

Introduction-

Understanding what influences academic success is key to improving learning, creating personalized education, and making better interventions. Traditional ways of evaluating performance, like focusing only on test scores and grades, often miss important aspects of a student's abilities and challenges.

This project uses machine learning to build a model that can predict student performance based on different factors, such as Study time (Weekly), Absences, attendance, Extracurricular activities, and behaviour. By analysing this information, the model can offer useful insights to help teachers as well as students to focus more on weaker part and adjust teaching methods to improve results

Flow chart-



Importing Data-

Importing and reading the csv file using pandas library.

Dataset used: [Student Performance Prediction](https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset?resource=download)

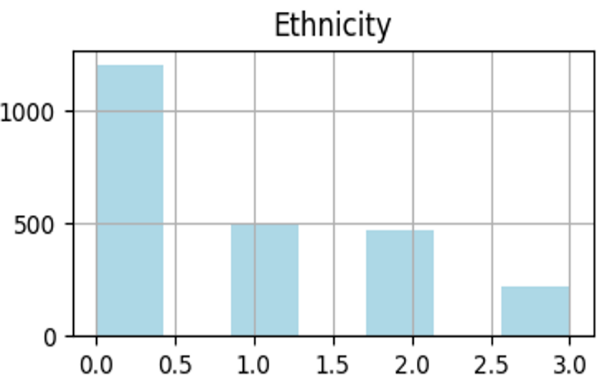
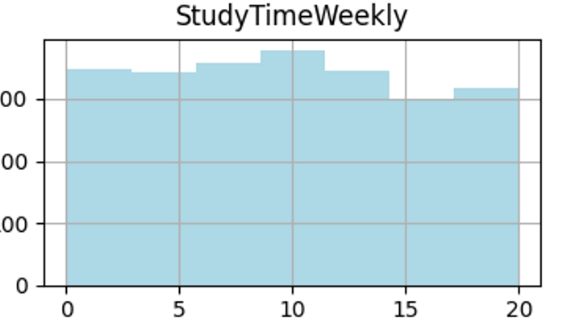
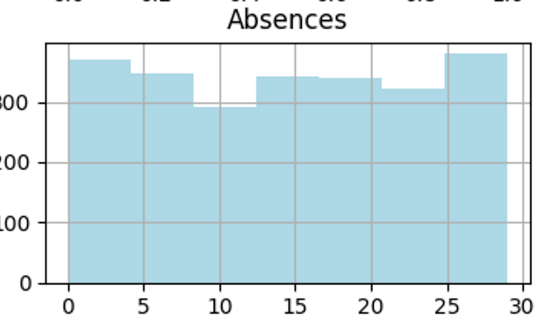
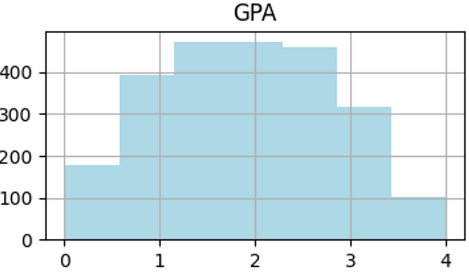
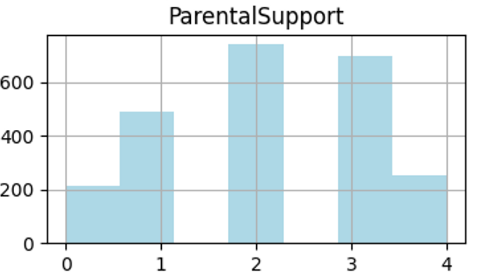
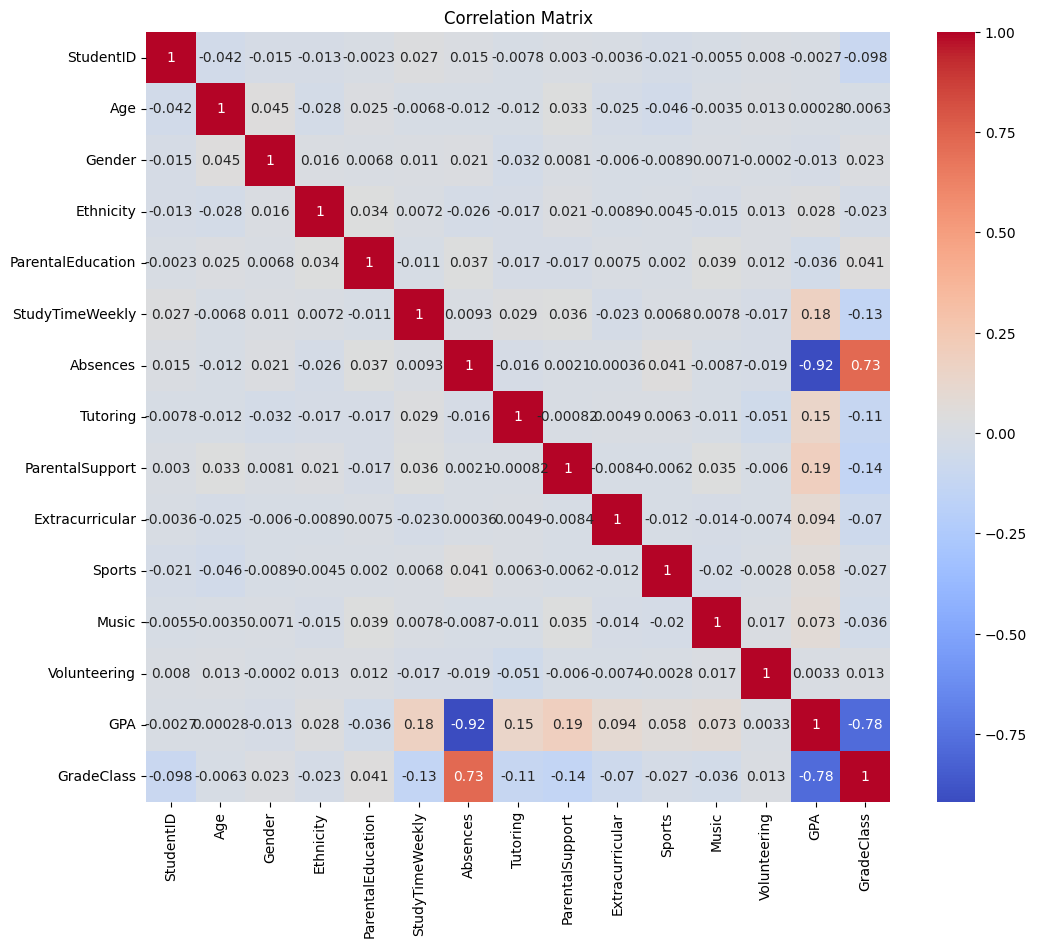
Preprocessing-

* Checking for Categorical and Numerical columns

We get-

* No categorical columns
* Numerical columns: (['StudentID', 'Age', 'Gender', 'Ethnicity', 'ParentalEducation', 'StudyTimeWeekly', 'Absences', 'Tutoring', 'ParentalSupport', 'Extracurricular', 'Sports', 'Music', 'Volunteering', 'GPA', 'GradeClass'])
* No Null Values

EDA-

* Histogram for Numerical columns-
* The age group is centred around 15-18 years.
* **Study time** over the week is distributed around 10 hours with no extreme outliers.
* Absences are skewed towards 0–10, but extends up to 30, indicating some students have frequent absences.
* Students have less participation for Extracurricular like such as Music and Sports or Volunteering. (majority is 0)
* Ethnicity is coded as- 0: Caucasian, 1: African American, 2: Asian, 3: Other
* Parental Education varies such as, 0: None, 1: High School, 2: Some college, 3: Bachelor’s, 4: Higher. With most parents having College level education.
* Parental Support is coded as- 0: None, 1: Low, 2: Moderate, 3: High, 4: Very High. With most parents supporting in range moderate to high. 
* Correlation heatmap-

1. Strong Positive Correlation with GradeClass:

* GPA (0.78) suggests that GPA has strong influence on a student’s grades.
* Absences (0.73) shows that students who more often remain absent are at a higher risk of obtaining lower grades. This relationship simply implies that maintaining good attendance is crucial for academic performance.

1. Strong Correlation between features:

* GPA and Absences (-0.92) students who miss classes are associated with lower GPA.
* Parental Support and GPA (0.19): Students receiving parental support tend to perform better.

1. Low correlation with GradeClass:

* Gender, Age show weak (near-zero) correlation.
* Both Sports and Volunteering have minimal correlation.
* Tutoring and GPA: Weak negative correlation (-0.11), implying that tutoring might be more common among students with lower GPAs.

**Results-**

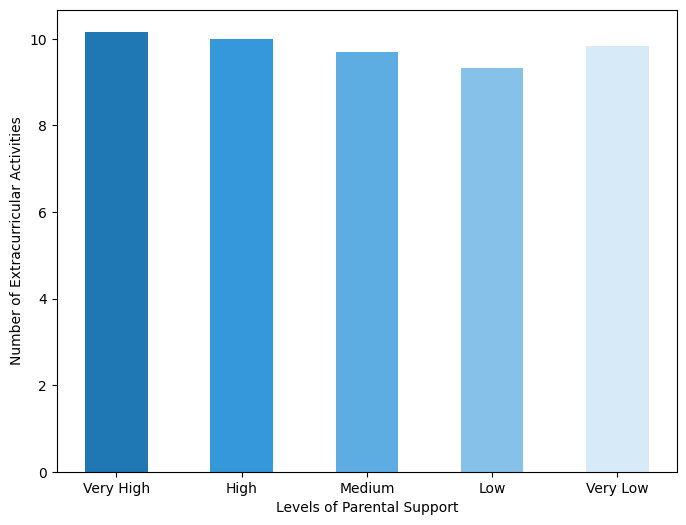
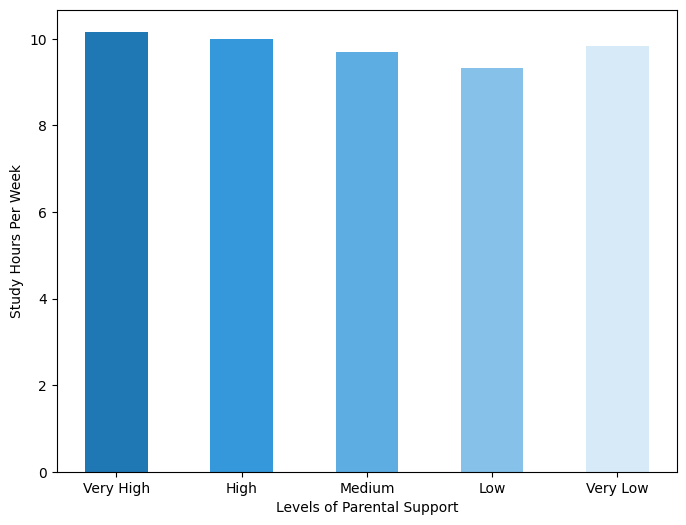
The features **StudentID**, **Music**, **Sports**, and **Volunteering** were dropped from the analysis due to their minimal or no correlation with our target variable, as shown in the correlation matrix. This removal simplifies the dataset for more focused insights.

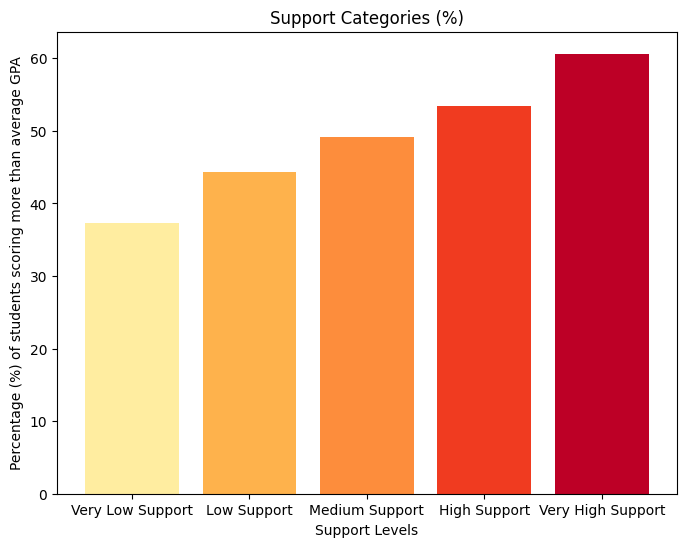
Boxplots-

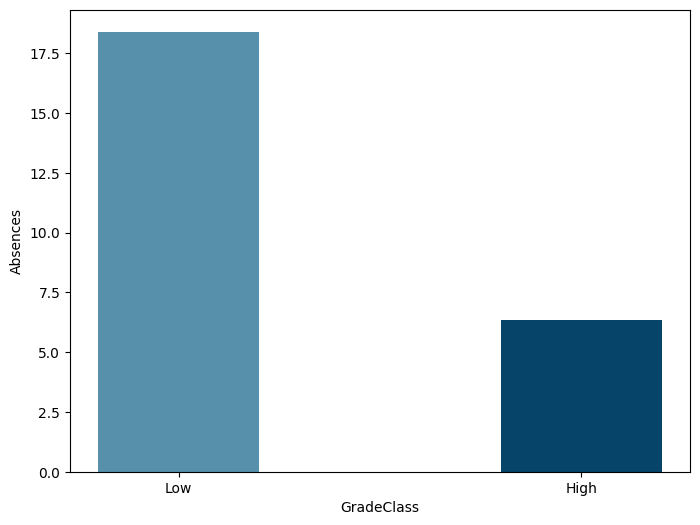
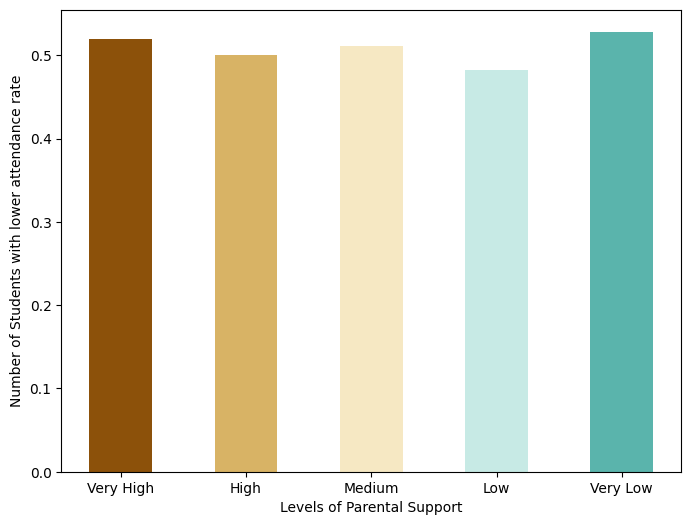
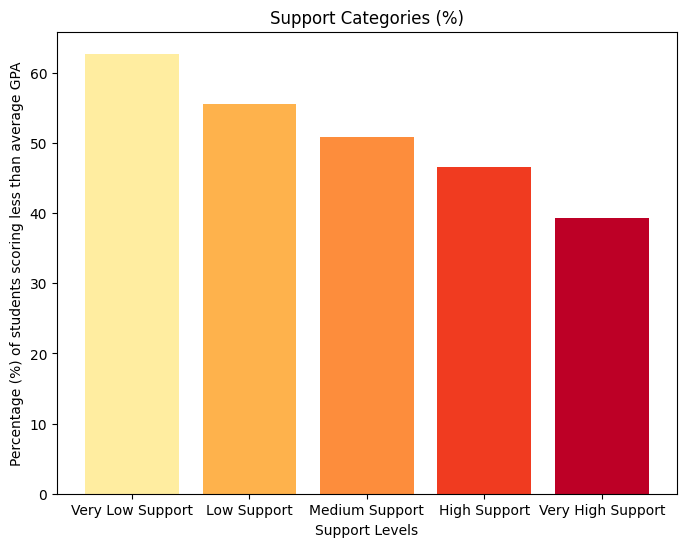
* They are used to visualize data distribution and mainly to detect outliers.
* The boxplots confirm what we see in the correlation matrix, showing that students are not very involved in activities like sports, music, and volunteering.
* Also, it is noted that there are no outliers in any of the features.

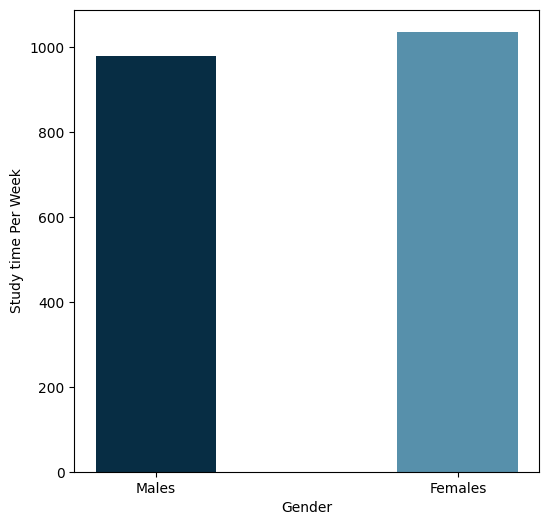
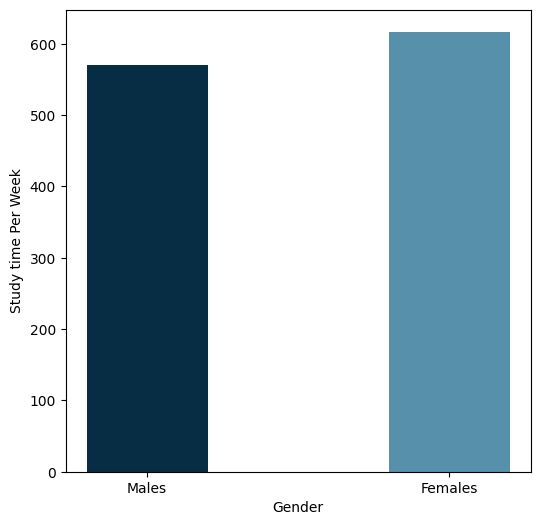
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

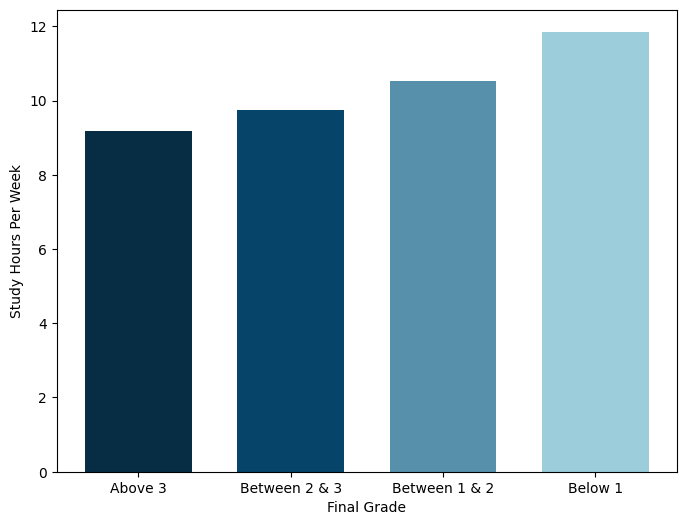
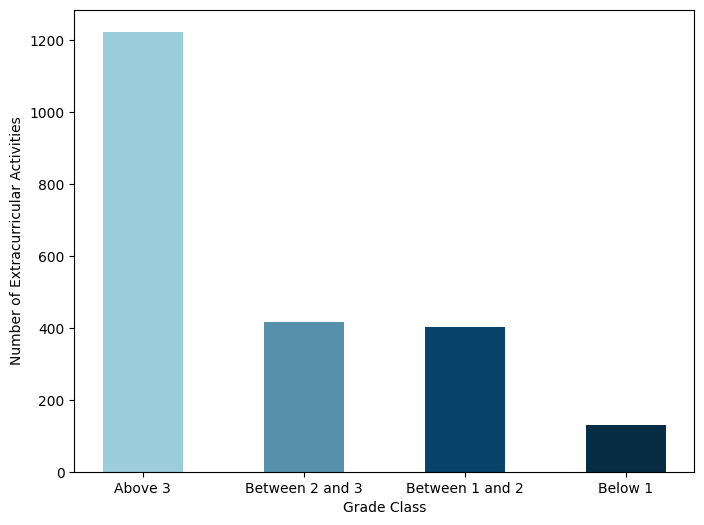
**Visualization of relation between features by plotting graphs showing their relations-**











-----> From above plots, we get the following inferences:

1. Weekly study hours of students are not influenced by the level of parental support.
2. Student participation in extracurricular activities is independent of parental support levels.
3. A higher percentage of students with above-average GPAs have strong parental support.
4. Conversely, a higher percentage of students with below-average GPAs have minimal parental support.
5. Attendance rates among students are unaffected by parental support levels.
6. Students with low final grade classes (above 3) tend to study fewer hours per week.
7. High-grade students have fewer absences compared to low-grade students.
8. Female students spend more time studying each week than male students.
9. The proportion of female students with above-average GPAs is greater than that of male students.
10. Students with lower grades (above class 3) are involved in more extracurricular activities, while those with higher grades (below class 1) are less engaged in such activities.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Standardizing Numerical Features-

* We use StandardScaler to standardize the numerical columns['Age', 'ParentalEducation', 'StudyTimeWeekly', 'Absences', 'ParentalSupport', 'GPA'

], which transform their mean to 0 and standard deviation to 1, making them easier to compare on a similar scale.

* This helps improve the performance of machine learning models by reducing the impact of differences in measurement units.

Train Test Split-

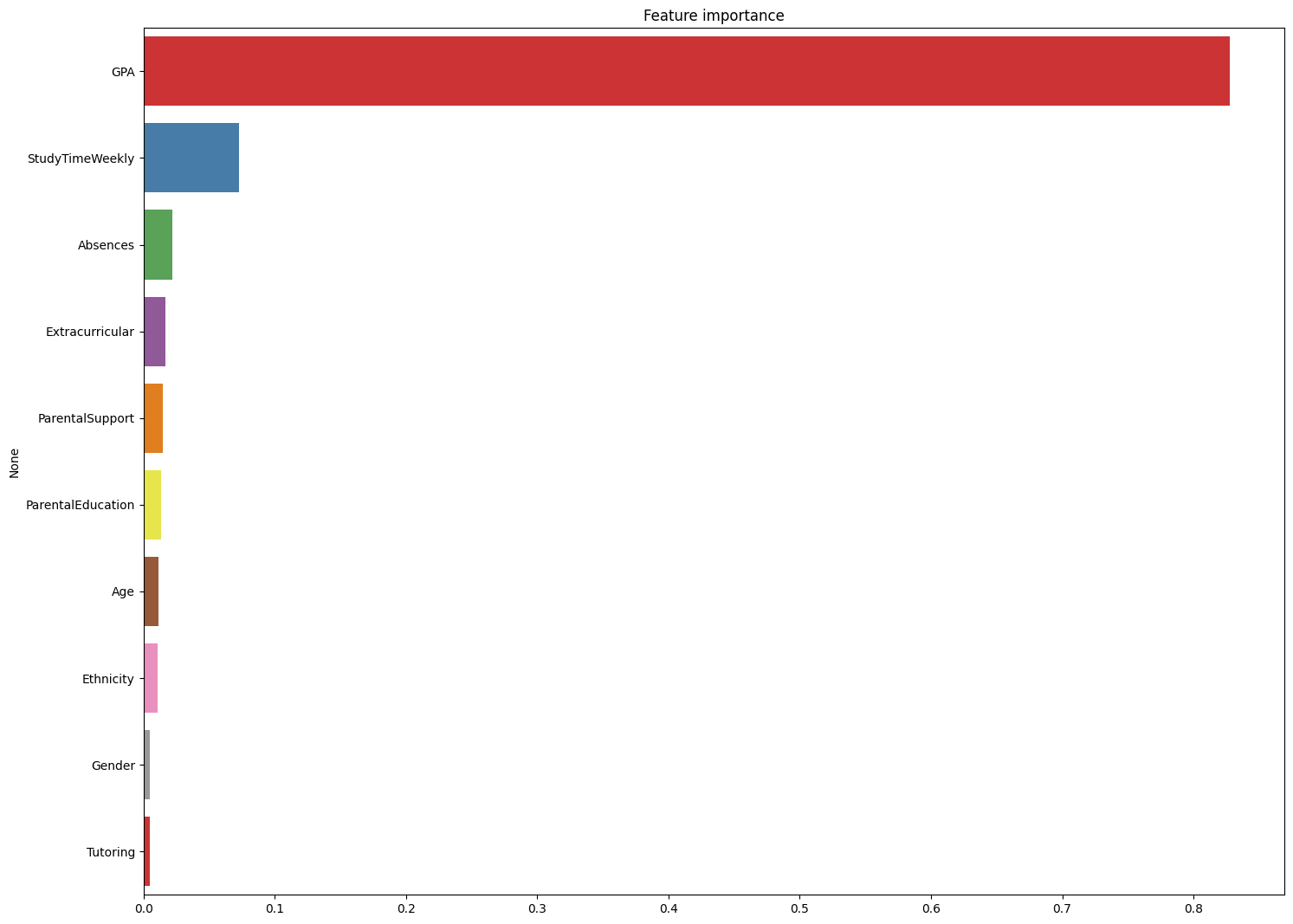
* Splitting the data into 2 parts-

80% for training and 20% for testing.

* Our target variable is **GradeClass**, and unnecessary columns like **StudentID**, **Volunteering**, **Sports**, and **Music** were removed as they are not contributing much.

Feature Importance-

* We use “RandomForestRegressor” to estimate feature importance.
* The following bar plot displays these feature importances, arranged from the most to the least significant.



Model training-

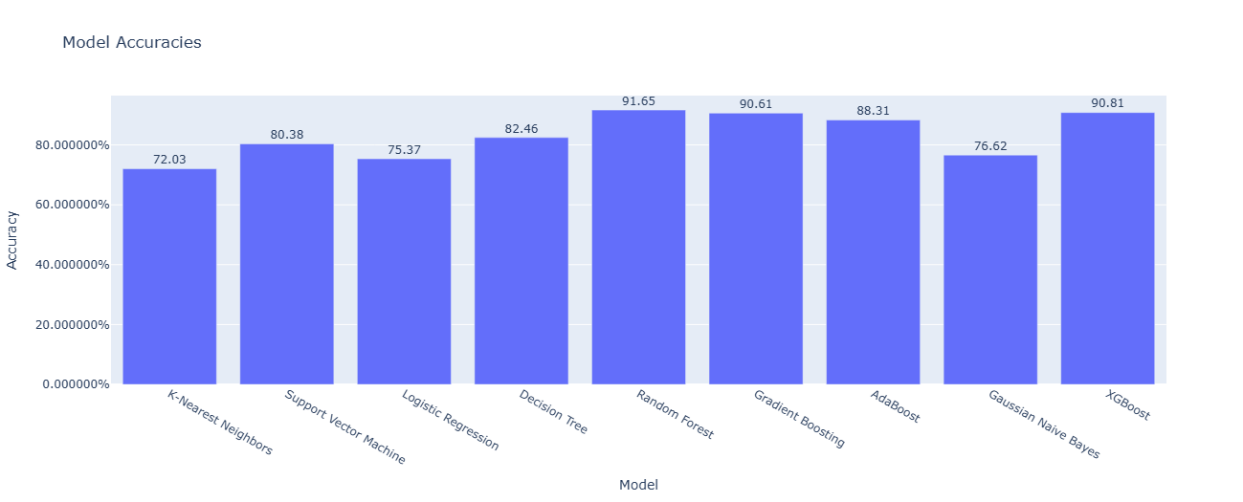
* Training various models to see which model is best for our dataset.

We use following models in our dataset-

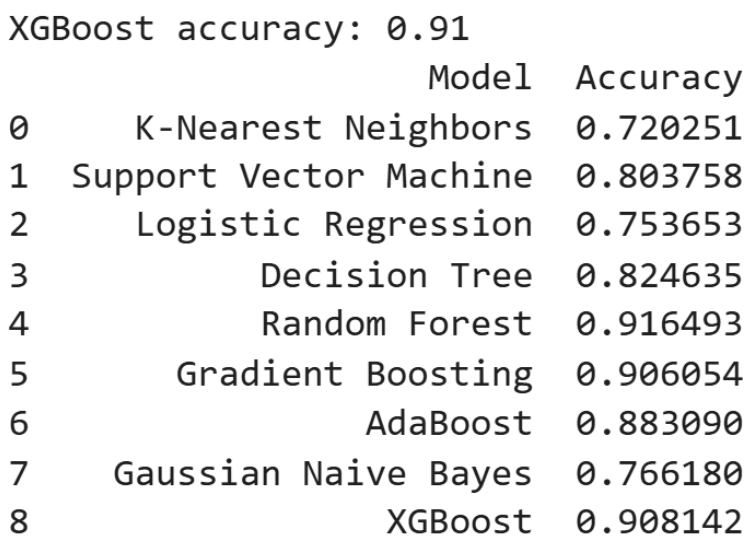
|  |  |
| --- | --- |
| S no. | Models |
| 1 | K-Nearest Neighbors |
| 2 | Support Vector Machine |
| 3 | Logistic Regression |
| 4 | Decision Tree |
| 5 | Random Forest |
| 6 | Gradient Boosting |
| 7 | AdaBoost |
| 8 | Gaussian Naive Bayes |
| 9 | XGBoost |

Each model was trained using the training set and tested on the test set to measure its accuracy. The bar chart shows the accuracy of each model as a percentage, making it easy to compare their performance.

* Accuracies for the following Models is given as-

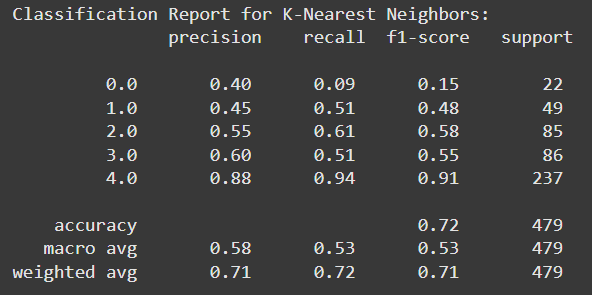


**Random Forest** has highest accuracy (91.65%) among all and **K-Nearest Neighbours** has least(72.03%).

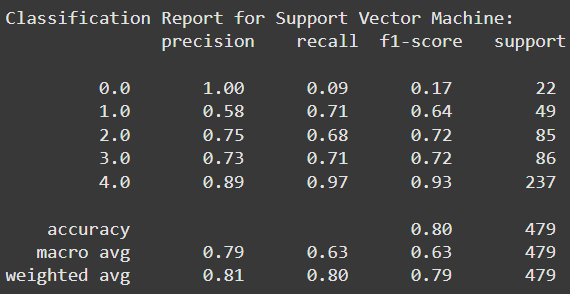


**Model Testing-**

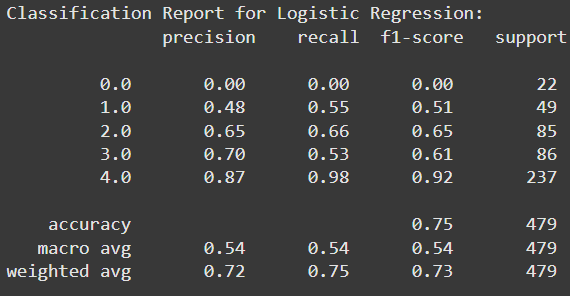
* K-Nearest Neighbours-



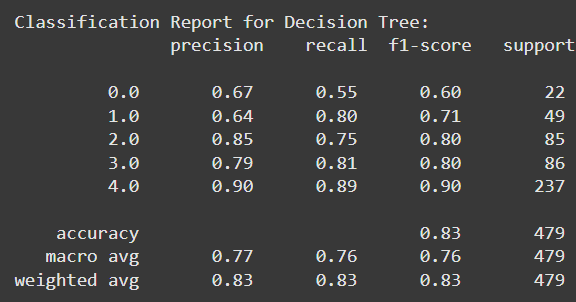
* Support Vector Machine-



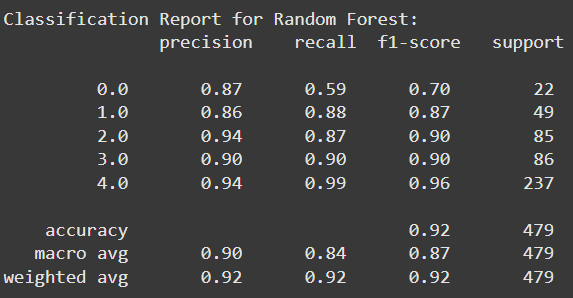
* Logistic Regression



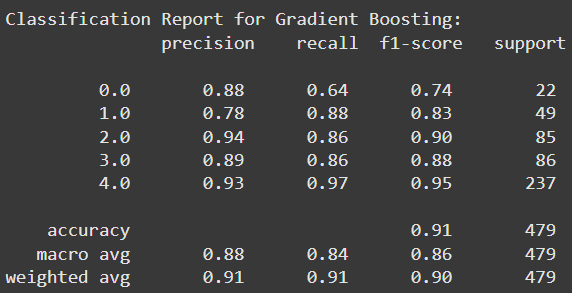
* Decision Tree



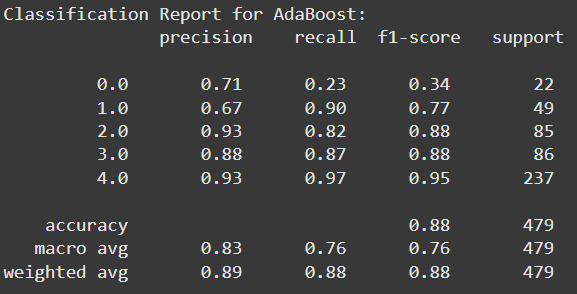
* Random Forest



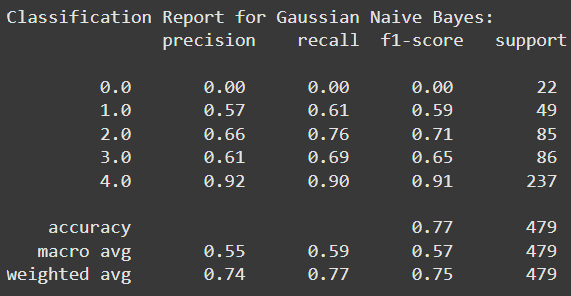
* Gradient Boosting



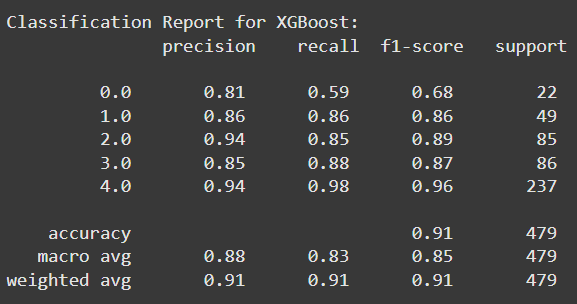
* AdaBoost



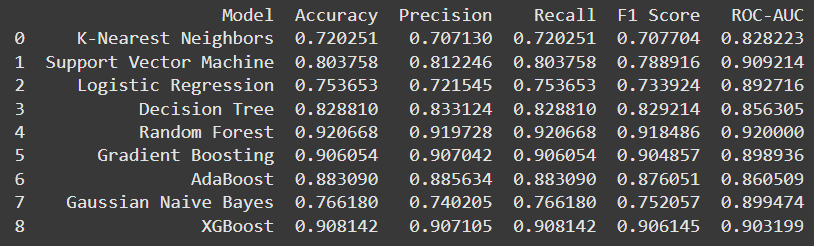
* Gaussian Naive Bayes

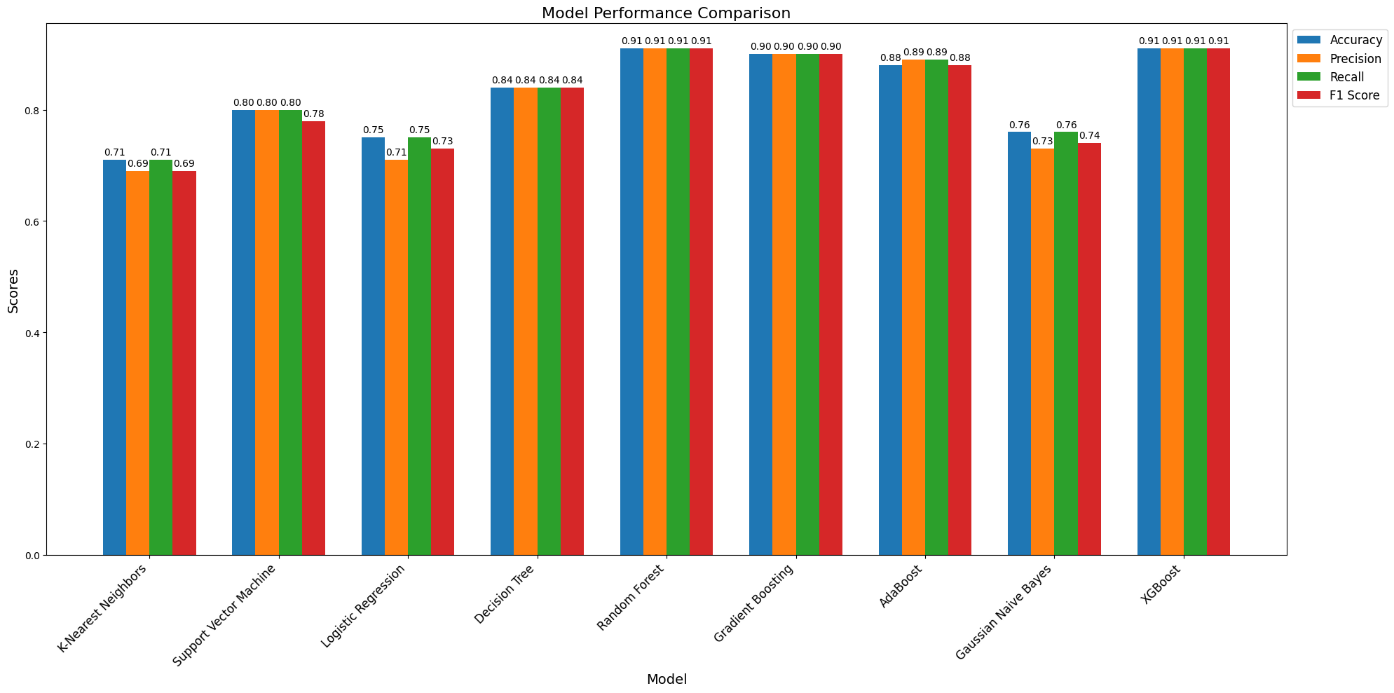


* XGBoost



Therefore,





Therefore, We have chosen 2 models for training the data i.e.

* 1. Random Forest
  2. XGBoost

Hyperparameter Tuning-

To improve the efficiency of the model we have use Hyperparameter tuning . We employed Grid Search Cross-Validation (GridSearchCV) to find the best hyperparameters for each model based on accuracy.

**Random Forest Classifier**

* **Tuned Hyperparameters**:
  + n\_estimators: 50, 100, 200
  + max\_depth: 10, 20, 30, None
  + min\_samples\_split: 2, 5, 10
* **Best Parameters Found**:

1.max\_depth: 10,

2.min\_samples\_split: 5

3.n\_estimators: 100

**XGBoost Classifier**

* **Tuned Hyperparameters**:
  + n\_estimators: 50, 100, 200
  + max\_depth: 3, 6, 10
  + learning\_rate: 0.01, 0.1, 0.2
* **Best Parameters Found**:

1.learning\_rate: 0.1

2.max\_depth: 3

3.n\_estimators: 50

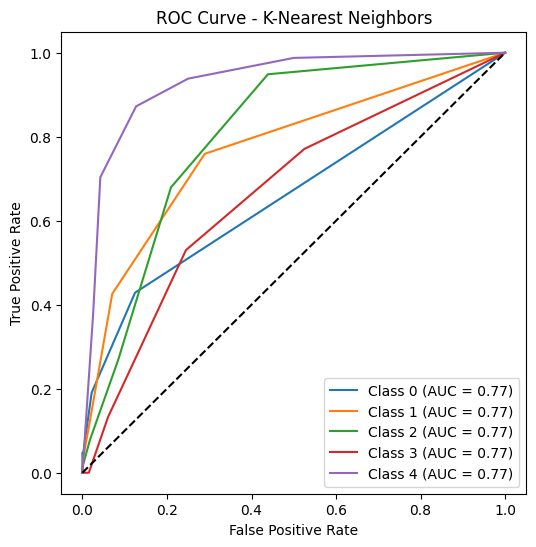
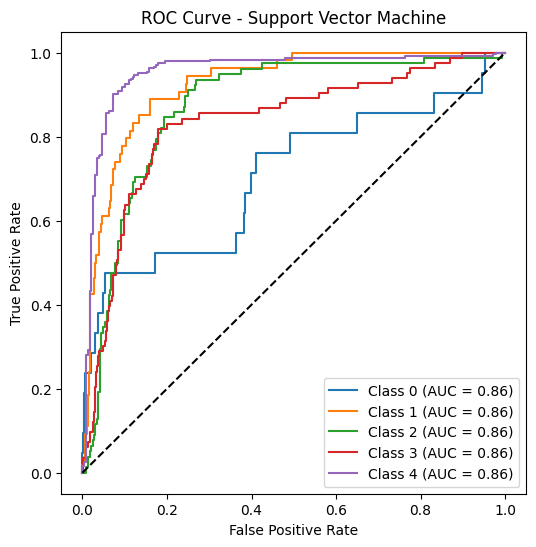
Model evaluation on different classification model and ROC and AUC Curves-

The process and results of evaluating multiple machine learning models based on key performance metrics, including Precision, Recall, F1-Score, and ROC-AUC. The analysis includes calibration of classifiers, calculation of metrics, and visualization of ROC curves and performance comparisons.

1. **Calibration and Training**: Each model was calibrated using `Calibrated ClassifierCV` if it didn’t support probability prediction. Models were trained on the training dataset and tested on the test set.  
 2. **Metrics Calculation**: Precision, Recall, and F1-Score were calculated to evaluate classification performance. ROC-AUC was used for assessing the model’s ability to distinguish between classes.

For each model, a ROC curve is plotted. The curve represents the trade-off between the true positive rate (sensitivity) and the false positive rate, giving a visual understanding of model performance at different thresholds. A higher ROC-AUC score (closer to 1.0) indicates better performance

3. **Visualization** :ROC curves were plotted to show performance at various thresholds, and bar charts provided comparisons of Precision, Recall, and F1-Score across models.

**Precision , Recall , F1 –score by the models**

