1 Introduction

Predicting grades from the given dataset is the primary goal of this project. Data was taken from Moodle. Some data visualization was performed to understand data. I employed two different predicting models, Linear Regression and Random Forest. In addition to this, I served preprocessing and compared the accuracy results before and after preprocessing.

2 Data processing

2.1 Data information: There were 48 columns/features and 107 rows. And I am considering this dataset minimal for AI models.

2.2 Missing value handling: From the data info, we understood that there was no missing value. So, I did not need to fill in or remove any data.

2.3 Outlier detection: I found left-skewed diagrams using different distribution graphs, but I did not modify or remove them as I already mentioned that the dataset is minimal. As data will lose its character. In the visualization part, we will discuss more about it.

2.4 Feature selection: As there are 45 different columns and I wanted to reduce the calculation complexity. I used a correlation matrix with the target variable. I set the threshold value to 0.4. I only considered the features with absolute correlation values above this threshold value. After allowing this condition, I have got 25 columns.

3 Data analysis with visualization

I performed multiple visualizations to find some patterns in the dataset.

3.1 Heatmap of the correlation (Fig 1):

Positive correlation: Features with a dark red color suggest a positive correlation with each other. When one feature increases, the other feature tends to increase.

Negative correlation: Features that show a dark blue color suggest a negative correlation with each other. When one feature increases, the other feature tends to decrease.

No or weak correlation: Features showing a light color suggest they have no or weak correlation.

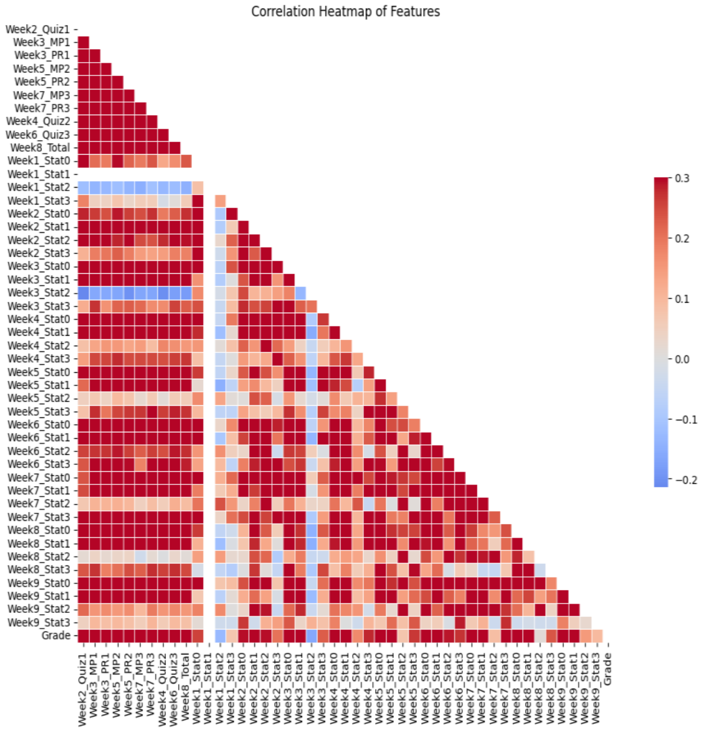

Fig 1: Heatmap of the correlation

3.2 Distribution graph: First, I will show the distribution of the target variable (Grade). Moreover, I chose the top three correlated columns with the target variable for distribution among 48 columns. They are Week8_Total, Week7_MP3 and Week5_MP2.

3.2.1 Distribution of Grade (Fig 2): The "Grade" distribution is left-skewed, and we can understand most of the students are getting good grades between 3 and 5 (Out of 5).

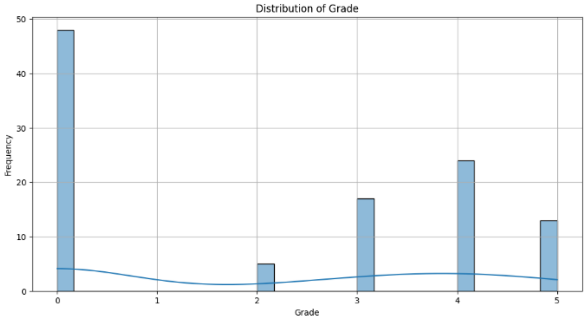

Fig 2: Distribution of Grade

3.2.2 Distribution of Week8_Total (Fig 3): This distribution shows left skewed. And most of the students have got good numbers between 60 and 100 (Out of 100).

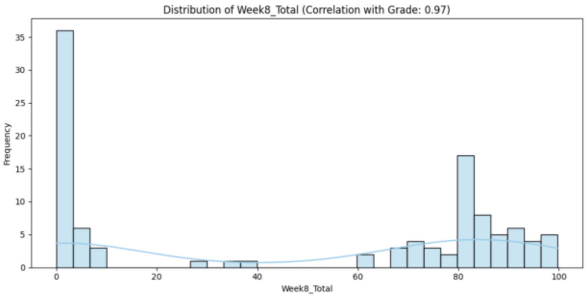

Fig 3: Distribution of Week8_Total

3.2.3 Distribution of Week7_MP3 (Fig 4): This distribution also shows left skewed. And most of the students have got good numbers between 25 and 35 (Out of 35).

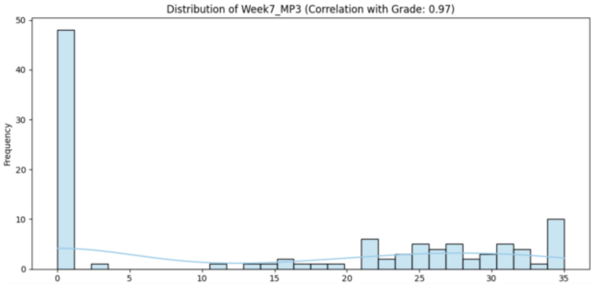

Fig 4: Distribution of Week7_MP3

3.2.4 Distribution of Week5_MP2 (Fig 5): This distribution also shows left skewed. And most of the students have got good numbers between 15 and 20 (Out of 20).

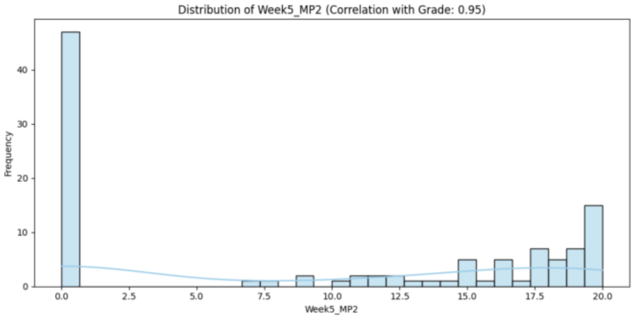

Fig 5: Distribution of Week5_MP2

From the distribution graphs we understood that a significant number of students (Around 48) did not continue the course and got Garde 0.

3.3 Scatter plot

I decided to show scatter plots between grade and the top correlated features.

3.3.1 Week8_Total vs. Grade (Fig 6): The below graph displays a strong linear correlation, meaning as "Week8_Total" increases, "Grade" tends to increase as well.

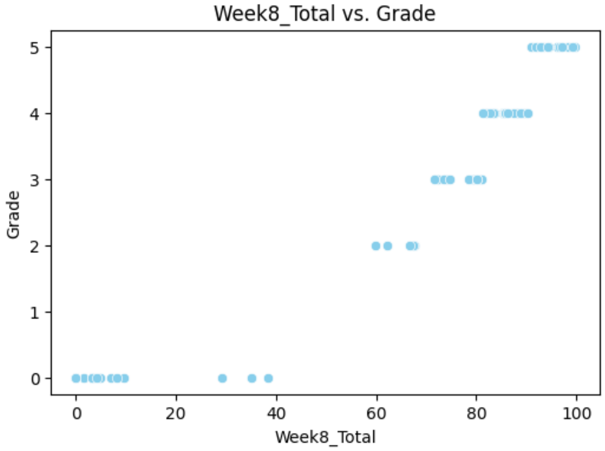

Fig 6: Week8_Total vs. Grade

3.3.2 Week7_MP3 vs. Grade (Fig 7): It also shows a strong linear relationship with "Grade," similar to "Week8_Total".

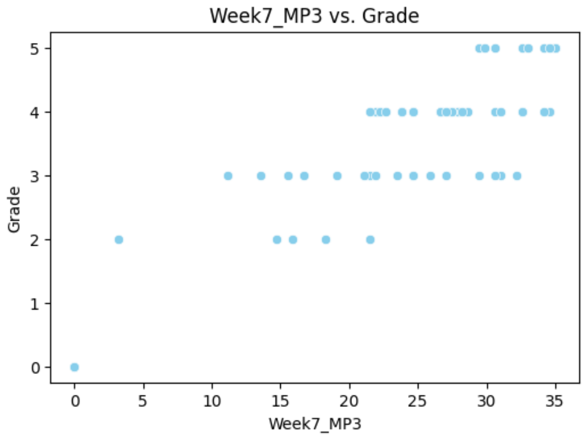

Fig 7: Week7_MP3 vs. Grade

3.3.3 Week5_MP2 vs. Grade (Fig 8): The below scatter plot exhibits a linear relationship but slightly more variability than the other two variables.

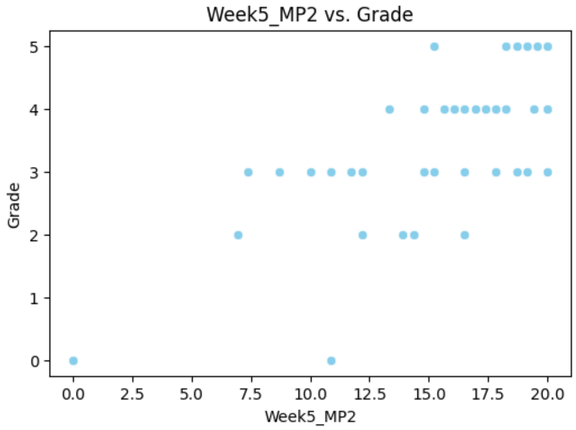

Fig 8: Week5_MP2 vs. Grade

4 Result comparison of two different models

4.1 Dataset: I used two different datasets for this analysis.

1. **Original dataset:** Here, I considered all the given dataset's features (48 features).
2. **Filtered dataset:** After feature selection, I got this filtered dataset (25 features), which I already discussed in Section 2.4.

4.2 Employed model: I used two different models for this analysis.

1. **Linear regression:** In machine learning and statistics, linear regression is a statistical technique used to describe the relationship between one or more independent variables (typically represented by the letter "x") and a dependent variable (commonly defined by the letter "y"). To help us comprehend the relationship between variables and make predictions, it seeks to identify the linear equation that best fits the data.
2. **Random Forest Regression:** An ensemble of decision trees is used in Random Forest Regression, a machine learning technique, to generate predictions for regression tasks. It integrates predictions from several trees, each trained on a different subset of data and attributes to increase accuracy and decrease overfitting. Because of its adaptability and strength, it can be used to solve a variety of regression issues.

4.3 Model accuracy result

4.3.1 Linear Regression on the original dataset (Fig 9): First I applied Linear Regression on the original dataset and found the accuracy R^2 of 78%. Below comparison line chart is showing the differences between actual and predicted data.

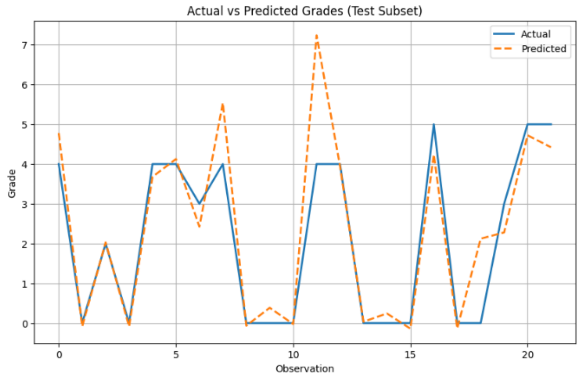

Fig 9: Actual vs Predicted grades.

4.3.2 Linear Regression on the filtered dataset (Fig 10): Again, I applied Linear Regression on the filtered dataset and interestingly, found an improved accuracy R^2 of 84.6%. The below comparison line chart shows the differences between actual and predicted data.

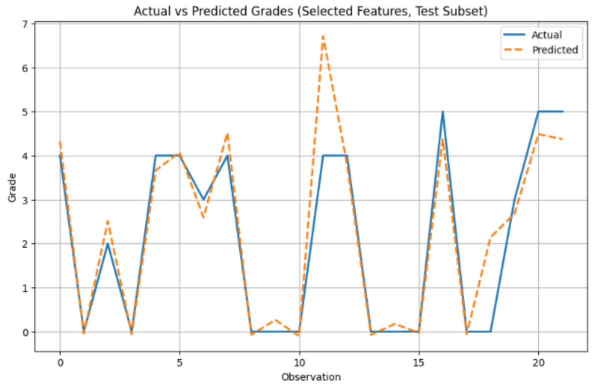

Fig 10: Actual vs Predicted grades.

4.3.3 Random Forest Regression on the original dataset (Fig 11): Then, I applied Random Forest Regression on the original dataset and found an excellent accuracy R^2 of 98.8%. The comparison line chart below shows the differences between actual and predicted data.

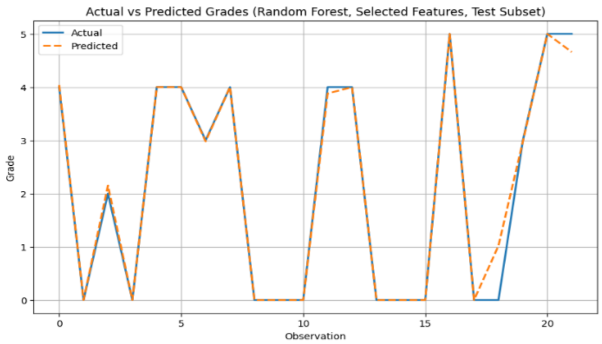

Fig 11: Actual vs Predicted grades.

4.3.4 Random Forest Regression on the filtered dataset (Fig 12): I applied Random Forest Regression on the filtered dataset and found a slightly decreased accuracy R^2 of 98.7%. The below comparison line chart is showing the differences between actual and predicted data.

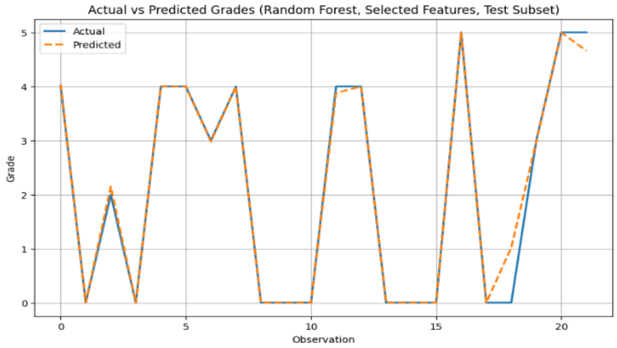

Fig 12: Actual vs Predicted grades.

A comparative result table is given below:

Dataset type	**The accuracy (as measured by the coefficient of determination R^2)**	
	Linear Regression	**Random Forest**
Original Dataset	78%	98.8%
Correlation-Based Filtered dataset	84.6%	98.7%

From the above observation, it is clear that the Random Forest algorithm outperforms Linear Regression in every cases. However, filtered data improved the Linear Algorithm's accuracy.

5 Important features

I found different features are liable to predict the grade for two different models. We can find the top three important features using the coefficient score in Linear Regression. Meanwhile, the Random Forest algorithm has its feature importance module to identify the top three features to predict the grade.

A feature importance table for different models is given below:

Priority	**Linear regression**		**Random Forest**	
	Col Name	**Coeff**	**Col Name**	**Importance**
1st	Week3_PR1	-0.45	Week8_Total	0.407530
2nd	Week3_MP1	0.138885	Week5_MP2	0.259435
3rd	Week7_PR3	0.093690	Week7_MP3	0.197750

6 Conclusion

The main aim of this project is to predict the students' grades. First, I visualize some relationships of different features with the target variable. Afterward, I did some feature engineering and considered the most correlated features (25 out of 48), then applied the linear regression and Random Forest before and after the feature extraction process. And found that the **Random Forest** algorithm performed more accurately even without the feature extraction process. Along this journey, I found some bottlenecks.

Minimal dataset: A small dataset can not give appropriate insights, and the model can perform worse against unseen data.

Unable to data manipulation: There were only 107 data rows. If I manipulate this small dataset, it might lead to lose its character.

Overfitting issue: I also tried to do another feature extraction process using correlation weight but found 100% accuracy in linear regression. So, I had to discard that process. Once again, this was the subsequence of a very small dataset.

However, while doing this project, I learned some data science concepts like data visualization, feature selection method, Machine learning models, and their applications.