STUDENT EXAM PERFORMANCE ANALYSIS AND GRADE PREDICTION USING MACHINE LEARNING

A report by Toheebat Yewande Tiletile, Data Science Enthusiast

INTRODUCTION

This report presents a detailed analysis of student performance across multiple subjects, identifying key patterns and correlations that can inform educational strategies. The analysis leverages data on scores across subjects like Math, English, Computer Science, and more, offering insights into areas of strength and opportunities for educational improvement.

PROBLEM STATEMENT

A school manager has received multiple complaints regarding inconsistent student performance across different classes. Some classes consistently perform well, while others struggle, leading to concerns about overall academic quality and the effectiveness of teaching methods. The manager seeks a data-driven solution to assess the performance of each class, identify top-performing classes, and pinpoint those that require intervention. The goal is to understand the key factors contributing to these disparities and address them to improve the overall academic outcomes.

DATA OVERVIEW

The dataset includes variables like age, gender, class, subject scores, and attendance, which were used to predict the final grade of each student. It contains 12 variables which are:

- Demographic Data: Age, Gender, and Class.
- Performance Data: Scores in Math, English, Computer Science, Sport, Art, Attendance, Total Score, and Average Score.
- Grade: The final grade assigned to each student (A, B, C, D, F)

DATA COLLECTION

The dataset was sourced from an open educational dataset available on Kaggle.

DATA PREPROCESSING:

Data preprocessing was a critical step to ensure data quality and consistency before analysis and model training.

- The dataset was first inspected for duplicates and all duplicates found were dropped.
- Missing values were imputed using appropriate techniques
    - Median was used for Age to reduce the impact of outlier
    - Mean was used for scores for consistency
    - Mode was applied to gender and class to fill the most common occurrence.

➢ Data type: Gender and class were initially converted to categorical data for easier visualization and were further encoded to numerical values for correlation.

FEATURE ENGINEERING

➢ Total and average scores were calculated based on the student's performance across various subjects but were not stored as separate columns in the dataset..

➢ A grading function was defined based on average score and applied, classifying students into grades (A, B, C, D, or F) but was not stored as a separate column in the dataset.

➢ Attendance group was created and attendance was categorized into three groups, Low, Medium and High to analyze student's regularity in school.

DESCRIPTIVE STATISTICS

Age: The students range from 14 to 18 years old, with an average age of 16.

Scores:

➔ Math, English, Computer, Sport: These scores range from 0 to 100, indicating a wide range of performance.

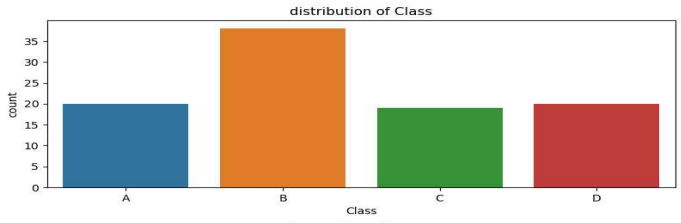
➔ Attendance Score: High average (82.41) with a narrow range from 50.75 to 99.93, suggesting most students have good attendance.

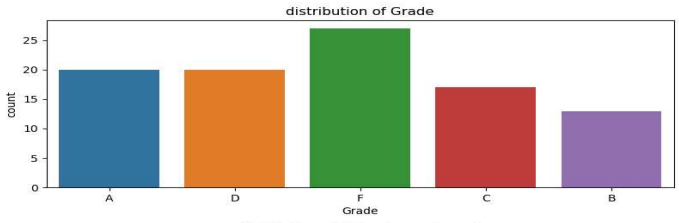
➔ Art Score: Varies significantly, with a mean of 49.6 and a wide range of 0 to 99.

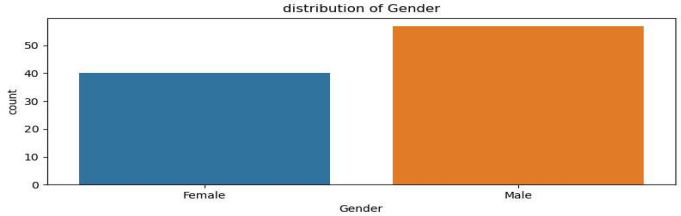
➔ Total and Average Scores: These also have a wide range, reflecting different student performance across subjects.

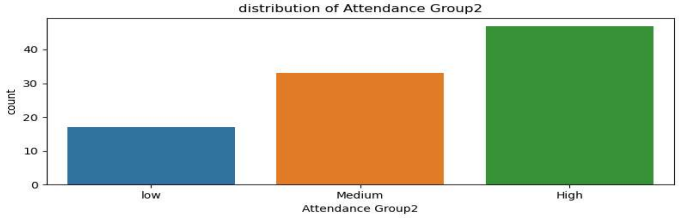
EXPLORATORY DATA ANALYSIS (EDA)

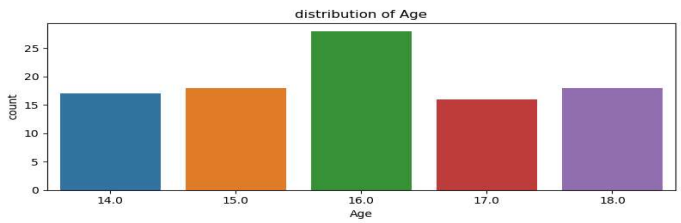
Distribution of Students











Class**:** The distribution of students across different classes is not even.Class B has the highest number of students, followed by class A, then class C and D that have almost the same number of students.

Gender**:** In this distribution, Male students outnumbered the female students.

Grade**:** F is the most common grade among students, followed by A and D**.** Only a few students got B and C.

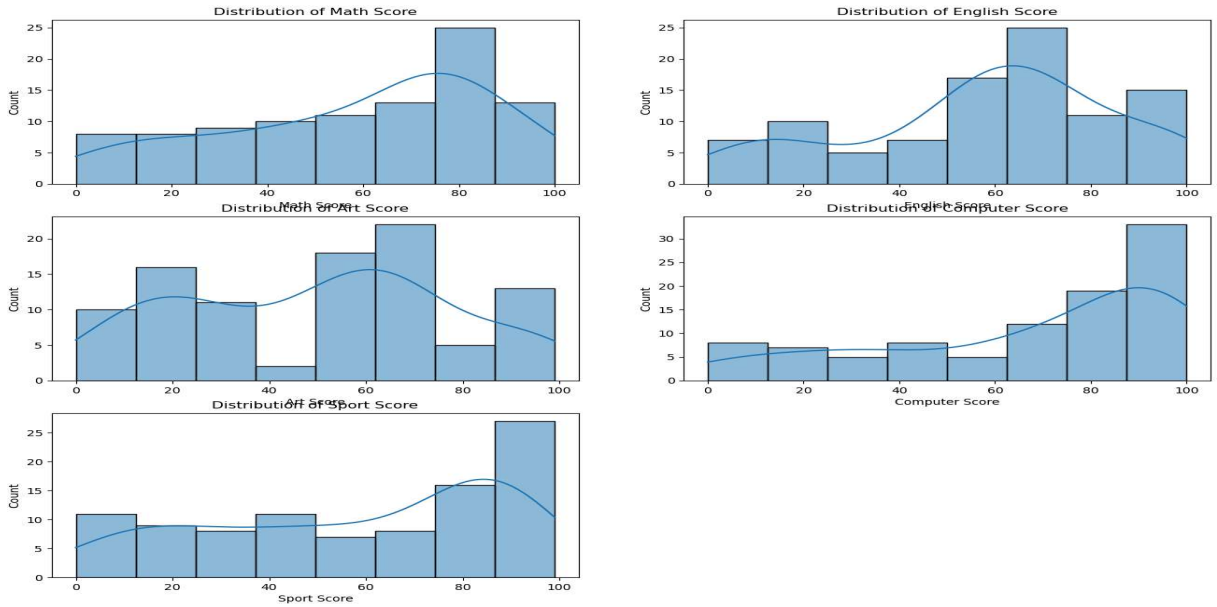
Attendance**:** Majority of the students have a high attendance, followed by medium and then low attendance as the smallest.

Age**:** Majority of students are 16 years of age followed by 14 and 15 who have almost equal distribution. Age 18 and 17 has the lowest distribution.

Insight: Class B has the highest number of students, which may require additional resources or attention to manage effectively. The high frequency of F grades suggests that a significant number of students are struggling academically, which could be related to several factors including attendance, student intellectual ability, among others. Gender imbalance could be a factor in classroom participation and overall student experience. Attendance appears to be generally good, with most students attending classes regularly but students in the low attendance group could be at a risk for lower academic performance.

Distribution of Scores

Math Score**:** The distribution shows that the majority of students score between 60 and 80 with a significant number scoring close to 80.



English Score**:** The distribution shows peak around 60 **-** 70 with few student scores below 40.

Art Score**:** The scores appear bimodal with peaks around 20**-**30 and 50**-**70.

Computer Score**:** The distribution is skewed towards the end, with the majority of students scoring between 80 and 100.

Sport Score**:** This distribution is also somewhat skewed to the higher end, with many students scoring near 100 which is similar to computer scores.

Generally**:** Most students perform well in Computer and Sport, there's balanced performance in English and Math and variability in Art that is, while some students are performing well others performed below average.

Insight: There's a significant number of students scoring below average in each subject as well as a few number of students scoring a higher mark. Attention should be paid to students scoring below average to develop their foundational skills and engage them with interactive and different learning approaches such as practical experiences to suit their learning styles. Likewise, learning activities should be enriched to enhance the performing student skills to continue performing better and score higher.

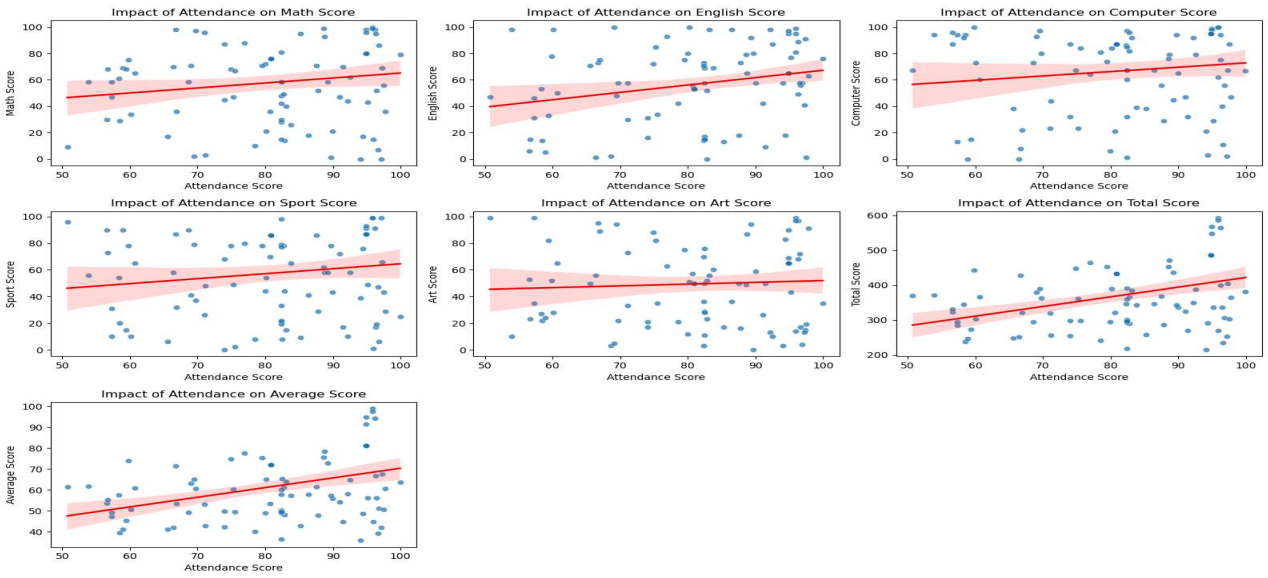
Impact of Attendance nn Subject Scores

Math Score: There's a moderate positive correlation between attendance and math scores. The regression line suggests that as attendance increases, math scores tend to increase.

English Score**:** There's also a moderate positive correlation similar to Math. Students with higher attendance tend to score better in English.

Computer Score**:** A weak positive correlation with attendance is shown. There's a slight upward trend in the data.

Sport Score: The trend line shows a higher sport score with increased attendance but the relationship is not very strong. There's a weak positive correlation between Sport score and attendance.



Art Score**:** Art score and attendance has a very weak positive correlation. The relationship between attendance and art score is minimal, with the regression line almost flat.

Total and Average Score**:** There's a strong positive correlation with attendance. The positive trend indicates that

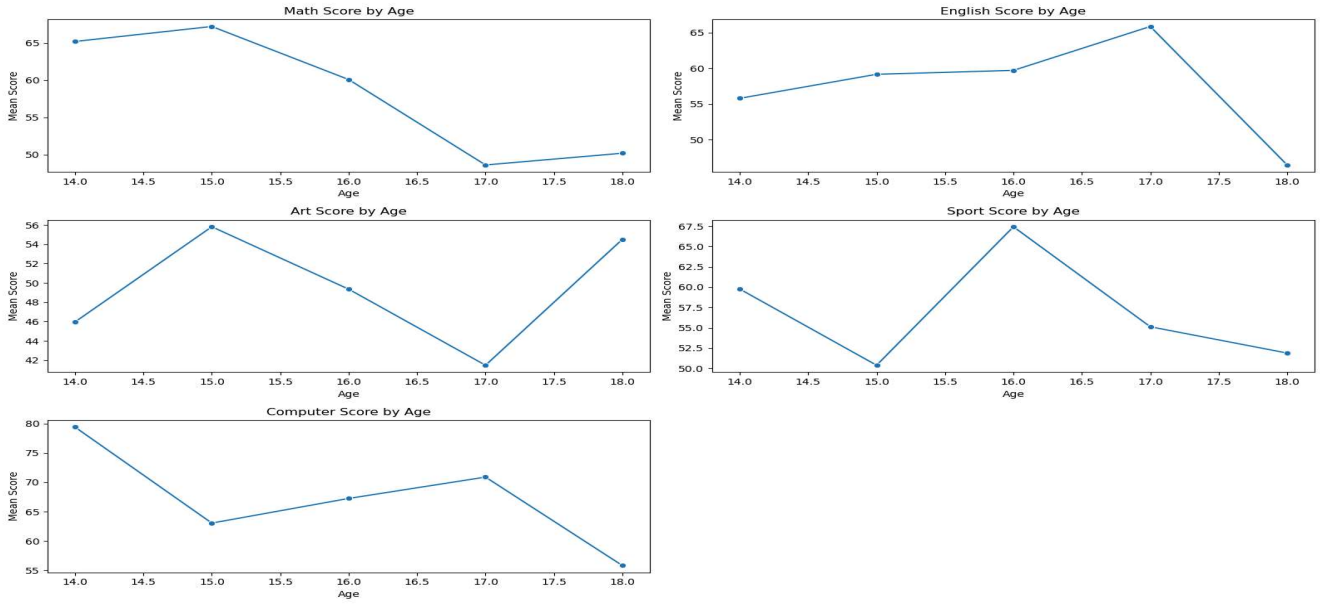
students with higher attendance tend to have a high performance.

Generally**:** There's a positive correlation between attendance scores and all academic performance metrics. As attendance increases, there's a trend of higher scores across all subjects, overall total and averages. There's a considerable scatter in all plots indicating that attendance is not the only factor that determines academic performance. There are also obvious outliers in most subjects where some students with high attendance have very low scores.

Insight: Attendance appears to be an important factor in academic performance but it doesn't guarantee a high score. The impact of attendance varies by subject, like art subject, there may be other factors more dominant than attendance. Other factors like learning environment, learning infrastructures, gender, study habits, individual aptitude, student's family background, teaching quality,etc. may likely play significant roles alongside attendance. Encouraging higher attendance may lead to improved outcomes but personal learning support might be necessary for students who attend regularly but still struggle academically.

Impact of Age on Scores

Math Score: The math score decreases gradually after peaking at age 15. The sharpest decline occurs after age 16, where it drops to the lowest at age 17.



English Score**:** English scores increase steadily, peaking at age 17 then drops drastically 18. This may be due to increased language proficiency and practice.

Art Score: Art scores vary significantly but there's a peak at age 15 then a drop at 17 followed by a sharp increase at age 18.

Sport Score: This score shows an irregular pattern with highest scores observed at 16 with a decline at age 17.

Computer Score: Scores start high at age 14, then drop at age 15 which remains stable before a slight increase 17.

Generally: Most subjects show a performance peak between age 15 - 17 after which there's a general decline. Subjects like Math, Sport and Computer scores show a declining trend as students grow older, with obvious drops at 17 or 18. English scores improve at age 17 while Art scores fluctuates

Insight: Across all subjects, age 17 seems to be a critical point where performance drops at various subjects except English.. This could suggest that older students may prioritize specific subject areas, face increased academic pressure or distractions that deter them from performing excellently.

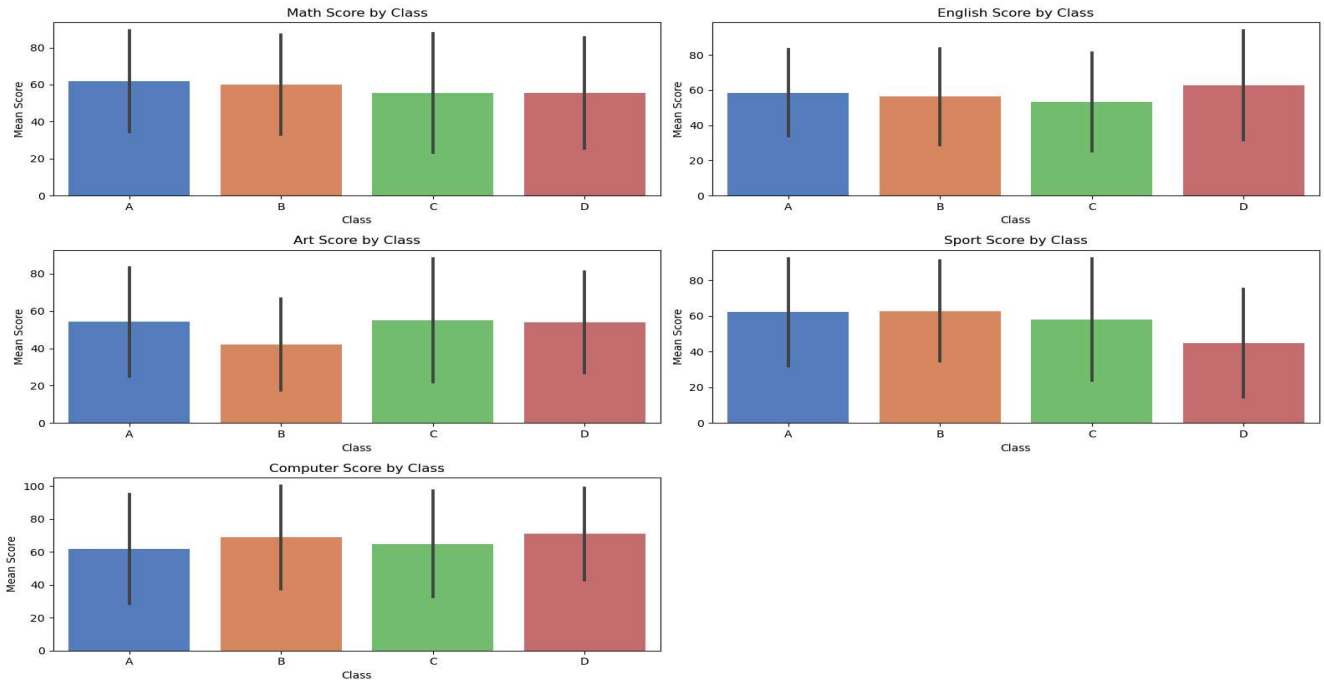
Class - Based Performance

Math**:** Class A has the highest mean score, while class C and D show lower averages.

English**:** Class D has the highest mean score in English while A, B and C have lower average scores with C has the least score.

Art**:** The mean scores for class B are lower while A, C and D have similar mean scores.

Sport**:** Class A and B have similar mean scores, while C is slightly higher than D which has a lower average.



Computer**:** The mean scores shows class B and D slightly outperforming

Classes A and C.

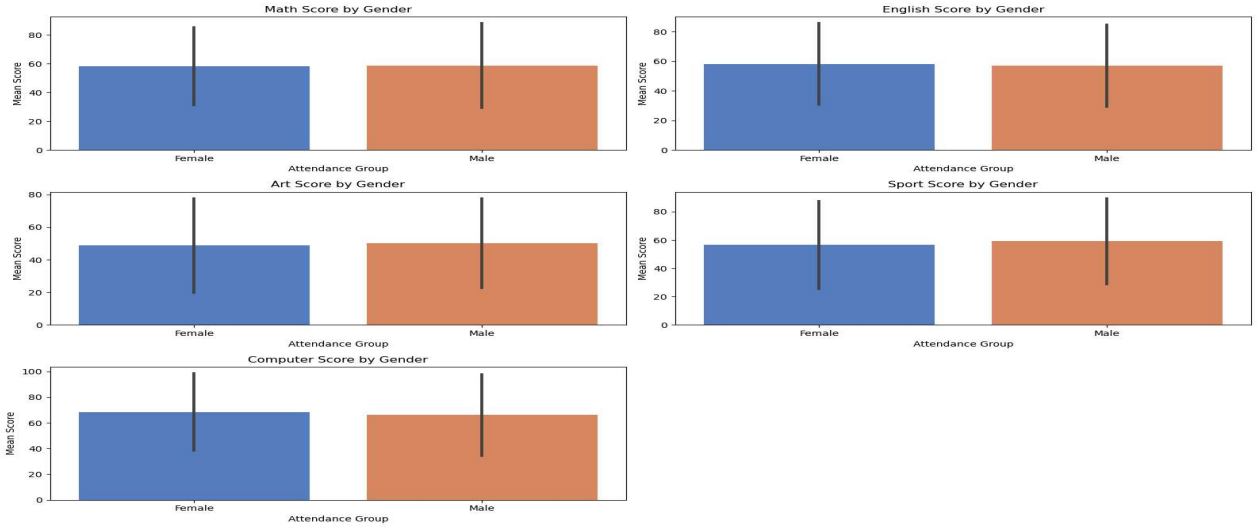
Generally**:** Across all subjects and classes, the error bar shows variability in student performance. This means that while the mean scores are close between classes, there are significant differences between individual students within each class. Despite the variability, the mean scores are quite consistent.

Insight: There are revealing differences between individual students in each class. The range of performance within classes may be due to different levels of understanding or external factors like the economic status of

parents affecting each student. Simply put, class performance is not the primary driver of performance. A class with a high mean score doesn't necessarily indicate every student in that class performs excellently.

Gender Based Performance

General: The mean score for both genders are relatively small and have almost the same score across all subjects.



Insight: The overall similarity in scores might indicate success is providing equal educational opportunities for both genders.

Correlation Analysis

- ➢ Subject Correlations with Total/Average Score:

Math has the highest correlation (0.64) with the total/average score, suggesting it's the strongest predictor of overall performance.

English (0.54) and Computer (0.62) scores also show strong correlations with the total/average score.

Sport (0.62) has a surprisingly high correlation, indicating athletic performance might be linked to overall academic achievement.

Art (0.58) and Science (0.40) have moderate correlations with the total score.

Attendance (0.40) also has a moderate correlation with total score, which shows it’s not the sole determinant of academic performance.

- ➢ Inter-subject Correlations:

The inter-subject subject correlation shows a weak to very weak positive relationship among each other. Although they’re positive, they have no significant impact on one another. Surprisingly, a high score in math doesn’t

guarantee the student will perform excellently well in computer, same as English with Art. All subjects are independent of each other. Though math is an essential component of computer science, it depends on the depth of the curriculum and levels of education.

[26]:

	Total Score	Average Score	gender	class	Math Score	English Score	Computer Score	Sport Score	Art Score	Attendance Score	Age
Total Score	1.000000	1.000000	-0.006111	-0.044172	0.635720	0.538619	0.618459	0.618139	0.577855	0.398643	-0.159463
Average Score	1.000000	1.000000	-0.006111	-0.044172	0.635720	0.538619	0.618459	0.618139	0.577855	0.398643	-0.159463
gender	-0.006111	-0.006111	1.000000	0.062550	0.008693	-0.020026	-0.038473	0.042288	0.025655	-0.081129	0.046910
class	-0.044172	-0.044172	0.062550	1.000000	-0.084538	0.039358	0.071280	-0.201418	0.068801	-0.055698	0.022375
Math Score	0.635720	0.635720	0.008693	-0.084538	1.000000	0.144302	0.283556	0.248685	0.273126	0.176533	-0.229216
English Score	0.538619	0.538619	-0.020026	0.039358	0.144302	1.000000	0.287303	0.032308	0.164962	0.269160	-0.066090
Computer Score	0.618459	0.618459	-0.038473	0.071280	0.283556	0.287303	1.000000	0.220482	0.057121	0.142259	-0.172675
Sport Score	0.618139	0.618139	0.042288	-0.201418	0.248685	0.032308	0.220482	1.000000	0.344434	0.161934	-0.047826
Art Score	0.577855	0.577855	0.025655	0.068801	0.273126	0.164962	0.057121	0.344434	1.000000	0.062685	0.016319
Attendance Score	0.398643	0.398643	-0.081129	-0.055698	0.176533	0.269160	0.142259	0.161934	0.062685	1.000000	-0.003929
Age	-0.159463	-0.159463	0.046910	0.022375	-0.229216	-0.066090	-0.172675	-0.047826	0.016319	-0.003929	1.000000

- ➢ Demographic Correlation

Gender has very weak correlations with all other variables values close to 0 while age has weak negative correlations with most subject scores and Class shows weak positive correlations with most subject scores. Notwithstanding, gender, age and class are factors that determines academic achievement but they are not the only factors and doesn't work independently but with other internal and external factors that might affect academic performance like low entry grades, family support, accommodation, previous grade in assessments, students' internal assessment grades, GPA, and students' e-Learning activity (Al Husaini, Y., & Shukor, N. S. A, 2023).

Insight: The strong correlation between Math and total score suggests that Math performance might be a good predictor of overall academic performance. The near-zero correlations with gender suggest that educational outcomes are relatively gender neutral in this dataset. The positive correlations between different subject scores indicate that students who perform well in one area tend to do well in others. The slight negative correlation between age and scores within classes might warrant investigation to determine what led to that outcome. It could point to issues with late school entry. While there are correlations between subjects, none are extremely high, indicating that each subject brings unique value to overall education.The positive correlations between class and scores suggest that the education system is generally effective in improving student performance as they progress.

PREDICTIVE MODELING: RANDOM FOREST CLASSIFIER

Model Setup: A Random Forest Classifier was used to predict student grades based on demographic and performance data. The following steps were taken:

- ➢ Target Variable: Final Grade (A, B, C, D, F).
- ➢ Features: Total Score, Average Score, Attendance, Age, Gender, and Class.

- ➢ Train-Test Split: The data was split into 70% training and 30% testing sets to evaluate model performance.
- ➢ Result**:** The model was trained and tested on the dataset, predicting the final grades for students in the test set.

Model Evaluation: The model achieved an accuracy of **97%** on the test set indicating it correctly predicted the final grade for 97% of the students in the test data.The precision for Grade A was 80% suggesting some students predicted to receive an A may have received a lower grade. The lower recall score for Grade B (80%) suggests that the model missed some students who should have been classified as B.

```
#Evaluate model's performnace with classification report
print (classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
A	0.80	1.00	0.89	4
B	1.00	0.80	0.89	5
C	1.00	1.00	1.00	4
D	1.00	1.00	1.00	6
F	1.00	1.00	1.00	11
accuracy			0.97	30
macro avg	0.96	0.96	0.96	30
weighted avg	0.97	0.97	0.97	30

CONCLUSION AND DISCUSSION

The analysis of student exam performance highlights the crucial role of math scores and attendance in determining overall academic success. Math proficiency and regular attendance are strong predictors of higher grades, suggesting that focusing on these areas can lead to significant improvements. The use of technology is a useful way to engage students because its incorporation can create fun and enhance their chances of being successful. Breaking down problems into smaller components, applying real-world examples, and one-on-one instruction are fundamental strategies to help students with their learning (Serin, H., 2023). Similarly, to improve attendance, strategies like support-based interventions, such as parental engagement, improving transportation, school bonding, and incentive-based strategies, are more effective than punitive approaches(Klein, M., & Sosu, E. M., 2023).

While math was identified as a key predictor, it’s essential to recognize that other subjects, such as English and computer studies, also play a vital role in a student's education. Balanced support across all subjects is important for overall student development. The predictive model used in this analysis provides valuable insights into which students may need additional support. By identifying these students early, schools can tailor interventions and allocate resources more effectively to address specific needs.

PROPOSED SOLUTIONS

- ➢ Class A should focus on improving English performance, as it lags behind their Math scores, D, on the other hand, excels in English but needs to improve their Math scores., B has a higher number of students and a varied performance. It's important to assess this class individually and provide personalized

interventions to close performance gaps.

- ➢ Students with lower attendance tend to perform worse, initiatives to increase regular attendance should be introduced.
- ➢ Students showing signs of struggle, particularly in subjects like Math and Computer Science, should have personalized intervention plans. These can include one-on-one tutoring, etc.
- ➢ Programs that will increase the engagement and motivation of the older student, such as career counseling, workshops on setting goals, and mentorship opportunities should be implemented.

FURTHER ANALYSIS

- ➢ Looking into other factors like parental involvement, extracurricular activities, and resources could give a clear and detailed picture of enhance academic performance
- ➢ The current dataset is small, so findings may not apply to all students. More data would improve the accuracy of predictions and recommendations.

TOOLS AND TECHNOLOGIES USED

- ➢ Python: Used for writing and executing code for data analysis and modeling.
- ➢ Pandas and NumPy: For data manipulation and analysis
- ➢ Matplotlib & Seaborn: Used for data visualization
- ➢ Scikit-Learn: For building and evaluating machine learning models, including the Random Forest Classifier used in this analysis.
- ➢ Jupyter Notebook: Used for interactive coding, visualizations, and documentation.

REFERENCES

1. Klein, M., & Sosu, E. M. (2023). School Attendance and Academic Achievement: Understanding Variation across Family Socioeconomic Status. International Journal of Social Sciences & Educational Studies, 97(1). https://doi.org/10.1177/00380407231191541

2. Serin, H. (2023). Teaching Mathematics: Strategies for Improved Mathematical Performance. Retrieved from https://www.researchgate.net/publication/367360842

3. Al Husaini, Y., & Shukor, N. S. A. (2023). Factors Affecting Students' Academic Performance: A Review Article. Arab Open University - Oman and UNISEL | Universiti Selangor.

4. Kaggle. (2024). Student Exam Dataset with Issues. Retrieved from https://www.kaggle.com/datasets/dinachanthan/student-exam-dataset-with-issues