**Student Performance Dataset Report**

We are given the student performance datasets for student’s performance in Math and Portuguese, and we must first combine the two to produce one single dataset for analysis. We conduct regression analysis and produce a model using a classification tree to determine which model can help us predict G3 with the highest accuracy. There may be certain drawbacks to using this method. Binding two datasets together may sometimes result in duplicated rows and values, which would reduce the accuracy of our prediction, and this should be considered when getting to a conclusion regarding which model we should use.

We begin by checking for any missing values or duplicate values in the dataset before we proceed with our analysis. In this case, we determine that the data is clean, we conduct Exploratory data analysis. From this we determine that the average grade for Student, G3, is most commonly 10 or 11. The predominant range for G3 is between 8-16. Very few students obtain less than 8 marks, and even fewer obtain 20.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Most students in the sample are between 15 and 18 years old, a very few are older than 18. This indicates, as proven by a graph, that very few students fail and repeat the exams, as there aren’t too many older students.

Chart

Description automatically generated

We conduct correlation analysis and note that there isn’t anything significant produced by it.

Less than 200 students out of the total 1044 students, fail. Approximately 80% of students have access to the internet. We assess how G3 regresses with age and notice that the grade for G3 is likely to be lesser the older a student is. G3 is positively correlated with the amount of time a student spends studying and our analysis also tells us that it is highly likely it is that the student receives lesser total for G3 depending on the number of times the student has failed. The more times the student has fails, the lower the total G3 that student receives. Another conclusion we make is that the higher the education level the mother of a student has, the better the total they receive for G3 is likely to be. There is not much to be determined from the correlation plot we have created that could help us determine the correlation between any of the factors and our target variable G3.

Chart

Description automatically generated with medium confidenceChart, line chart

Description automatically generated

Graphical user interface, chart, line chart

Description automatically generatedGraphical user interface, chart, line chart

Description automatically generated

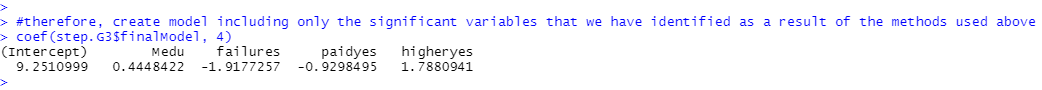
**Linear Regression:**

We proceed with Linear Regression to form a model that help us predict G3 based on certain factors that we identify using stepwise prediction. The variables for the model- MEdu, Paid, Higher and Failures- that are identified through stepwise regression are consistent with some of the conclusions we make based on the exploratory data analysis above. that are the best performing ones when included in the model. We pick the most accurate one and choose that as our model. We use stepwise regression in this case in order to pick a useful subset of predictors from the long list of variables that we have. We define a linear model for G3 and perform stepwise regression on it. Then, to determine accuracy we check what the ‘best tuned’ model is out of all the models produced by it. The model with the lowest ‘Root Mean-Squared Error’ (RMSE) is the most accurate one.

****

Model 4, with RMSE 3.52, the lowest RMSE out of any model we obtain, is the most accurate model we obtain from stepwise regression.

Table

Description automatically generated

The final model we choose would be:

G3= 9.251+0.445MEdu-1.918failures-0.930paid+1.788higher. Here, if there is one additional unit of failures for a given individual, then the value of G3 reduces by 1.918, ceteris paribus.  
Paid is a dummy variable, and in this model, we interpret it as follows: since the coefficient estimate is negative, we understand that students that have paid for extra classes are likely to have a lower G3 than students that have not. Same interpretation goes for the variable ‘higher’.

With an RMSE of 3.521 and a Normalized RMSE of 0.176.

**Classification Tree:**

We create a classification tree model in order to assess whether it can be used to predict G3 based on the factors that are given to us.

Diagram

Description automatically generated

Using this we can predict G3 based on the value of certain factors. We now proceed to assess how accurately we can predict G3 using this model using the testing data set.

Table

Description automatically generated

This method would only help us predict the value of G3 with an accuracy of 8.25%, and hence should not be used

We conclude that the best method to predict G3 is the 4th model that we have produced above, with an RMSE of 3.521, while conducting Linear Regression as it is much more accurate than the classification tree method we did above.