Title: Prediction of students’ performance using random forest regression

**Question 1**

**Introduction:**

Random Forest Regression is used for predicting the academic performance of students on the basis of the dataset. Machine learning techniques are used to attain insights and better forecasts about the final grades of the students. The efficacy of Random Forest Regression is explored.

**Prediction Tasks:**

Two key aspects revolve around prediction tasks which are Classification Task, Regression Task.

* **Classification Task:**

Categorized the students into three categories: Poor Achieving, Average Achieving, and Well Achieving on the basis of their final grades.

The Random Forest Classification model is used for achieving accurate predictions.

The performance of the model is assessed by the matrices, recall, precision, F1-score, and accuracy.

* **Regression Task:**

The Random Forest Regression model is used to predict the final year grades (G3) of the students. The performance of the model is evaluated to know its ability for prediction. The importance of multiple features is visualized in this model by feature importance graphs.

**Random Forest for Prediction:**

Random Forest is used to handle classification and regression tasks as it is considered one of the powerful ensemble learning methods. Random Forest is applicable to better predict the performance of students as it can manage complex relationships in data and can reduce overfitting. For educational data analysis, this model is considered the correct model for it can handle categorical and numerical features. Random Forest can accurately predict which factor is important in assessing the success of students.

**Question 2**

**Exploratory Data Analysis (EDA):**

2.1.1 Size of the data

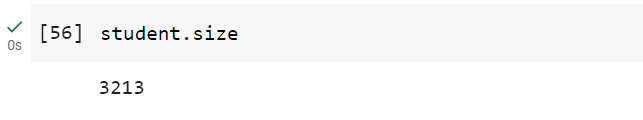
****

Figure 1: Indicates that the size of the data is 3213.

2.1.2 Number of rows and columns in the data:

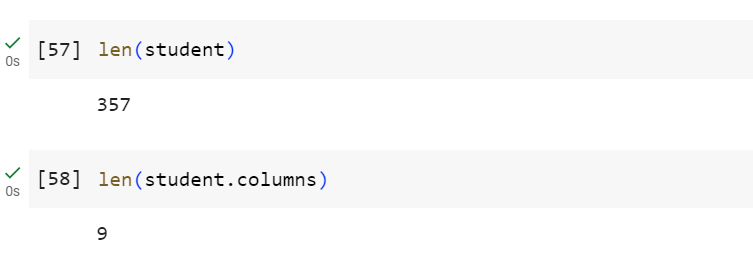
****

Figure 2: A Python code snippet showing the number of rows and columns in the data.

According to Figure 2, there are 357 rows or samples and 9 columns or features in the data.

2.1.3 The Grades scored by the students:

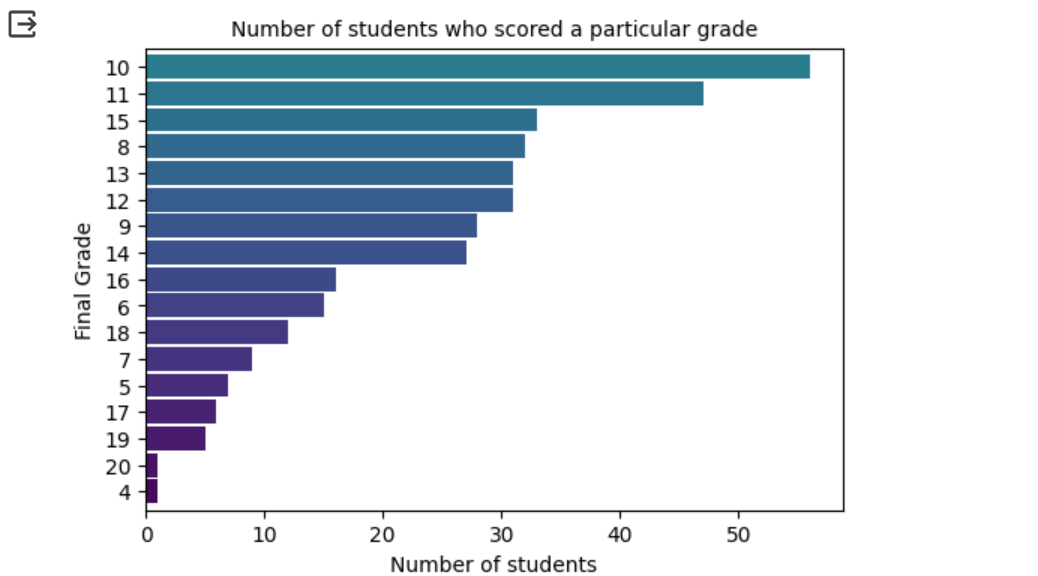


Figure 3: This code creates a horizontal bar plot for visualization of the distribution of final grades (G3) in the dataset, the y-axis showing the number of students and the x-axis showing the final grade. The color palette is 'viridis'.

2.1.4 Checking the distribution of Age along with gender

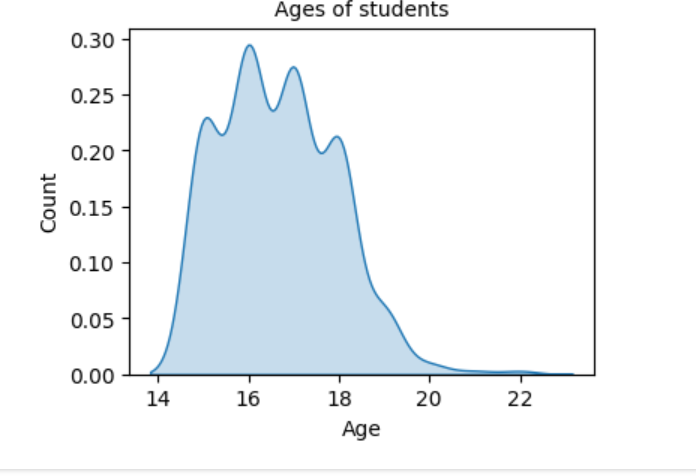


Figure 4: Ages of students

2.1.5 Does age affect final grade?

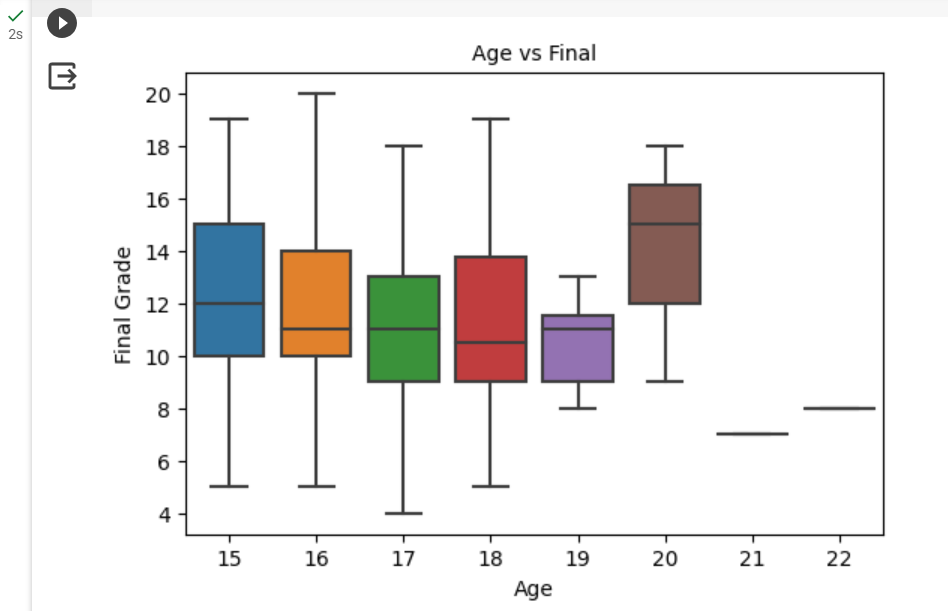


Figure 5: Age vs Final

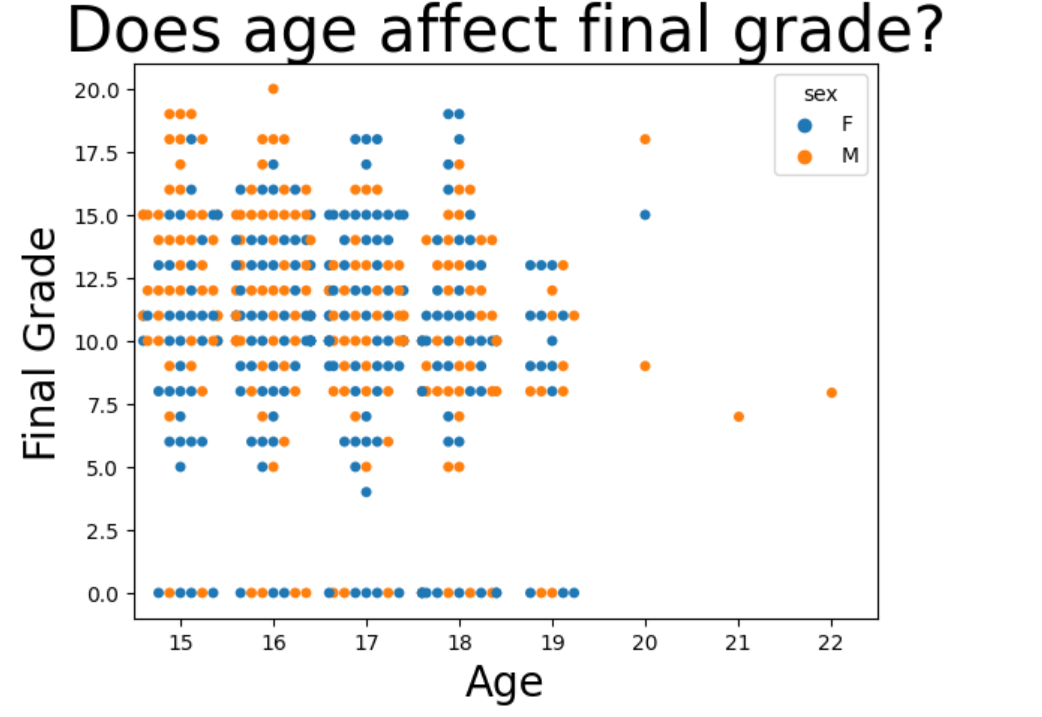
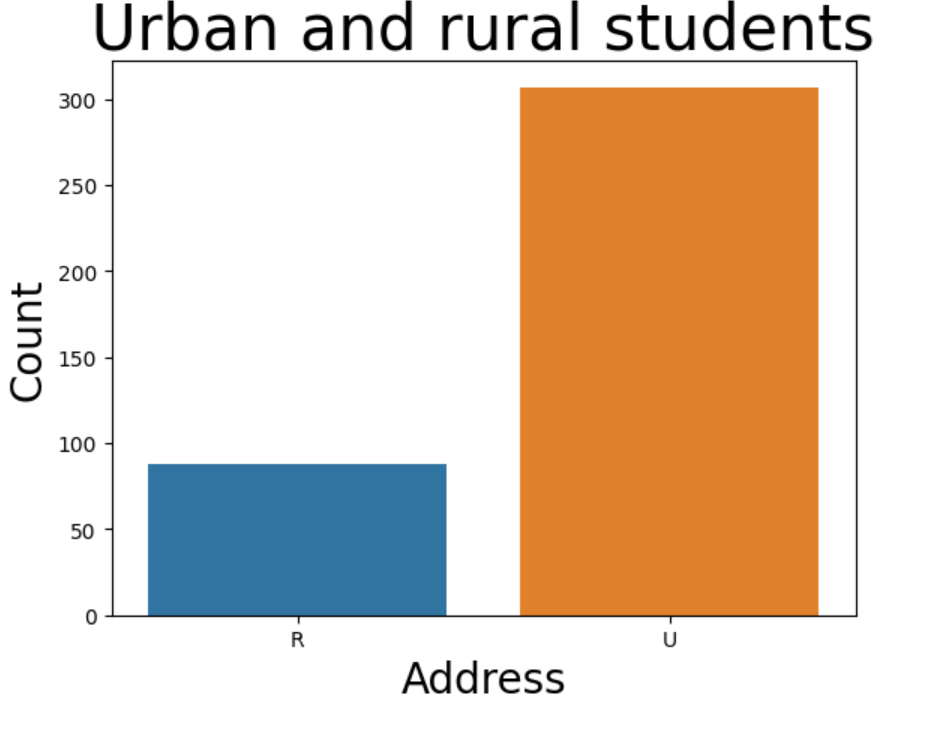


Figure 6: Plotting the distribution.

2.1.6 Count of students from urban and rural areas



By plotting the graph to know that does it make a difference on the grades because of the location.

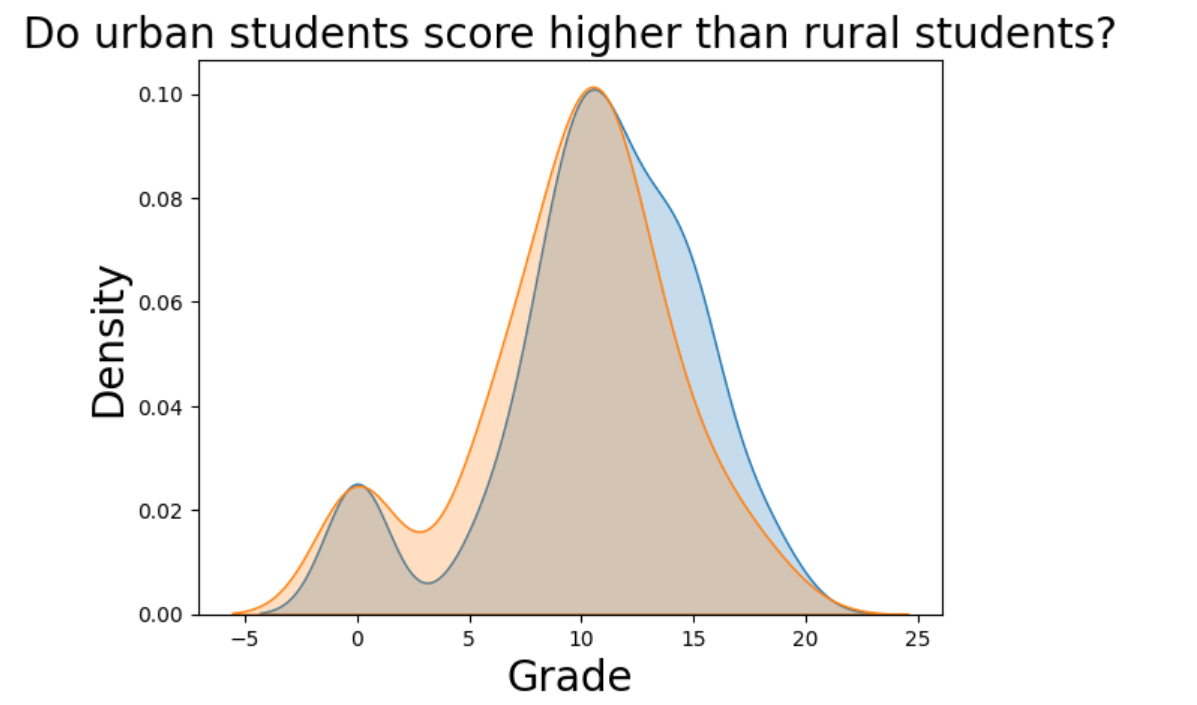


Figure 7: Urban VS Rural Students.

2.1.7 Finding Correlation

Finding a correlation between numeric values, applying one hot encoding, and dropping the G1, and G2 even if they are highly correlated with G3, but dropping them will be more useful to know that the other factors are also effecting the grades.

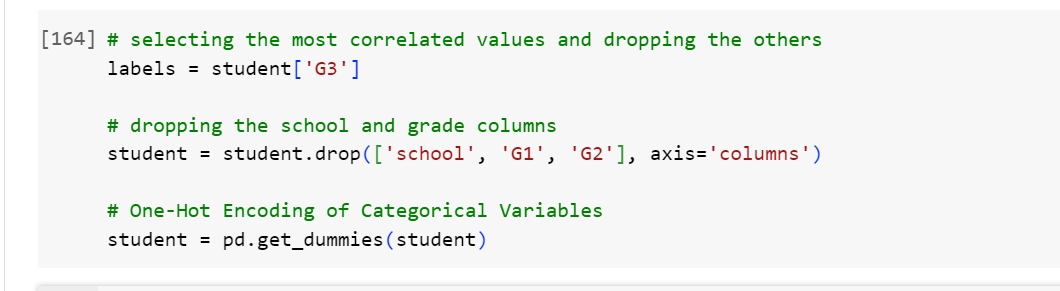


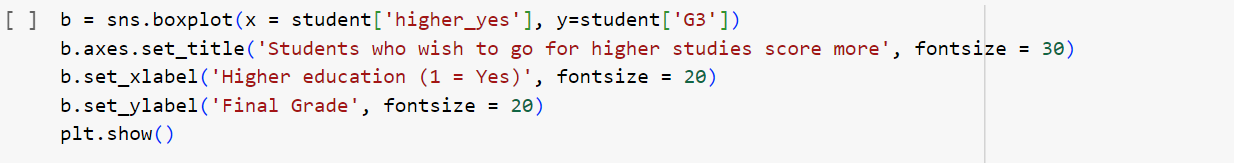
Figure 8: Dropping G1 & G2

2.1.8 Analyzing the variables and training the model

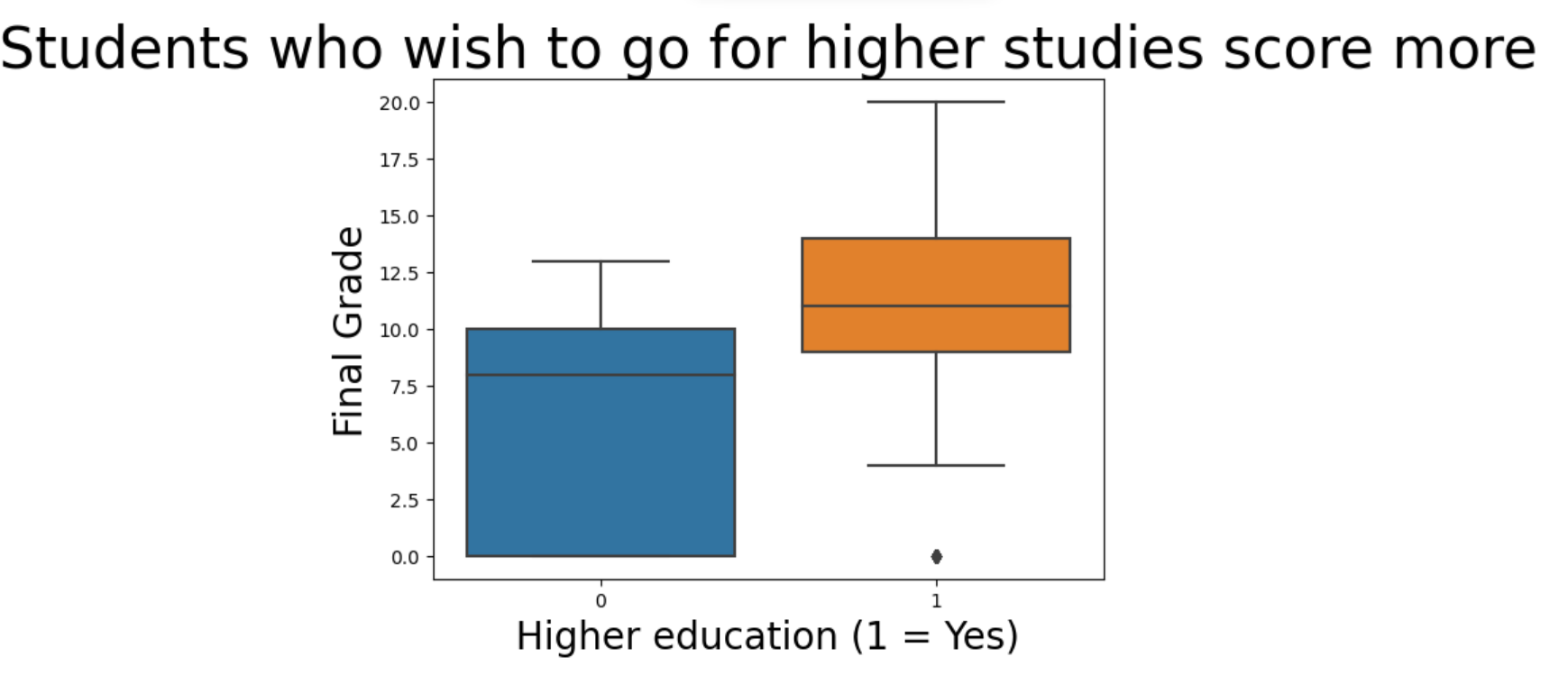


Figure 9: Students who have faced less failures will score even better and better.

2.1.9 Higher Education

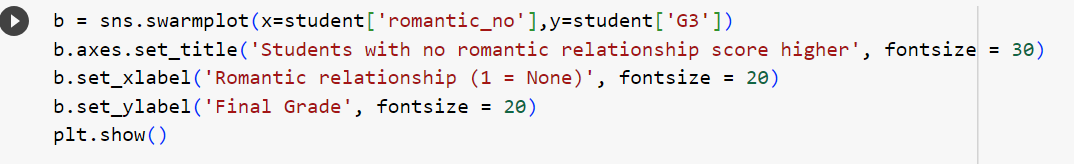


So below Figure displays whether this feature (which labels whether students want to go for higher education or not) affects the final grades or not. In the dataset this feature was given categorical values as "yes" or "no" but by applying one hot encoding we converted it into 2 variables,

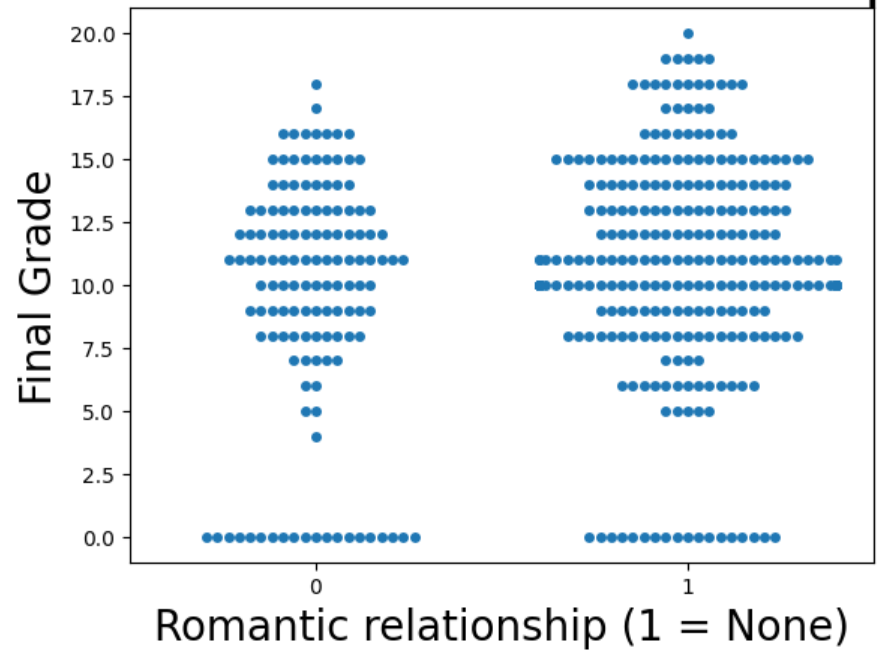


So, this feature is affecting the final grades.

2.1.10 Having Romantic Relationship



So, the below figure displays that the students who are having romantic relationship scores more.



2.1.11 Modeling (Regression)

The model achieved an MAE of approximately 3.70, and RMSE is approximately 4.75.



**Advanced Question 1: Regression Task**

**Preparation of Data:**

The dataset is split into features/training set X, and target variable/Test set Y. (Splitting the data into train (75%), test (25%).)

**Random Forest Regression Model:**

By training data X, the Random Forest Regression model is trained.

**Evaluation of Model:**

On the test set evaluation is done by using MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error).

**Feature Importance Visualization:**

Feature Importance Visualization is to visualize which features are important for the prediction.

**Optimization Tips:**

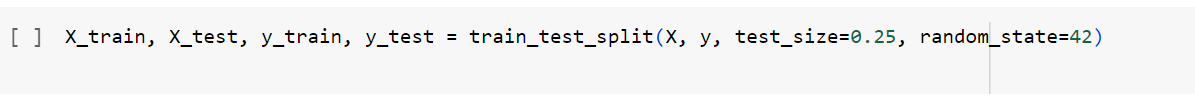
There is space for optimization, we can apply hyper-parameter tuning for even better performance of the mode.

**Advanced Question 2: Classification Task:**

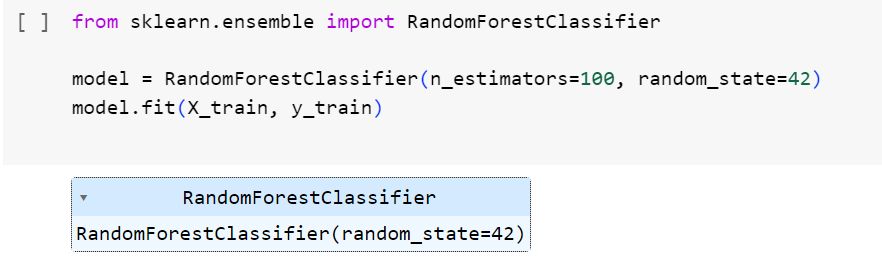
**Preparation of Data:**

The dataset is split into features/training set X, and target variable/Test set Y. Three categories are made on the basis of final grades, 'poor achieving students', 'average achieving', and 'well achieving'.

student['Achievement\_Category'] = pd.cut(student['G3'], bins=[0, 8, 14, 20], labels=['Poor Achieving', 'Average Achieving', 'Well Achieving'])

**Random Forest Classification Model:**

By training data X, the Random Forest Regression model is trained.



**Evaluation of Model:**

The model is evaluated by accuracy, and the performance of the model is assessed by the confusion matrix.

Accuracy, Classification report, and confusion matrix are printed.

Code:

from sklearn.metrics import classification\_report, confusion\_matrix, accuracy\_score

y\_pred = model.predict(X\_test)

# providing precision, recall, F1-score for each class

classification\_rep = classification\_report(y\_test, y\_pred)

confusion\_mat = confusion\_matrix(y\_test, y\_pred)

accuracy = accuracy\_score(y\_test, y\_pred)

print("Classification Report:")

print(classification\_rep)

print("Confusion Matrix:")

print(confusion\_mat)

print("Accuracy:", accuracy)

# Feature importances

feature\_importances = model.feature\_importances\_

**Comments:**

Classification report, confusion matrix, and XAI are given:

**Classification Report:**

* Precision

Precision tells positive predictions, Precision is higher for Average achieving students as per my dataset, which is 0.70, which means it can predict the students correctly 70% of the time.

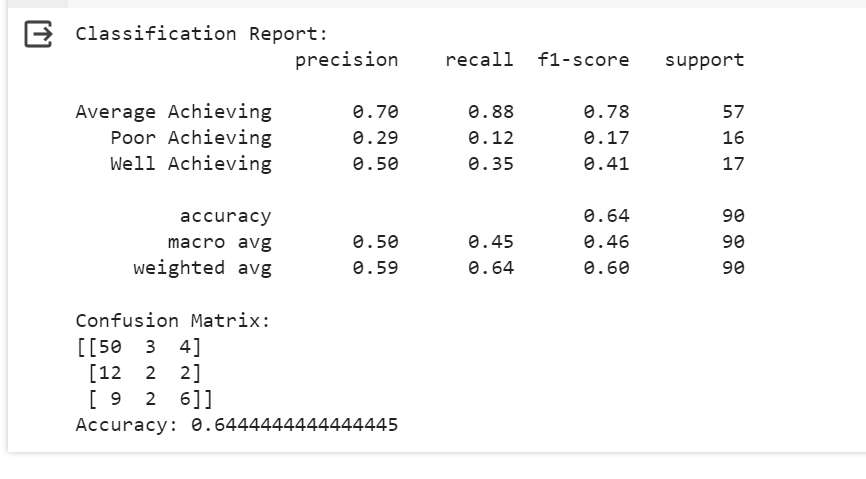


Figure: Classification Report

* Recall

Recall can perfectly capture all of the relevant instances. The model is performing well for average-achieving students as it is 0.88 but struggling with predicting poor-achieving and well-achieving students.

* F1-score

The F1-score is the harmonic mean of recall and precision, again, the F1-score of average achieving students is highest which is 0.78.

* Accuracy

The overall accuracy of the model is 0.64 which is 64%.

**Confusion Matrix**

To know the detailed breakdown of the model, the confusion matrix is created which includes, TP, TN, FP, and FN.

**Explainable AI (XAI)**

Feature visualization is one of the crucial methods of XAI, as it can clearly state which features highly affect the prediction. So I have used this feature to assess the impacts of several features on the final grades as I have shown in plotting the charts (Question 1).