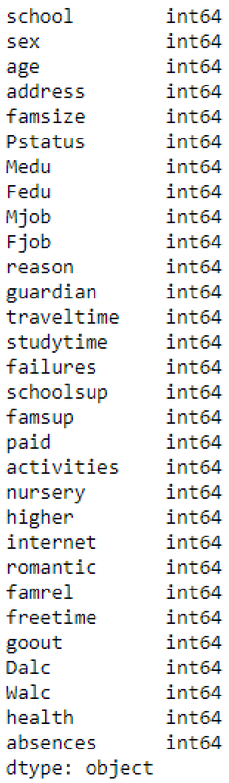
Predict Student Grades using Decision Tree Classification and Regression Tree (CART) and Specify Good variables for these algorithms

Introduction

This dataset contains the performance of student mathematics and background variables with 33 variables including first grade, second grade, and final grade as the main topic of this report. The dataset has been split into 80% train and 20% test for machine learning algorithm with random state=1. This dataset has been rechecked on the dataset and there have no missing values. Plus, the dataset has been changed into binary, numerical, and nominal variables to make sure the machine learning can learn very well.

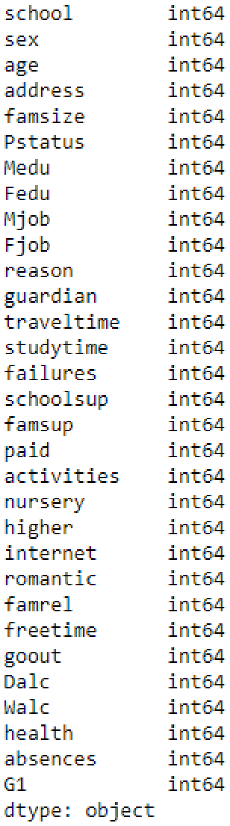
The variables G1, G2, and G3 need to remove since they are our target variables. By doing this, any possible data leaking is avoided since the model is prevented from having direct access to the G1 variables during training.



```
school        int64
sex           int64
age           int64
address       int64
famsize       int64
Pstatus       int64
Medu          int64
Fedu          int64
Mjob          int64
Fjob          int64
reason        int64
guardian      int64
traveltime    int64
studytime     int64
failures      int64
schoolsup     int64
famsup        int64
paid          int64
activities    int64
nursery       int64
higher        int64
internet      int64
romantic      int64
famrel        int64
freetime      int64
goout         int64
Dalc          int64
Walc          int64
health        int64
absences      int64
dtype: object
```

The picture above shows independent variable train test data for the variables for target variables G1

 However, the G1 must conclude this G2T train test data set since is the previous Grade before G2T.



```
school        int64
sex           int64
age           int64
address       int64
famsize       int64
Pstatus       int64
Medu          int64
Fedu          int64
Mjob          int64
Fjob          int64
reason        int64
guardian      int64
traveltime    int64
studytime     int64
failures      int64
schoolsup     int64
famsup        int64
paid          int64
activities    int64
nursery       int64
higher        int64
internet      int64
romantic      int64
famrel        int64
freetime      int64
goout         int64
Dalc          int64
Walc          int64
health        int64
absences      int64
G1            int64
dtype: object
```

The picture above shows independent variable train test data for the variables for target variables G2T

For part, G2T variables, the G2 variables which are numerical variables have been transformed into 5 categories and changed to G2T.

The main objective of this report is to specify and investigate which variables are essential for predicting grades G1 and G2T by using the Decision Tree Classification and Regression Tree (CART).

Specify which variables are important for predicting the grades

Feature Importance

 The feature importance identifies the variables that matter. By using feature selection, it can aid in greater comprehension of the problem that has been handled and occasionally result in model enhancements. This feature importance using the Gini index which is computed from the Random Forest structure.

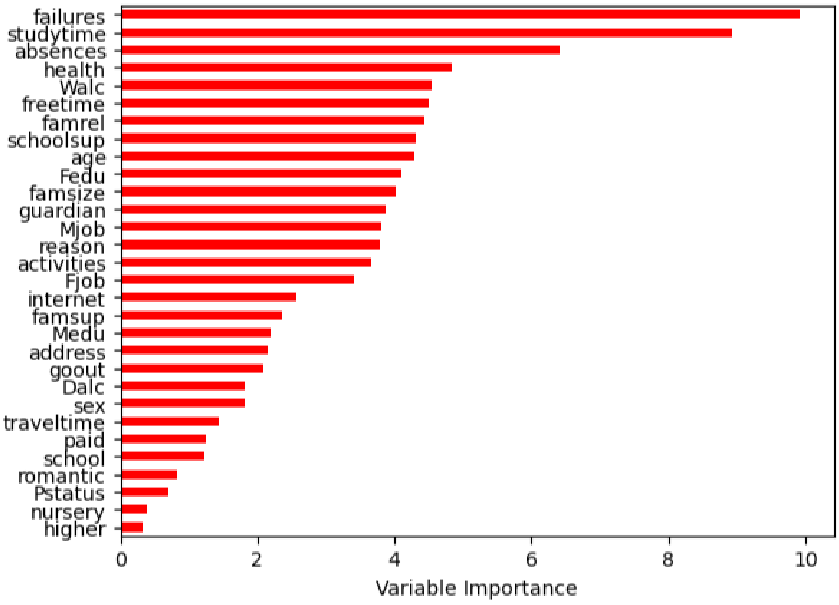
 For variable G1 feature importance, we can see clearly that failures and studytime are the main factors for G1, nursery, and higher are the less important in this graph. In the decision

Tree, the highest feature importance is the main upper decision-making category to predict the risk of student grades.

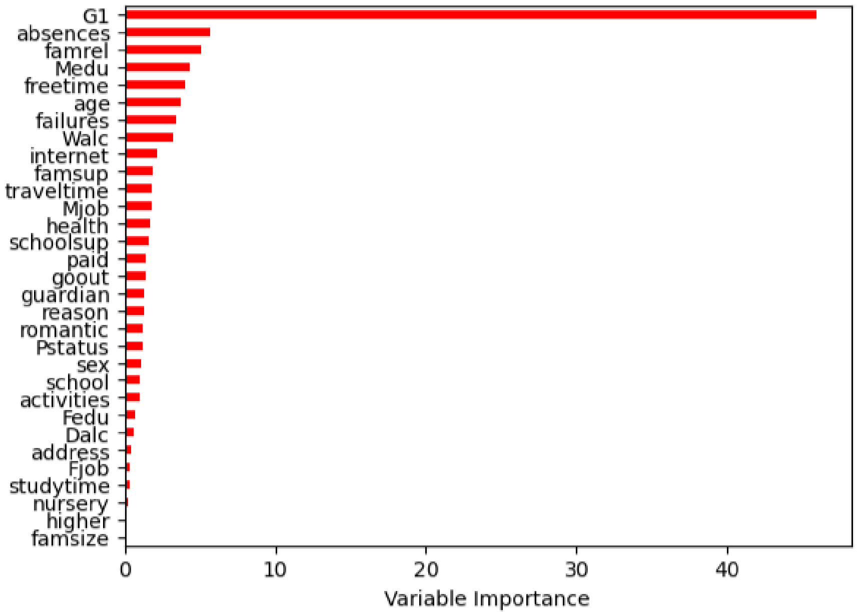
Next, variable G2T features a slightly different result from G1. As we can see, the main factor for G2T is G1 and absences which show quite interesting. The absences of variables indicate that the number of absences will affect the grade on G2T.

Another exciting part of both graph feature importance is the parent's educational background. The parent's educational background plays an important role in our target variables, especially the mother's background. The mother's educational background is much more important than the father's. The variables that stand out with ‘Medu’ are the mother's education which indicates that it is important to have education in our parents.

However, plotting all variables can cause overplotting. When several data points are plotted on top of or very near to one another, overplotting occurs, making it difficult to tell apart individual data points. This may occur when there is a significant concentration of data in one area of a plot or when the data has a narrow range of values.



The graph above shows variable importance for G1 without max_depth



The graph above shows variable importance for G2T without max_depth

MSE and Accuracy Score with Max_depth

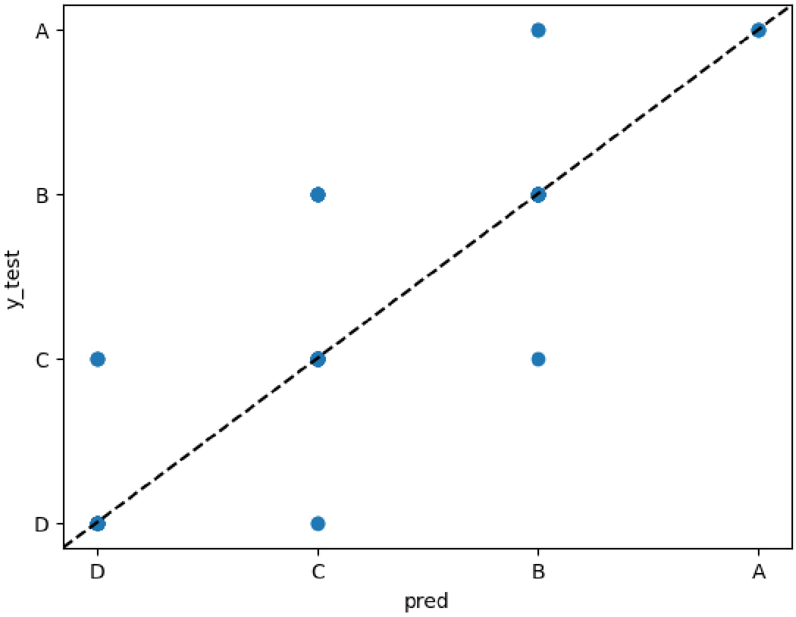
Regression Tree

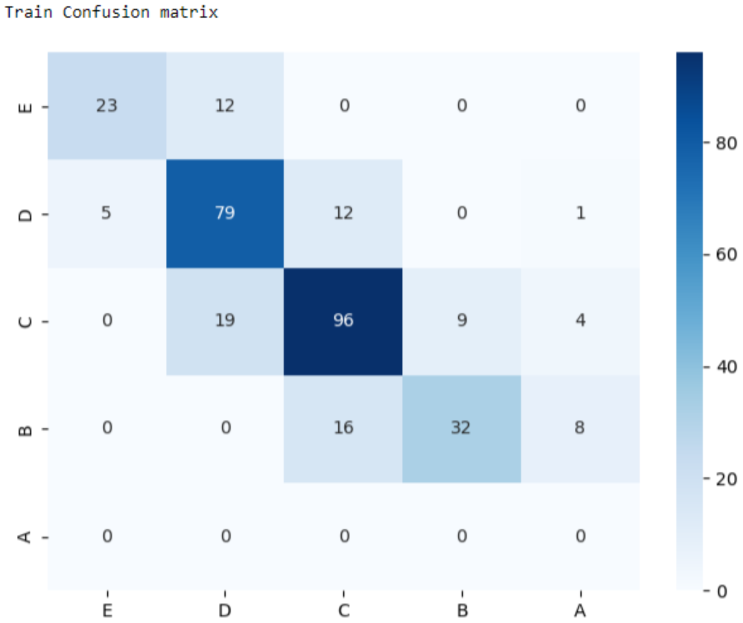
Max_depth	MSE
1	8.19612173774661
2	**7.338965243637666**
3	7.596388788522728
4	10.487372703685322
5	10.562287645642032
6	3.239790091580701

According to the result obtained from the table, we can see that the best max_depth for the regression tree in max_depth is 2 which gives us 7.339 MSE, and based on the regression tree observation we see that failures and schoolsup are the main variables to make upper decision making for G1. From the observation tree regression, the student who has failures 0.5 and schoolsup<=0.5 will obtain high-value marks of around 11.732. This shows that the failures and schoolsup can give effect to our main target variable G1.

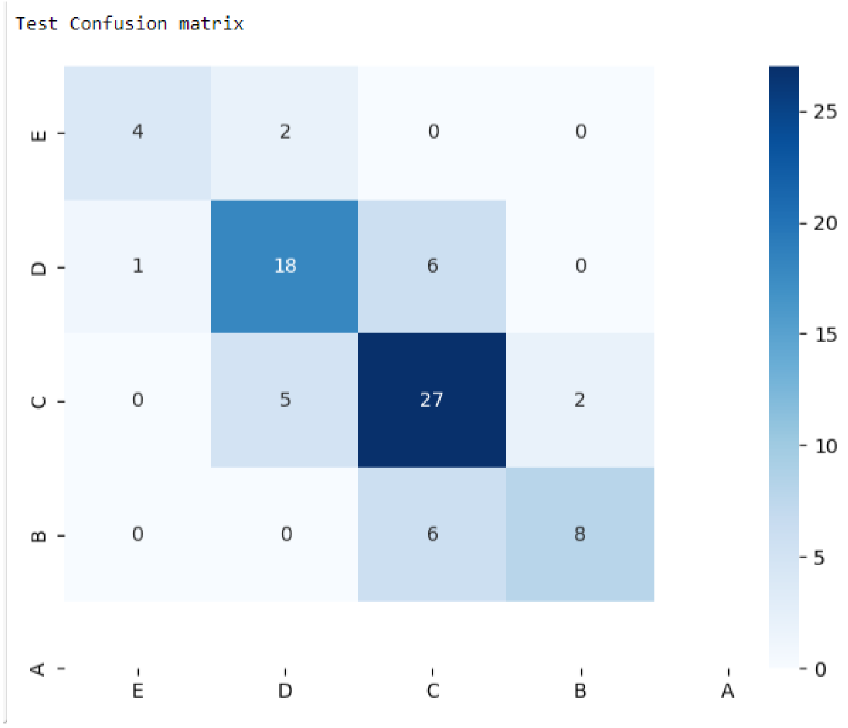
Classification Tree

Max_depth	Accuracy Score
1	0.6708860759493671
2	**0.7215189873417721**
3	0.7088607594936709
4	0.7088607594936709





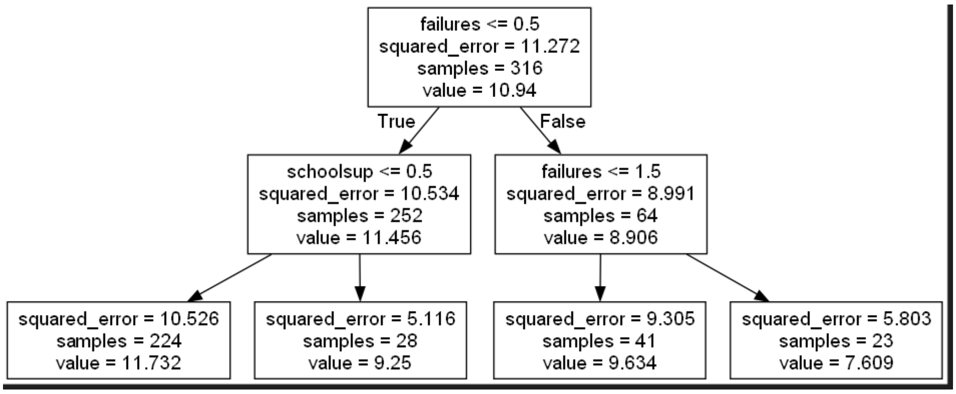
The graph above shows the confusion matrix for Train data



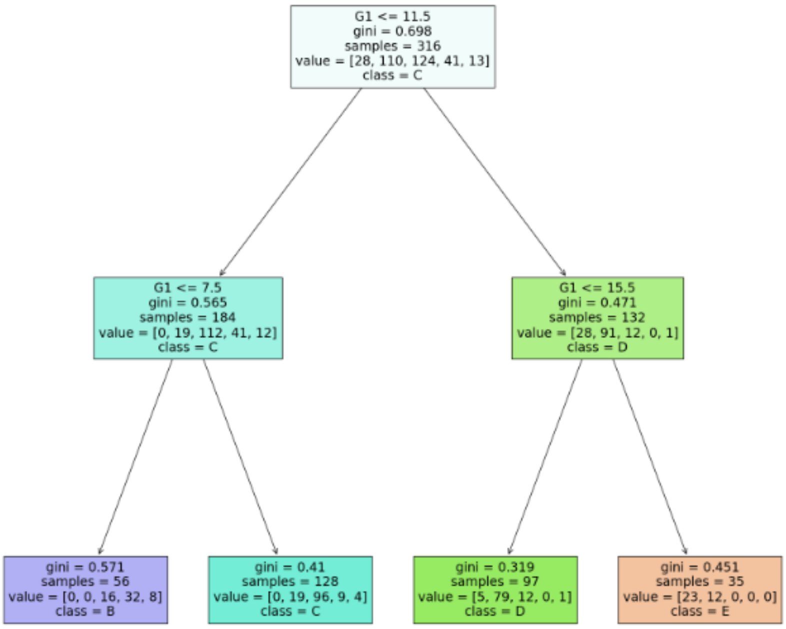
The graph above shows the confusion matrix for Test data

According to the result obtained from the table, we can see that the best max_depth for the classification tree in max_depth is based on the best accuracy score of around 0.7215, and based on the classification tree observation we see that G1 are the main variable to make upper decision making for G2T. From the observation classification regression, the student who has marks below 11.5 and 7.5 in G2 will obtain B grade value marks. This shows that the grade student can give effect to our main target variable G2T.

Classification and Regression Tree



Regression Tree graph with max_depth=2



Classification Tree graph with max_depth=2

Conclusion

In conclusion, the feature importance analysis shows that "failures" and "studytime" is the most crucial variable for G1 and "G1" and "absences" is the most important for G2T after using the random forest method on the dataset. After using max_depth, we find out that the variables which obtain low MSE and better accuracy show "failures" and "school" for G1 and “G1” for G2T. Plus, both of them are using max_depth=2.

This suggests that "G1" for G2T has the greatest influence on the random forest model's overall performance in terms of correctly predicting the target variable. Understanding the importance of "G1" might assist concentrate efforts on maximizing and using this specific characteristic to improve the system's overall performance and the model's ability to predict the future.