A DATA-DRIVEN STRATEGY FOR EDUCATIONAL INTERVENTIONS: USING RANDOM FOREST CLASSIFIER TO FORECAST STUDENTS' ACADEMIC PERFORMANCE (SAP)

Sameer Jain[1]

Abstract

This study presents an original data-driven framework for predicting student academic performance using a random forest classifier (RFC). The research, by incorporating socio-demographic, behavioural, and academic factors in a single model, transcends individual factors to offer holistic solutions for early academic interventions. Using a sample from 2,392 students, the model is highly accurate in predicting grades from A to F. The research also shows that the most significant factors for academic success are weekly study habits, attendance, and parental support, with gender and ethnicity having a lesser impact. From a methodological standpoint, the research illustrates the potential for understandable machine learning in decision-making in education. From a practical standpoint, the research supports evidence-based interventions for improved student outcomes.

Keywords: Student Performance Prediction, Academic Achievement, Machine Learning, Random Forest Classifier, Higher Education

INTRODUCTION

Student academic performance is a vital metric that determines the success of an educational system as well as individual students. With the advent of data-driven decision-making in the field of education, there is a growing need to utilize existing data related to students to identify patterns that significantly affect academic performance. Several socio-demographic and behavioural factors are believed to significantly influence academic performance. These factors are interconnected in a complex manner. The purpose of this research is to utilize a real-world dataset that includes various academic as well as personal attributes of 2,392 students. The objectives of this research are as follows:

a) To analyse the importance of various factors related to academic performance.
b) To build a robust predictive model that classifies students according to their grades using a Random Forest Classifier.
c) To identify factors that are significantly related to low academic performance among students.
d) To narrow the gap between data-driven decision-making and educational policy by developing effective intervention strategies.

By using categorical variables (such as gender and ethnicity) and continuous variables (such as study time and absences), we will be able to determine the most significant predictors of student success. The main purpose of the research is to offer significant and useful findings that could be used to guide teachers, parents, and educational authorities on how to enhance student performance.

Though various researchers have extensively studied the different factors that contribute to academic performance individually, most of them have studied the factors as independent variables rather than an integrated system. With the increase of digital learning environments and data analysis in learning institutions, the need for integrated models that could simultaneously consider various academic, behavioural, and social factors is on the increase. Additionally, the statistical analysis of the factors is mostly linear and therefore of little use.

---

[1] Faculty at NICMAR Business School, NICMAR University, India

This study is significant because it operationalizes a holistic, predictive, and interpretable machine learning approach to student performance, moving beyond correlation-based analysis to actionable forecasting. By employing a Random Forest Classifier, the research not only predicts academic outcomes with high accuracy but also identifies the relative importance of different predictors, thereby enabling data-informed policy and institutional decision-making.

LITERATURE REVIEW

Significant research has been conducted on the determinants of academic performance among students, and various statistical and machine learning methods have been applied. Academic performance is considered a complex construct, and its determinants are diverse, ranging from demographic, psychological, and socio-economic factors.

Demographic and Social Factors**:** According to Tinto, social integration and family background are significant factors in student retention and academic performance. Sirin also focuses on the impact of socio-economic status on academic performance. He has related parental education, income, and occupation with academic performance. Gender is another factor, and research has shown that in academic settings, females tend to perform better than males.

Study Behaviour and School Engagement**:** Research by Kuh et al. (2008) indicates that student engagement—including time spent on homework and participation in extracurricular activities—correlates positively with GPA. Attendance and class participation are particularly strong predictors, as identified by Credé, Roch, and Kieszczynka (2010), who found a significant link between absenteeism and lower academic performance.

Machine Learning in Education**:** Machine learning models have become popular tools for educational data mining. Romero and Ventura (2010) reviewed the use of classification algorithms, such as decision trees and support vector machines, in predicting student performance. More recently, Random Forests have emerged as effective classifiers due to their ability to handle mixed data types and prevent overfitting (Breiman,2001). Academic results have been successfully predicted using student background, behavioural data, and academic records in studies like those conducted by Al-Barrak and Al-Razgan (2016), who used Random Forests. Jaya Prakash et al. (2014, 2018) demonstrated that early-warning predictive systems using ensemble models could identify at-risk students with greater accuracy than traditional statistical methods. Delen (2019) employed Random Forest and Gradient Boosting to predict student attrition, highlighting the value of feature importance analysis in institutional decision-making. Baker and Siemens (2020) emphasized the role of learning analytics in personalized education and recommended predictive modeling for proactive intervention. Zhang et al. (2021) used ensemble machine learning to classify academic risk levels, finding absenteeism and engagement as dominant predictors. This research is trying to integrate various factors such as demographic, behavioural, and institutional factors into a single model to increase the accuracy of predictions and obtain a more complete picture, while earlier research has analyzed individual predictors in isolation.

RESEARCH METHODOLOGY

The approach used in this research includes data preparation, exploratory analysis, model construction, and assessment.

Dataset Description: The dataset contains data from 2,392 students, each with 33 characteristics, including demographic information (such gender, ethnicity), academic metrics (such GPA, Grade Class), behavioural data (such study time, absences), and variables related to parental and institutional support through Kaggle datasets.

Data pre-processing: The raw dataset included numerical and categorical data. The main actions were as follows:

a) Handling missing values (if any).

b) Converting categorical variables (e.g., 'gender', 'ethnicity', 'parental level of education') into numerical form using Label Encoding.

c) Dropping irrelevant features like 'student and' which do not contribute to model learning.

d) Visualizing correlation matrices to understand feature interdependencies.

Exploratory Data Analysis (EDA): To illustrate distributions and correlations, heatmaps, distribution plots, and bar plots were utilized. These visual aids helped in determining key predictors, such as the amount of time spent studying, absences, and the level of parental education.

Model Selection and Training: Because of its robustness and interpretability, a Random Forest Classifier was selected. With 80:20 ratio, the dataset was divided into training and testing groups. Hyperparameters, like the quantity of estimators, were chosen experimentally, and the default settings offered adequate performance.

Model Evaluation: Accuracy, classification of report measures (precision, recall, and F1-score), and confusion matrix were used to assess the trained model. The Random Forest model had a high accuracy on the test set, suggesting that it has strong generalization ability.

Feature Importance**:** To enhance interpretability, feature importance values were extracted from the model to rank the top predictors influencing Grade Class. This helped identify actionable variables that educational institutions can target for improvement.

Research Hypotheses

H1: Higher weekly study time is positively associated with better academic performance.

H2: Higher absenteeism is negatively associated with academic performance.

H3: Greater parental support is positively associated with higher GPA and grade classification.

H4: Parental education level significantly influences student academic outcomes.

H5: Behavioural engagement (tutoring, extracurriculars, sports) positively correlates with performance but with weaker impact than study time and attendance.

STEPS IN RESEARCH EXECUTION

Step 1: Obtained the Dataset and Loaded the Data Using Python

Step 2: Explored, Cleaned, and Pre-processed the Data

Attributes and Their Descriptions

The dataset includes the following key variables:

a) StudentID: Unique identifier for each student (dropped for modeling).

b) Age: Age of the student.

c) Gender: Binary categorical feature indicating male or female.

d) Ethnicity: Categorical feature representing demographic background.

e) Parental Education: Education level of the student's parents.

f) StudyTimeWeekly: Average number of hours studied per week.

g) Absences: Number of classes missed.

h) Tutoring: Binary variable indicating participation in tutoring.

i) Parental Support: Level of parental support (from None to Very High).

j) Extracurricular: Participation in extracurricular activities.

k) Sports: Involvement in sports.

l) Music: Participation in music-related activities.

m) Volunteering: Whether the student is involved in volunteering.

n) GPA: Grade Point Average (continuous variable).

o) Grade Class: Categorical class (A, B, C, D, F) derived from GPA.

```
RangeIndex: 2392 entries, 0 to 2391
Data columns (total 15 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   StudentID          2392 non-null   int64
 1   Age                2392 non-null   int64
 2   Gender             2392 non-null   int64
 3   Ethnicity          2392 non-null   int64
 4   ParentalEducation  2392 non-null   int64
 5   StudyTimeWeekly    2392 non-null   float64
 6   Absences           2392 non-null   int64
 7   Tutoring           2392 non-null   int64
 8   ParentalSupport    2392 non-null   int64
 9   Extracurricular    2392 non-null   int64
 10  Sports             2392 non-null   int64
 11  Music              2392 non-null   int64
 12  Volunteering       2392 non-null   int64
 13  GPA                2392 non-null   float64
 14  GradeClass         2392 non-null   float64
```

Figure 1: Data type for Student Academic Performance

Data Cleaning and Feature Engineering

a) Missing values were checked and found to be minimal; handled through imputation or removal.

b) Categorical variables were converted using label encoding and one-hot encoding.

c) Irrelevant columns such as StudentID were dropped.

d) Data normalization was applied using StandardScaler to scale numerical features for model training.

Descriptive Analysis of Student Attributes

	StudentID	Age	StudyTimeWeekly	Absences	Tutoring	Extracurricular	Sports	Music	Volunteering	GPA
count	2392.000000	2392.000000	2392.000000	2392.000000	2392.000000	2392.000000	2392.000000	2392.000000	2392.000000	2392.000000
mean	2196.500000	16.468645	9.771992	14.541388	0.301421	0.383361	0.303512	0.196906	0.157191	1.906186
std	690.655244	1.123798	5.652774	8.467417	0.458971	0.486307	0.459870	0.397744	0.364057	0.915156
min	1001.000000	15.000000	0.001057	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1598.750000	15.000000	5.043079	7.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.174803
50%	2196.500000	16.000000	9.705363	15.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.893393
75%	2794.250000	17.000000	14.408410	22.000000	1.000000	1.000000	1.000000	0.000000	0.000000	2.622216
max	3392.000000	18.000000	19.978094	29.000000	1.000000	1.000000	1.000000	1.000000	1.000000	4.000000

Figure 2: Descriptive Analysis of Student Attributes

An initial descriptive analysis was conducted on 2,392 student records to understand the distribution of key numerical variables related to demographics, behaviour, and academic performance. The average student age was 16.47 years, with the majority falling between 15 and 18 years. The range of weekly study time varied from close to zero to around 20 hours. The average was 9.77 hours. There was considerable variability in absenteeism as well. On an average, the students were absent 14.54 times. This indicates that absenteeism could be an important factor that impacts performance. There was low participation in support and enrichment activities. Around 30% of the students were tutored, 38% participated in extracurricular activities, 30% participated in sports, 19.7% were exposed to music, and merely 15.7% were exposed to volunteering activities. The GPA on a scale of 4.0 had an average of 1.91 and a standard deviation of 0.91. The median GPA was 1.89. The GPA of the students was concentrated at the lower end of the scale. The 75th percentile of the GPA was at 2.62.

Step 3: Data Visualization

The bar chart illustrates the distribution of students across different age groups (15–18 years) with their corresponding academic performance, categorized by grade class (A to F).

a) Age 15has the highest concentration of students receiving an “F” grade (over 300 students), indicating early high school may be a critical period for intervention.

b) Across all age groups, a majority of students fall into the F and D categories, suggesting persistent academic challenges across grades.

c) Grade “A” remains the least common across all ages, while grades “B” and “C” show modest and relatively stable representation.

d) Age 17 shows a slightly higher count of students achieving grade “D”, hinting at possible academic improvements during this phase.

e) The “Caucasian” group constitutes the largest ethnic category, and within this group, the most frequent grade is “F”, followed by “D” and “C”.

f) Students identified as “Asian” and “African American” also show a strong skew towards lower grades (“F” and “D”), though their total counts are significantly smaller.

g) The “Other” ethnic group is the least represented, and similar to other groups, the frequency of high academic achievers (Grade A) is minimal.

h) Across all ethnic groups, Grade A remains rare, indicating that ethnicity alone may not be a decisive factor for top academic performance but reflects broader systemic trends in underachievement.

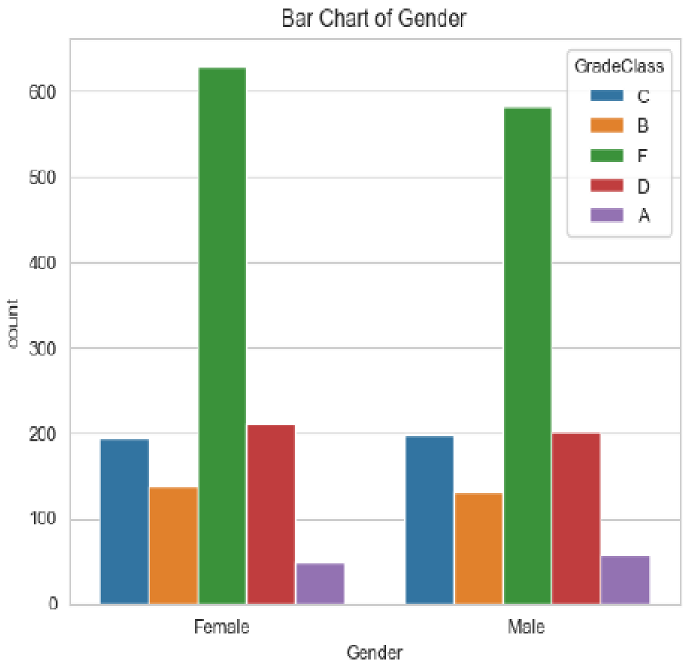
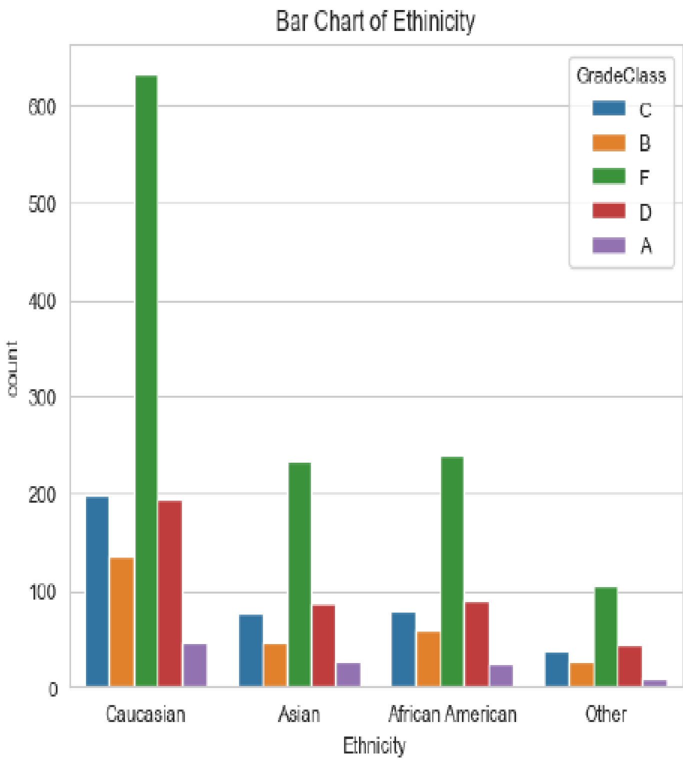


Figure 3.1: Bar Chart of Gender vs Grade Class Figure



3.2: Bar Chart of Ethnicity vs Grade Class

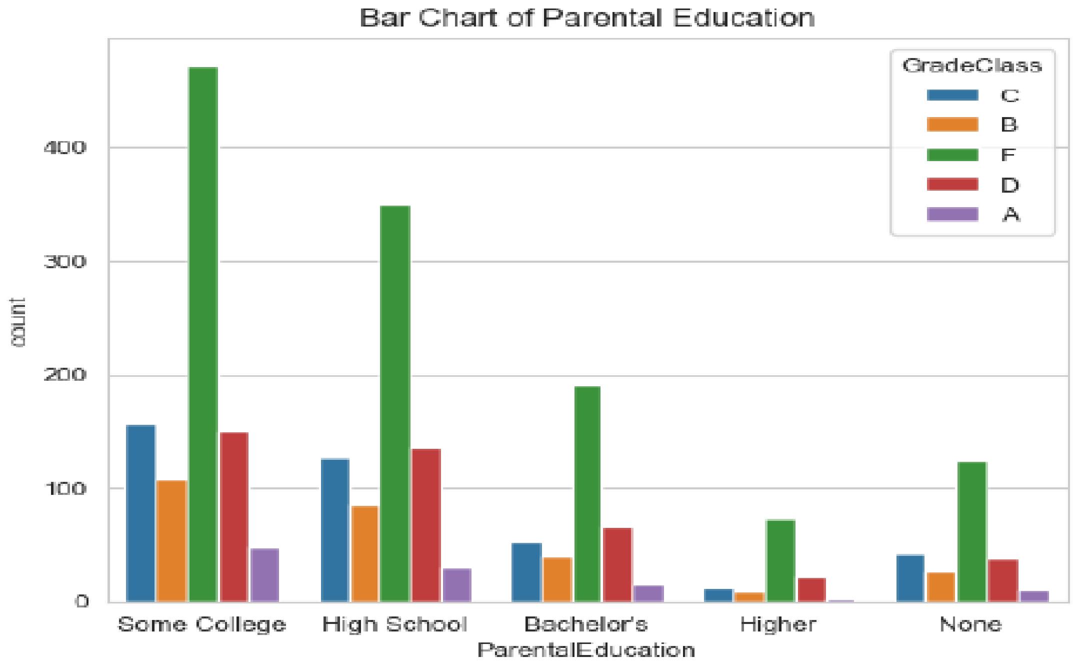


Figure 3.3: Bar Chart of Parental Education vs Grade

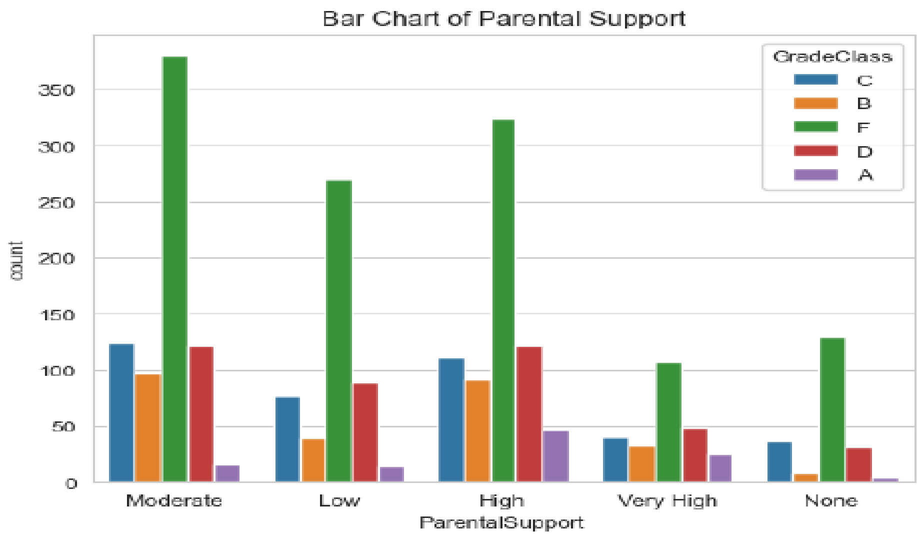


Figure 3.4: Bar Chart of Parental Support vs Grade Class

This bar chart illustrates the relationship between parental education levels and student academic performance, categorized by grade class (A to F).

a) Students whose parents have a level of education categorized as "Some College" and "Bachelors" dominate the figure, but a majority of these students also receive low grades, mainly “F” and “D”.

b) Students whose parents have a high level of education categorized as "Higher" dominate the figure for the highest grades, A and B.

c) Students whose parents have no education and high school education dominate the figure for low grades, mainly “F”.

d) From the chart, it is clear that there is a moderate and positive relationship between the education level of parents and the grades received by students.The bar chart shows the distribution of student grades, A to F, according to levels of parental support.

e) Students who receive moderate and high parental support dominate the figure for high grades, A, B, and C. There is a significant decline in the number of students who receive grade F.

f) Students who receive no parental support dominate the figure for low grades, mainly F and D.

g) The largest number of students who receive grade A is from the "Very High" support group, but this is a small fraction. These patterns indicate a positive association between increased parental support and improved academic outcomes, highlighting the role of family engagement in student success.

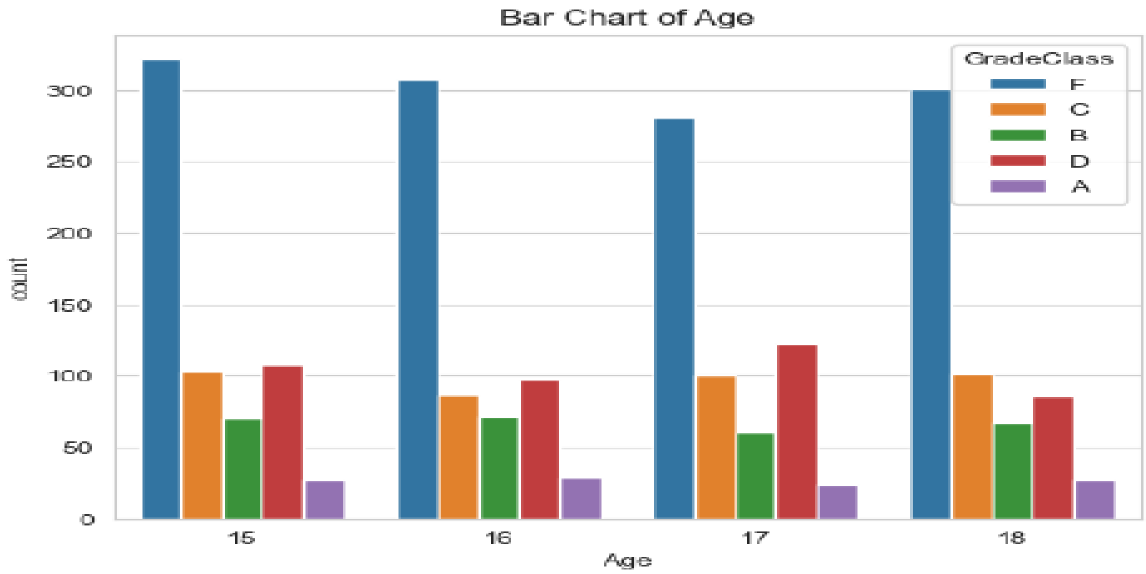


Figure 3.5: Bar Chart of Age vs Grade Class

This bar chart displays the distribution of student grades (A to F) across different age groups ranging from 15 to 18 years.

a) 15-year-old students represent the largest group and also record thehighest number of low grades, particularly Grade F.

b) Across all age groups, Grade F consistently dominates, indicating widespread academic challenges irrespective of age.

c) A slight improvement in performance is observed in older students (ages 17 and 18), with marginal increases in Grades A and B. However, high-performing students (Grade A) remain in a minority in all age brackets.

d) Overall, the chart suggests no strong age-based trend in academic performance but does indicate that academic underperformance is prevalent across the adolescent spectrum**.**

Figure 4.1: Scatter Plots of GPA vs Study Time Weekly with Grade Class

This scatter plot visualizes the relationship between students' GPA and their weekly study time, differentiated by grade class (A to F).

a) A positive trend is evident: students with higher GPAs (Grades A and B) generally report greater weekly study time, often exceeding 15 hours per week.

b) Conversely, students with lower grades (particularly Grade F) tend to cluster around lower GPA values and reduced study time, typically below 10 hours weekly.

c) The plot shows distinct separation in study behaviour among grade classes, indicating that consistent study time is a significant factor in academic success.

d) Overall, the visualization supports the hypothesis that greater academic engagement through study time is correlated with higher academic achievement.

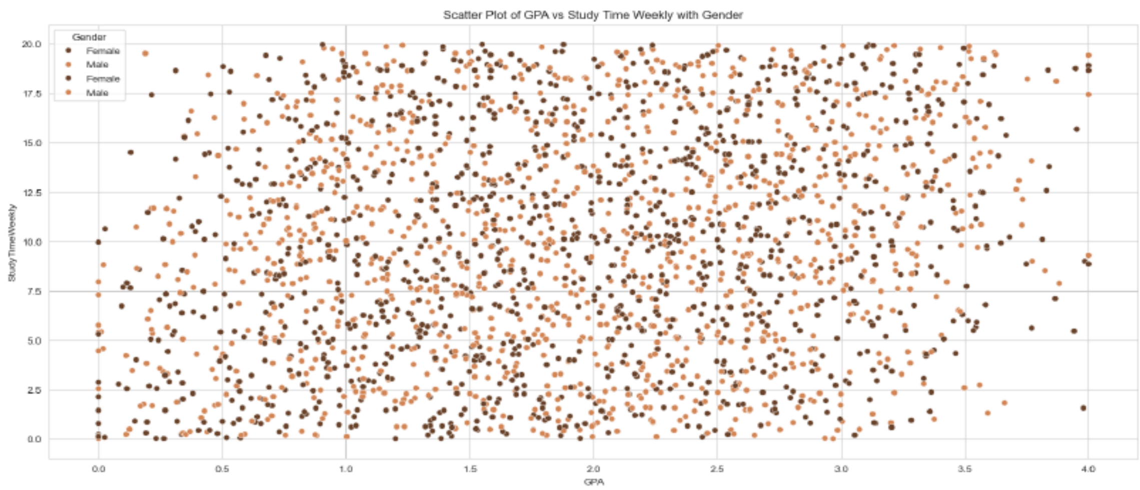


Figure 4.2: Scatter Plots of GPA vs Study Time Weekly with Gender

This scatter plot visualizes the relationship between students' GPA and their weekly study time, with data points color-coded by gender.

a) Both male and female students exhibit a wide spread in GPA and study time, indicating substantial individual variability.
b) A slight upward trend is observed among female students, suggesting that higher GPA values are often associated with increased study hours.
c) Male students appear more concentrated in the lower GPA and study time region, though many still span the entire range of performance.
d) No extreme clustering or outliers dominate either group, implying that gender alone does not dictate performance, but may interact subtly with study behaviour.

Overall, the chart suggests mild gender-based differences in study habits and performance, with females potentially showing a marginal edge in academic engagement.

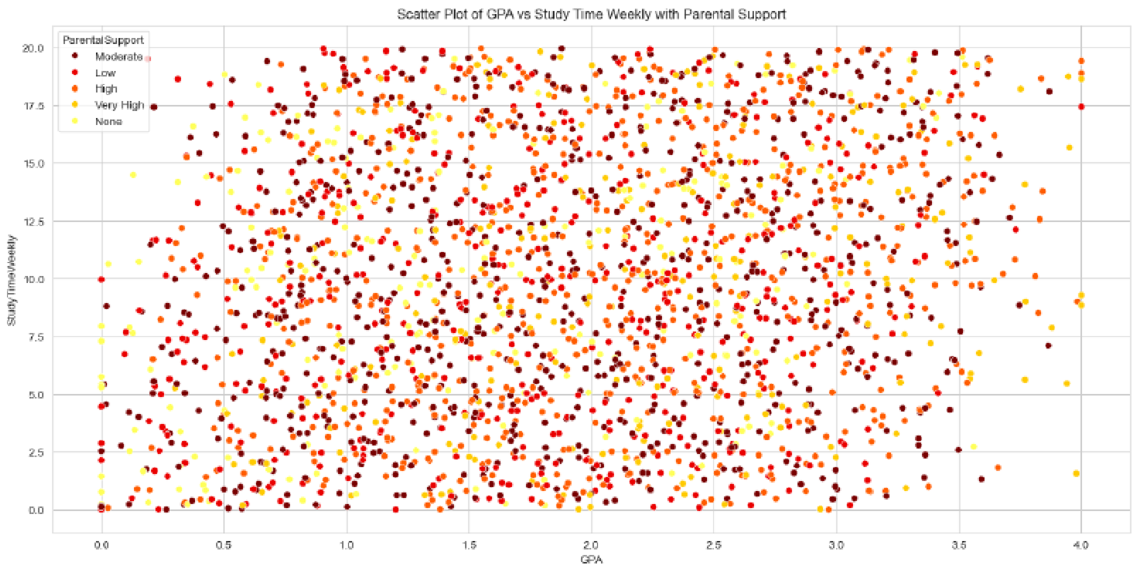


Figure 4.3: Scatter Plots of GPA vs Study Time Weekly with Parental Support

This scatter plot illustrates the relationship between students' GPA and their weekly study time, categorized by levels of parental support (ranging from 'None' to 'Very High').

a) Students who receive higher levels of parental support tend to demonstrate moderately higher GPA values and more consistent study time, often above 10 hours per week.
b) Those with low or no parental support are more concentrated in the lower GPA and low study time range, indicating a potential lack of academic reinforcement at home.
c) A clear gradient is visible, where higher support levels generally align with improved academic metrics, reinforcing the importance of family involvement.
d) However, overlaps across support categories suggest that while parental support is influential, it interacts with other factors in shaping student performance.
e) Overall, the plot underscores a positive association between parental support and academic engagement, particularly in terms of study commitment and GPA.

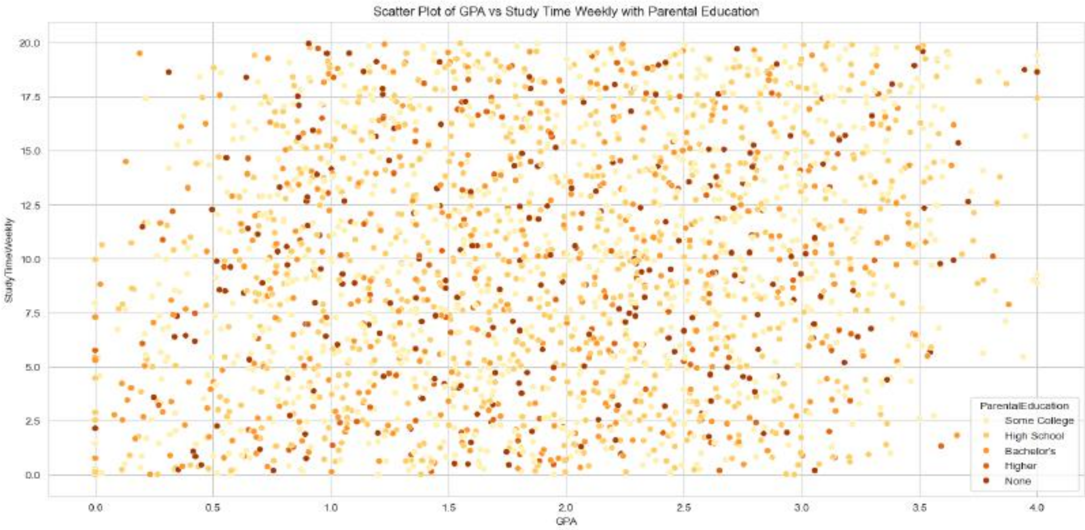


Figure 4.4: Scatter Plots of GPA vs Study Time Weekly with Parental Education

This scatter plot displays the relationship between students' GPA and their weekly study time, differentiated by their parents' highest level of education (ranging from "None" to "Higher").

a) Students whose parents hold higher educational qualifications (Bachelor's or Higher) are generally observed in higher GPA bands, often coupled with moderate to high study times.

b) Conversely, those from families with no formal education or only high school education are more concentrated in lower GPA and lower study time zones.

c) The dispersion pattern indicates that parental education positively correlates with academic performance and study commitment, though the relationship is not strictly linear.

d) Some overlap is present across all education levels, implying that while parental education influences academic habits and outcomes, it does not entirely determine them.

e) In summary, the plot suggests that higher parental education is associated with better academic outcomes and more disciplined study behaviours, contributing to improved student performance.

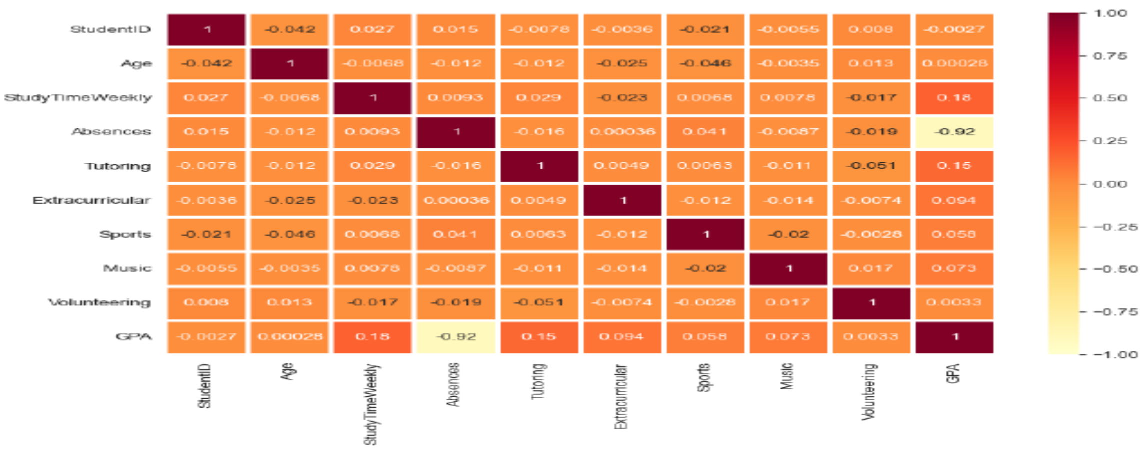


Figure 5: Heatmap of Feature Correlation (Heatmap of Correlation Matrix/)

This heatmap visualizes the pairwise Pearson correlations among the numeric variables in the student performance dataset, helping to identify the strength and direction of linear relationships.

a) The strongest positive correlation is observed between GPA and Grade Class, as expected, since both represent academic achievement in different formats.

b) StudyTimeWeekly exhibits a moderate positive correlation with GPA, indicating that students who dedicate more time to study tend to perform better.

c) Absences are negatively correlated with GPA, suggesting that higher absenteeism is associated with lower academic performance.

d) Tutoring, Sports, and Extracurricular activities show weak positive correlations with GPA, reflecting their marginal yet supportive roles in academic outcomes.

e) Most of the remaining features, including Music and Volunteering, have low or near-zero correlations, implying minimal direct linear influence on GPA.

Overall, the heatmap reveals that study time, attendance, and academic engagement factors play relatively stronger roles in predicting performance, offering key insights for feature selection in predictive modelling.

Random Forest Classifier (RFC): Random Forest is an ensemble machine learning algorithm that constructs multiple decision trees using bootstrap sampling and aggregates their predictions through majority voting. It reduces overfitting, handles mixed data types, and captures non-linear relationships among variables. Unlike single decision trees, Random Forest improves generalization and stability while also providing feature importance measures that enhance interpretability.

Model Evaluation and Performance Metrics

The performance of the trained Random Forest Classifier was evaluated on both training and testing datasets using standard classification metrics. The classification report reveals class-wise performance in terms of precision, recall, and F1-score:

a) Class 4.0 (Grade F), which is the largest class, achieved the highest F1-score (0.96), with precision and recall values above 94%.

b) Class 3.0 (Grade D) and Class 2.0 (Grade C) also showed high recall and F1-scores (~0.88–0.89), indicating strong model reliability for mid-performing students.

c) Class 1.0 (Grade B) achieved an F1-score of 0.79 with good recall (0.86), but slightly lower precision (0.74).

d) Class 0.0 (Grade A) had the lowest performance, with a high precision (1.00) but low recall (0.27), likely due to the small support (22 instances), making it harder for the model to learn minority class patterns. Theconfusion matrix further highlights that while the model classifies most instances correctly, it struggles with the minority class (Grade A), a common issue in imbalanced datasets.

Table 1: Model Evaluation Metrics (Random Forest Classifier)

Grade Class	Precision	Recall	F1-Score	Support
A (0.0)	1	0.27	0.43	22
B (1.0)	0.74	0.86	0.79	49
C (2.0)	0.91	0.86	0.88	85
D (3.0)	0.89	0.9	0.89	86
F (4.0)	0.94	0.99	0.96	237
Accuracy			0.9	479
Macro Average	0.89	0.77	0.79	479
Weighted Average	0.91	0.9	0.89	479

The macro average F1-score is 0.79**,** and the weighted average F1-score is 0.89**,** confirming the model's strong performance across all grade categories, especially for more populated classes.

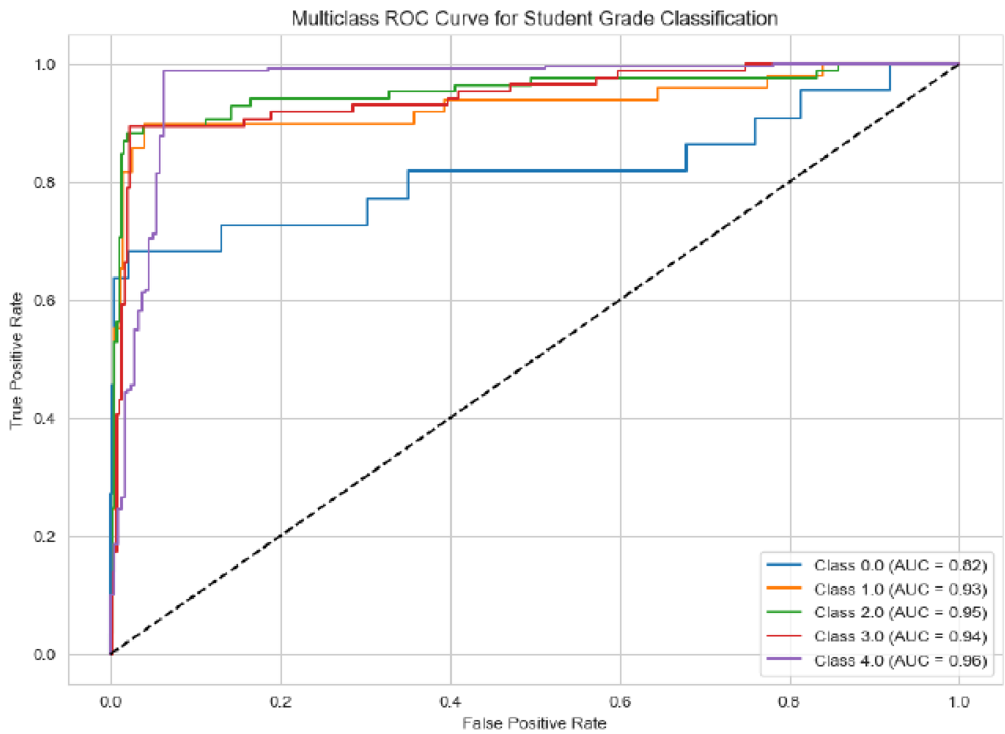


Figure 6: ROC Curve for Student Grade Classification

Explanation of Multiclass ROC Curve for Student Grade Classification

The graph shown is a Multiclass ROC (Receiver Operating Characteristic) Curve, used to evaluate the performance of the Random Forest model in predicting student academic performance across five grade categories (0.0 to 4.0), which typically represent grades from F (0.0) to A (4.0).

Key Observations:

a) X-axis: False Positive Rate (FPR) — the proportion of actual negatives incorrectly classified as positives.

b) Y-axis: True Positive Rate (TPR) — also known as recall or sensitivity, the proportion of actual positives correctly identified.

c) Diagonal dashed line: Represents a random classifier (AUC = 0.5). Any curve above this line indicates better-than-random performance.

d) Colored curves: Show ROC for each grade class, with Area Under the Curve (AUC) values as follows:

i) Class 0.0 (F): AUC = 0.82 – Good performance, but relatively lower than other classes.

ii) Class 1.0 (D): AUC = 0.93 – High discrimination ability.

iii) Class 2.0 (C): AUC = 0.95 – Excellent classification performance.

iv) Class 3.0 (B): AUC = 0.94 – High predictive capability.

v) Class 4.0 (A): AUC = 0.96 – Highest performance, indicating the model is most accurate at predicting students with top grades.

INTERPRETATION

The model demonstrates strong discriminative power across most grade classes, especially in higher performance categories (A, B, and C). The relatively lower AUC for Class 0.0 (F) suggests the model finds it more challenging to identify students at risk of failure, possibly due to class imbalance or feature overlap.

The findings of this study have several important policy implications for educational institutions and policymakers:

RESULTS, DISCUSSION & POLICY IMPLICATIONS

a) Early Warning Systems: Schools and universities should implement predictive analytics dashboards that flag students with high absenteeism and low study engagement for timely intervention.

b) Targeted Academic Support: Students identified as high-risk should receive personalized tutoring programs rather than generic remedial classes.

c) Parental Engagement Programs: Institutions should actively involve parents in student progress monitoring, particularly for students with low academic performance.

d) Attendance Monitoring Policies: Implementation of stricter attendance policies and academic counselling could lead to fewer dropouts and better academic performance.

e) Data-Driven Decision Making: Educational institution management should consider incorporating machine learning algorithms into their decision-making strategies.

CONCLUSION

The study has shown that machine learning, and specifically random forest classification, can be an effective tool in predicting students' academic performance and informing evidence-based practice. This study has made an important contribution to both theory and practice in that it has developed an integrated framework that incorporates all student characteristics. Future research should extend the model across institutions and longitudinal data.

LIMITATIONS AND FUTURE SCOPE

a) The dataset was limited to a single geographical or institutional setting; generalization needs to be verified using different datasets.

b) The potential for using deep learning techniques or time-series Modeling for academic progression could be a direction for further research.

c) The use of real-time management systems could be leveraged to automate trigger signals.

REFERENCES

1. Ahmed, A. B. E. D., & Elaraby, I. S. (2014). Data mining: A prediction for student's performance using classification method. World Journal of Computer Application and Technology, 2(2), 43–47. https://doi.org/10.13189/wjcat.2014.020203

2. Al-Barrak, & Al-Razgan, M. A. (2016). Predicting students' performance through classification: A case study. Journal of Theoretical and Applied Information Technology, 88(1), 1–7.

3. Baker, R. S. J., & Siemens, G. (2014). Educational data mining and learning analytics. In K. Sawyer (Ed.), Cambridge Handbook of the Learning Sciences (pp. 253–274). Cambridge University Press.

4. Breiman, L. (2001). Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

5. Credé, M., Roch, S. G., &Kieszczynka, U. M. (2010). Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics. Review of Educational Research, 80(2), 272–295. https://doi.org/10.3102/0034654310362998

6. Dimitriadis, Y. (2024). A review of machine learning methods used for educational data. Education and Information Technologies, 29, 22125–22145.

7. Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Computers & Education, 61, 133–145. https://doi.org/10.1016/j.compedu.2012.08.015

8. Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. Journal of Learning Analytics, 1(1), 6–47. https://doi.org/10.18608/jla.2014.11.3

9. Kuh, G. D., Kinzie, J., Buckley, J. A., Bridges, B. K., & Hayek, J. C. (2008). *What matters to student success: A review of the literature.* National Postsecondary Education Cooperative (NPEC). (Retrieved May 10, 2025, fromhttps://nces.ed.gov/npec/pdf/kuh_team_report.pdf)

10. Kotsiantis, S., Pierrakeas, C., &Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. Applied Artificial Intelligence: AAI, 18(5), 411–426. https://doi.org/10.1080/08839510490442058

11. Parack, N., Zahid, M., & Merchant, S. (2012). Application of data mining in predicting student performance. International Journal of Computer Science and Information Technology Research, 1(1), 22–28.

12. Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man and Cybernetics. Part C, Applications and Reviews: A Publication of the IEEE Systems, Man, and Cybernetics Society, 40(6), 601–618. https://doi.org/10.1109/tsmcc.2010.2053532

13. Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. Review of Educational Research, 75(3), 417–453. https://doi.org/10.3102/00346543075003417

14. Voyer, D., &Voyer, S. D. (2014). Gender differences in scholastic achievement: a meta-analysis. Psychological Bulletin, 140(4), 1174–1204. https://doi.org/10.1037/a0036620

15. Vandamme, J.-P., Meskens, N., &Superby, J.-F. (2007). Predicting academic performance and student dropout: A case study. European Journal of Operational Research, 181(2), 749–770.

16. Wagenaar, T. C., & Tinto, V. (1988). Leaving college: Rethinking the causes and cures of student attrition. Contemporary Sociology, 17(3), 414. https://doi.org/10.2307/2069700

17. Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9(1). https://doi.org/10.1186/s40561-022-00192-z