, "School absenteeism and academic achievement: Does the reason for absence matter?," AERA Open, vol. 8, 2022. doi: 10.1177/23328584211071115

[22] F. T. Johora, R. Mahbub, M. A. Roushan, M. M. Rahman, M. B. Mollah, M. S. Rahman, M. A. Rahman, and M. M. Rahman, "A comprehensive benchmark study of machine learning and fairness for predicting academic performance," Appl. Soft Comput., vol. 146, p. 110871, 2023. doi: 10.1016/j.asoc.2023.110871

[23] B. Albreiki, M. Zohdy, and H. Lutfiyya, "A comparison of machine learning algorithms for predicting student academic performance in higher education," Int. J. Emerg. Technol. Learn., vol. 16, no. 4, pp. 76–90, 2021. doi: 10.3991/ijet.v16i04.17123

[24] M. J. Khan, M. A. Khan, and M. S. Hossain, "Machine learning algorithms for student performance evaluation: A case study," Educ. Inf. Technol., vol. 26, pp. 3271–3289, 2021. doi: 10.1007/s10639-020-10406-6

[25] F. T. Johora, M. N. Hasan, and A. Rajbongshi, "An explainable AI-based approach for predicting undergraduate students' academic performance," Array, vol. 26, p. 100384, 2025. doi: 10.1016/j.array.2025.100384

[26] M. Gusnina, Wiharto, and U. Salamah, "Predicting student performance at Sebelas Maret University using the random forest algorithm," Ingénierie des Systèmes d'Information, vol. 27, no. 3, pp. 507–514, 2022. doi: 10.18280/isi.270321