CSL 2010

STUDENT PERFORMANCE PREDICTION PRESENTATION

Rahul Ahuja(B23CH1037)
Manan Ajmera(B23CH1026)
Vishwendra Pratap Singh(B23CH1048)
Chirag Bhutra(B23MT1012)

-

INTRODUCTION-

- Our machine learning project, "Student Performance Analysis", aims to predict students' academic outcomes by analyzing various factors such as attendance, study habits, parental support, and extracurricular involvement. This project leverages machine learning algorithms to uncover patterns and insights, helping educators and institutions identify students who may require additional support or tailored resources.

- 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

PREPROCESSING

1. 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'])

2. Checking for null values-

- No Null Values

3. Dataset shape-

- Total rows : 2392 entries, 0 to 2391 index
- Data columns (total 15 columns)

4. Scaling -

- Scaled numerical features ['Age', 'ParentalEducation', 'StudyTimeWeekly', 'Absences', 'ParentalSupport', 'GPA'] by standard scaler

DATA ANALYSIS RESULTS (EDA)

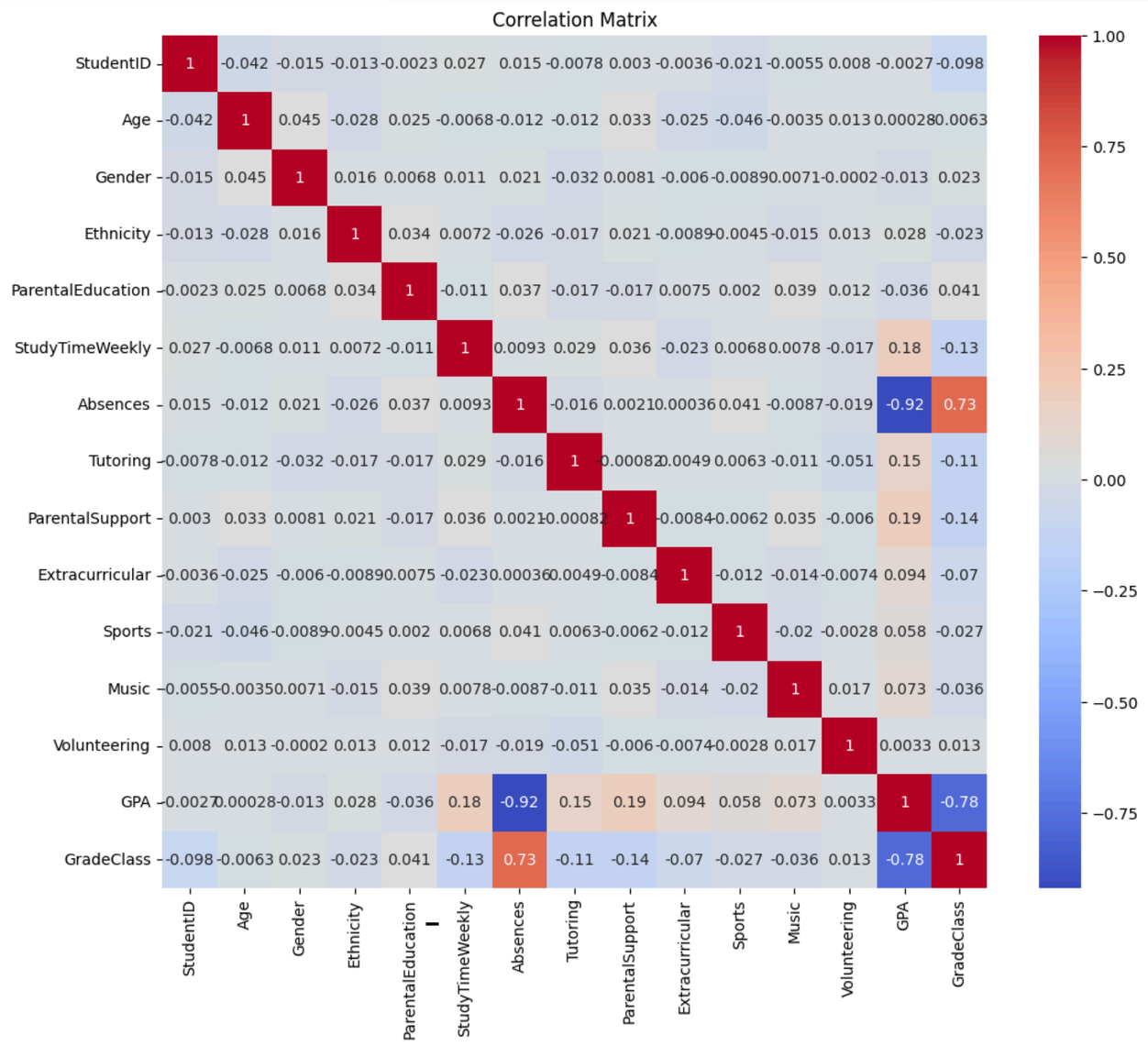
1. Strong Positive Correlation with GradeClass:
o GPA (-0.78) suggests that GPA has strong influence on a student's grades.

o 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.

2. Strong Correlation between features:
o GPA and Absences (-0.92) students who miss classes are associated with lower GPA.

3. Low correlation with GradeClass:
o Gender, Age show weak (near-zero) correlation.
o Both Sports and Volunteering have minimal correlation.
o Tutoring and GPA: Weak negative correlation (-0.11).

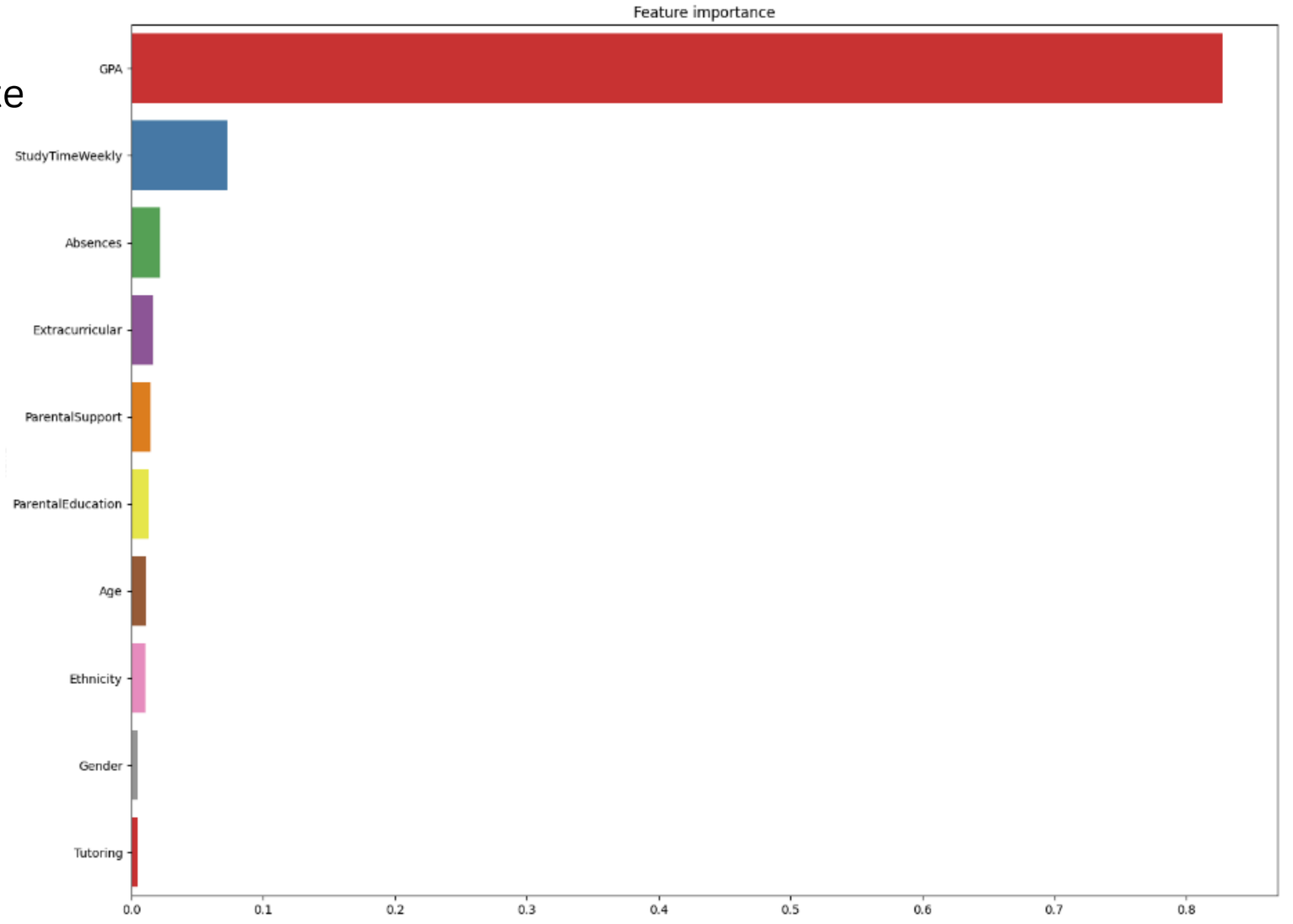
The features StudentID, Music, Sports, and Volunteering were dropped from the analysis due to their minimal or no correlation with our target variable, as shown in the correlation matrix.



FEATURE IMPORTANCE

- We have used “RandomForest” to estimate feature importance.

- Thus we concluded that GPA is most important feature, while tutoring is least



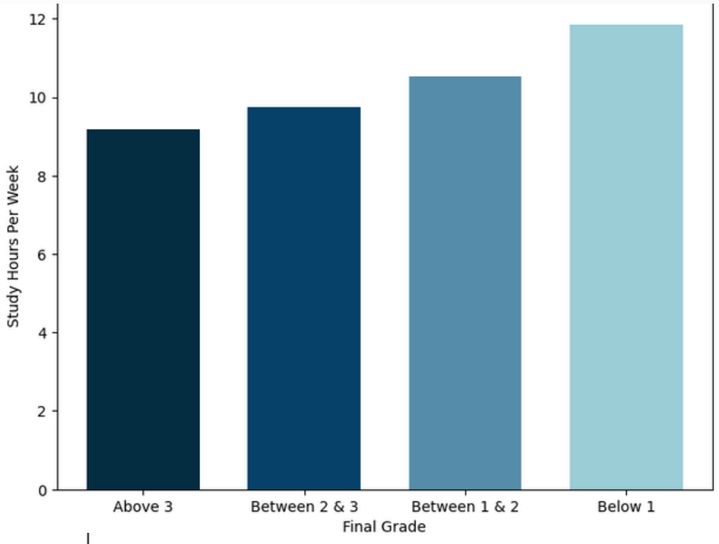
RELATION BETWEEN FEATURES

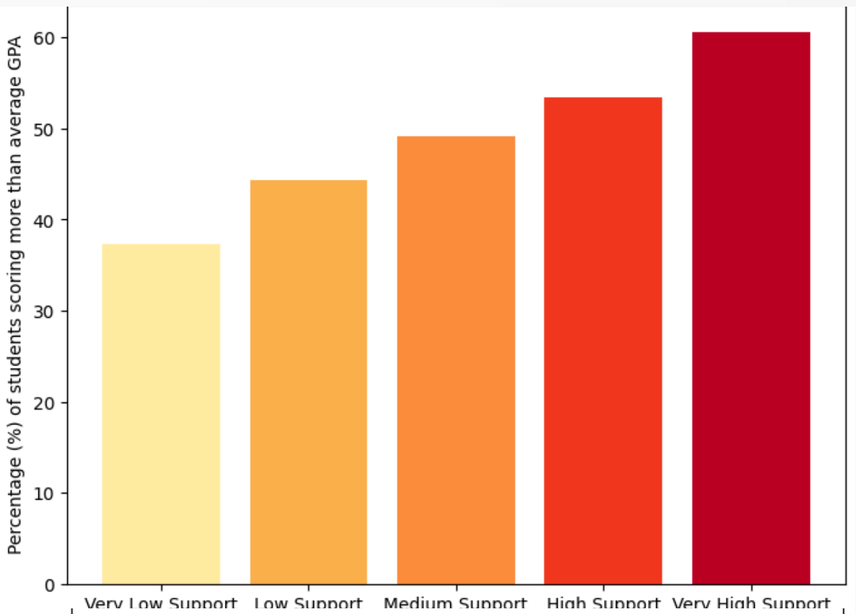
- Attendance rate is independent of Parental Support

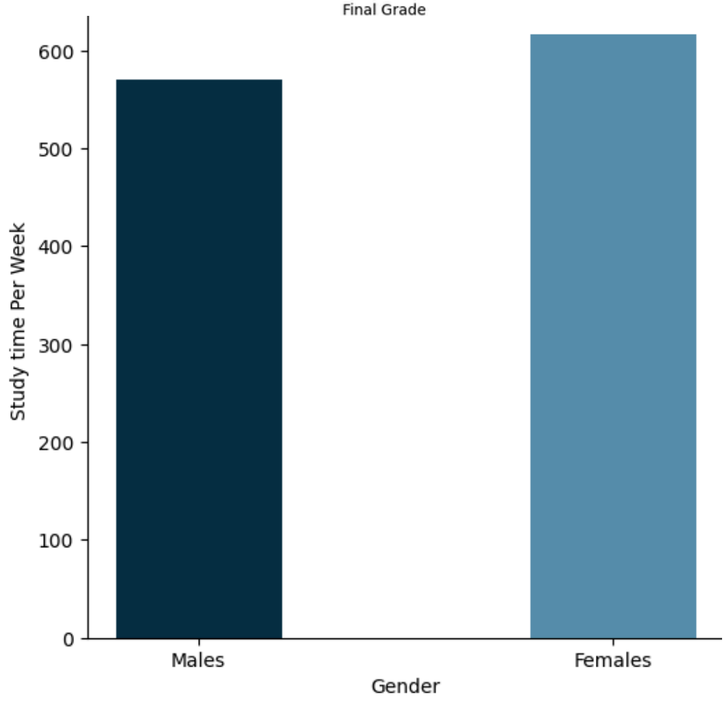
- More is the study hours per week more better will be the Grade

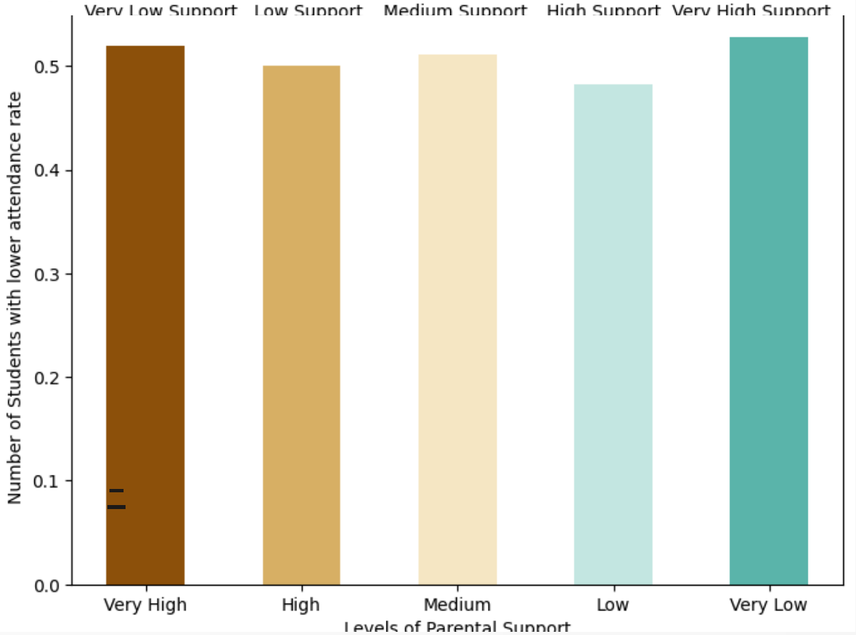
- Number of students who is scoring more than average gpa is larger for females

- It is noticed that a higher percentage of students with final GPA above average have very higher parental support.









MODELS USED

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

Model Accuracy

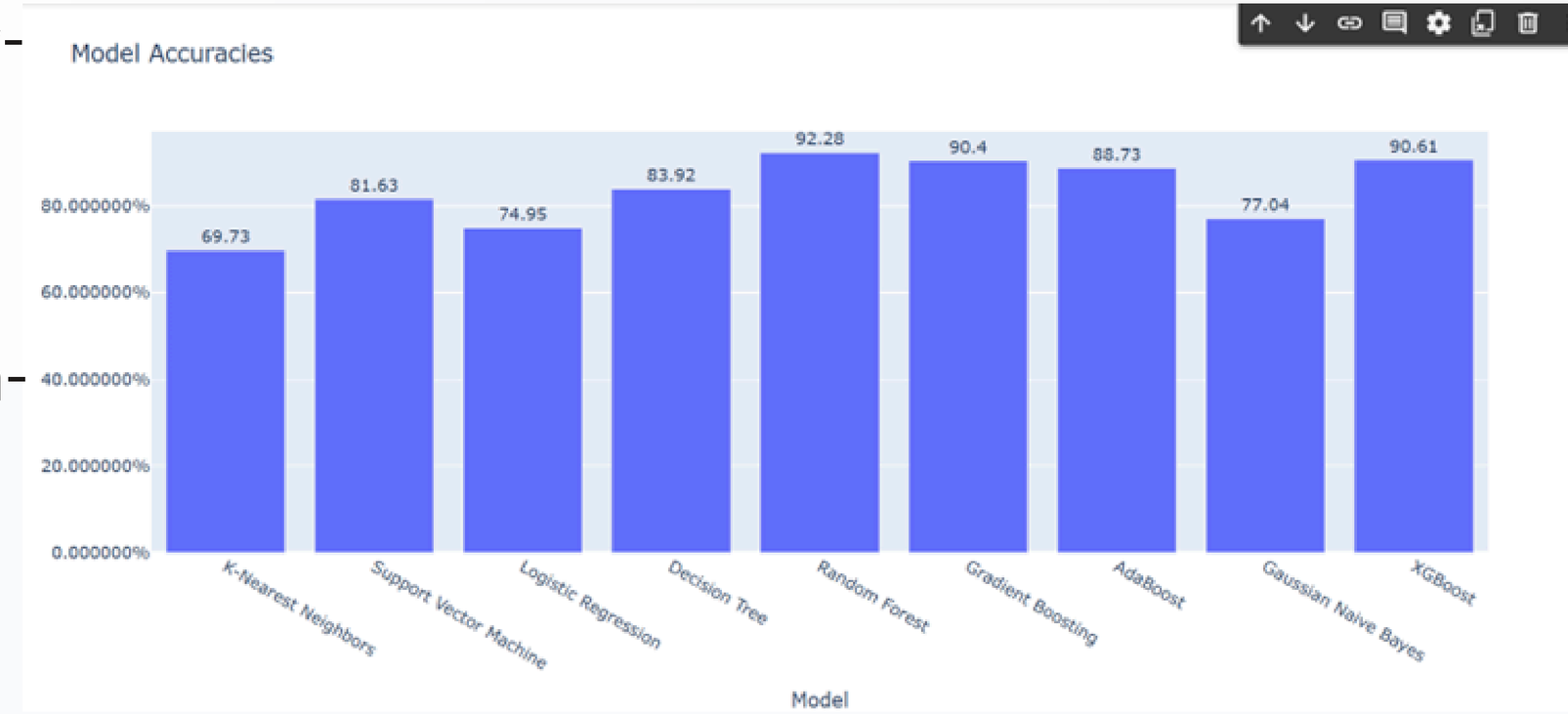
1. K-Nearest Neighbors- 0.697286
2. Support Vector Machine- 0.816284
3. Logistic Regression- 0.749478
4. Decision Tree- 0.839248
5. Random Forest- 0.922756
6. Gradient Boosting- 0.903967
7. AdaBoost- 0.887265
8. Gaussian Naive Bayes- 0.770355
9. XGBoost- 0.906054

Therefore we chose RandomForest and XGBoost as our best models

MODEL PERFORMANCE COMPARISION

Highest accuracy is obtained by-
1. Random Forest
2. XGBoost

Least accuracy is obtained from-
1. KNN
2. Logistic Regression



ACCURACY MATRIX

	Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
0	K-Nearest Neighbors	0.697286	0.691175	0.697286	0.689827	0.831309
1	Support Vector Machine	0.816284	0.827685	0.816284	0.802512	0.917966
2	Logistic Regression	0.749478	0.714032	0.749478	0.728252	0.892948
3	Decision Tree	0.832985	0.837038	0.832985	0.833235	0.862169
4	Random Forest	0.920668	0.919915	0.920668	0.918836	0.921391
5	Gradient Boosting	0.906054	0.907010	0.906054	0.904905	0.912971
6	AdaBoost	0.887265	0.887558	0.887265	0.875741	0.851674
7	Gaussian Naive Bayes	0.770355	0.746119	0.770355	0.756784	0.899337
8	XGBoost	0.906054	0.905210	0.906054	0.904284	0.913114

Random Forest, Gradient Boost and XGBoost achieves highest performance among the models (all the scores varying from 90%-92%).
This demonstrates excellent balance between precision, recall, and overall classification capability, which indicates that the models can accurately predict both positive and negative classes.

Logistic Regression and Gaussian Naive Bayes both have lower accuracies (0.75 and 0.77, respectively), indicating that they may not capture the complexity in the data as effectively as other

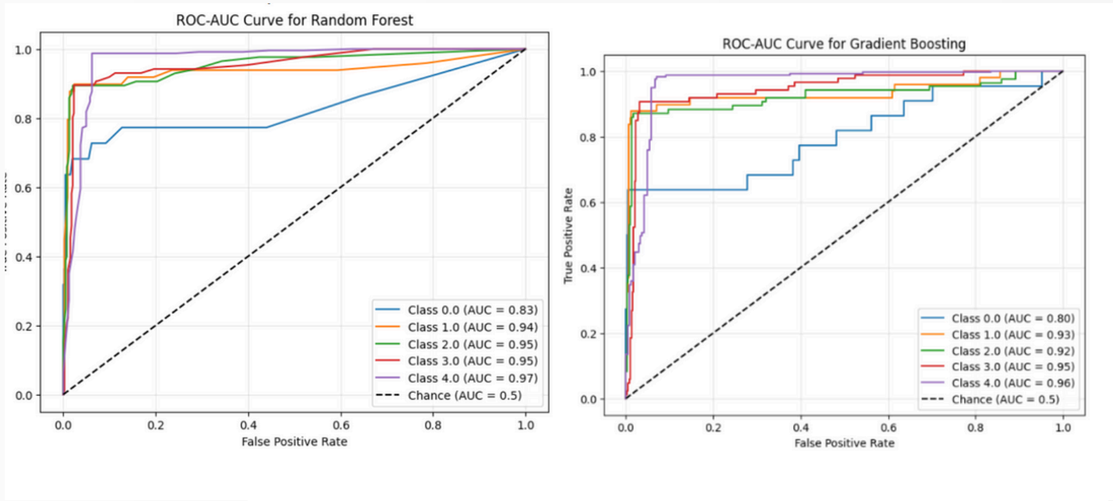
ROC-AUC PLOTS FOR SOME MODELS

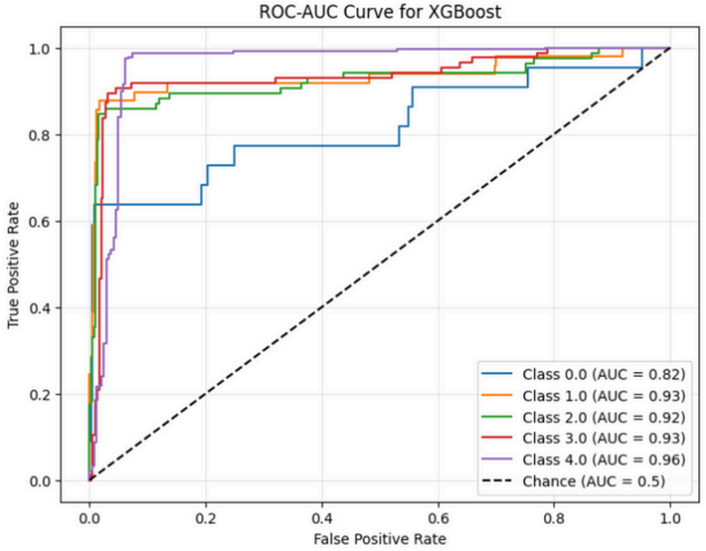
- All three models achieve high AUC values for most classes, indicating strong performance in this classification task.

- Random Forest slightly outperforms the other models in overall AUC, especially for Class 4, which has the highest AUC across all models (0.97).

- Class 0 is consistently challenging for all models, with AUC values around 0.80 to 0.83, suggesting that distinguishing this class may require additional tuning or feature engineering.

- These plots demonstrate that while all models perform well, there are some class-specific variations in performance. Random Forest appears to have a slight edge in overall discriminative power.



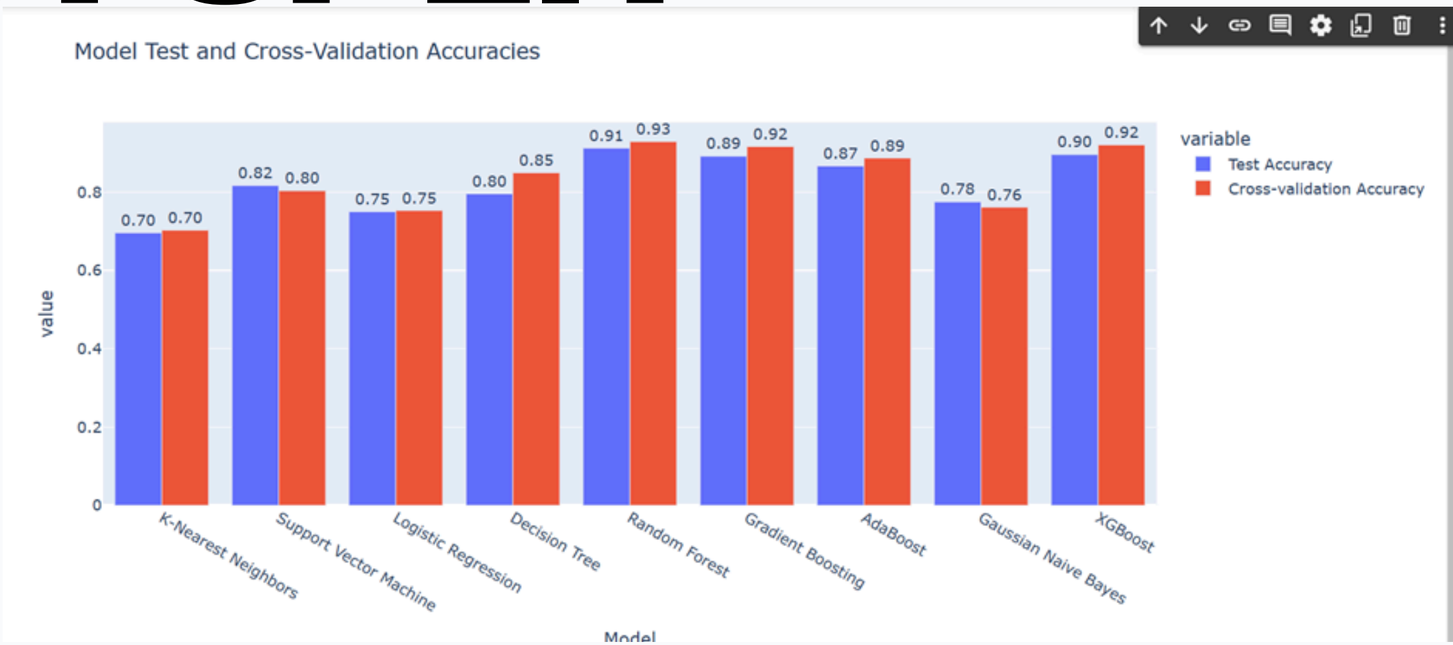


TRAIN-VALIDATION-TEST SPLIT

- Split the data into 80-10-10, then again train the models for to prevent from overfitting.

- Finally print and plot the accuracies before and after applying validation.

- Overall the accuracies increase from 2-5% for each of the models



HYPERPARAMETER TUNING

The Random Forest model is tuned using a parameter grid that includes the number of estimators (n_estimators), maximum depth of trees (max_depth), and minimum samples required to split an internal node (min_samples_split).

```
Best Random Forest Parameters: {'max_depth': 20, 'min_samples_split': 10, 'n_estimators': 100}
Best Random Forest Cross-Validated Accuracy: 0.9279
Random Forest Test Accuracy: 0.9207
```

For XGBoost, the tuning parameters include n_estimators, max_depth, learning_rate, and gamma.

```
Best XGBoost Parameters: {'gamma': 0, 'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 200}
Best XGBoost Cross-Validated Accuracy: 0.9273
XGBoost Test Accuracy: 0.9144
```

The hyperparameter-tuned Random Forest and XGBoost models achieved similar cross-validated accuracies (approximately 0.93) on the training data, demonstrating strong performance.

CONCLUSION

- The dataset is quite balanced with all class having considerate amount of instances, also no null values to be deal with.
- There are a few columns which have very minimal correlation which target variable hence are dropped.
- The data is scaled by standard scaler.
- Feature importance shows that GPA is most important in the prediction of GradeClass.
- We train on various models with highest accuracy for random forest and lowest for knn.
- Further To assess the effectiveness of various classification models in predicting student grade classifications, a range of performance metrics were calculated for each model, including accuracy, precision, recall, F1 score, and ROC-AUC.
- Hyperparameter tuning for XGBoost and Random Forest, best parameters found which increases the accuracy.
- Cross-Validation performed on all the models which also increases the accuracies.

THANK YOU!

Pranav Pant for such an excellent guidance