

1. Student name: Homer Ayuste
Student ID: 190658320
Email: ayus8320@mylaurier.ca

2. Student name: Brady Loenhardt
Student ID: 190905340
Email: loen5340@mylaurier.ca

3. Student name: Vinand Panchal
Student ID: 201526670
Email: panc6670@mylaurier.ca

CP322 Project Report

Abstract: The project's objective is to determine whether we can predict secondary school students' final grades based on a dataset containing students' features including demographic, prior grades, and other features. We used 2 different models, including a multiple linear regression model and a decision tree classifier, to predict final grades and found that 8 features out of 32 in the dataset were relevant enough to help predict final grades. The mean absolute error in our first two models were 1.15 and 1.10 which shows that our model was sufficiently accurate in predicting final grades, being only roughly 1 grade point off from students' true grades on average. And in our decision tree model, we achieved an accuracy of 80.67% in predicting what qualification a student will achieve.

Description of Applied Problem:

It will be possible to determine which factors in a student positively or negatively affect their grade so that these factors can be encouraged or discouraged for students. As well as being able to predict a student's final grade based on the information about them and finding what features are relevant enough in aiding our predictions. Based on additional factors that have a high correlation to final grades, we are attempting to predict final grades. As they would have less time to study if they traveled more, we predicted that "traveltime" may be substantially correlated with predicting final grades. We also assumed that first and second-period grades i.e. 'G1' and 'G2' would have the most effects. After assuming that there is a correlation between significant attributes and final grades, we believe that visualizing the relationship between various factors and final grades can be considerably aided by plotting scatterplots. Then, there should be enough information to use the multiple linear regression algorithm to predict final grades that are the closest to the actual final grades to see if our model works.

Description of Available Data:

The data contains many attributes describing a student. Data attributes include student grades, demographic, social, and school-related features. The dataset is about student achievement in mathematics and it was gathered from two Portuguese secondary schools. The target feature in this dataset is G3.

Attribute Information:

- 1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
- 2 sex - student's sex (binary: 'F' - female or 'M' - male)
- 3 age - student's age (numeric: from 15 to 22)
- 4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
- 5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
- 6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
- 7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 "5th to 9th grade", 3 "secondary education" or 4 "higher education")
- 8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 "5th to 9th grade", 3 "secondary education" or 4 "higher education")

- 9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- 10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
- 11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
- 12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
- 13 travelttime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- 14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- 15 failures - number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
- 16 schoolsup - extra educational support (binary: yes or no)
- 17 famsup - family educational support (binary: yes or no)
- 18 paid - extra paid classes within the course subject (Math) (binary: yes or no)
- 19 activities - extra-curricular activities (binary: yes or no)
- 20 nursery - attended nursery school (binary: yes or no)
- 21 higher - wants to take higher education (binary: yes or no)
- 22 internet - Internet access at home (binary: yes or no)
- 23 romantic - with a romantic relationship (binary: yes or no)
- 24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- 25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- 26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- 27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- 28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- 29 health - current health status (numeric: from 1 - very bad to 5 - very good)
- 30 absences - number of school absences (numeric: from 0 to 93)
- # these grades are related with the course subject (in this case, math):
- 31 G1 - first period grade (numeric: from 0 to 20)
- 32 G2 - second period grade (numeric: from 0 to 20)
- 33 G3 - final grade (numeric: from 0 to 20, output target)

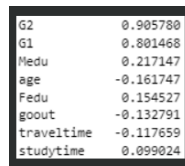
For context, the academic grading in a typical Portuguese high school is a 20-point scale and it be can described as such:

Grade	Qualification
18 - 20	Excellent
16 - 17	Very good
14 - 15	Good
10 - 13	Sufficient
4 - 9	Weak
0 - 3	Poor

Analysis techniques:

We had two options for machine learning: clustering and regression. First, we'll explain our

multiple linear regression model. The code that can demonstrate a correlation between the "G3" and other features was run on the cleaned data. In that, regardless of whether they have a positive or negative impact on "G3" we selected the strongest correlations (absolute values that are closest to 1). Studytime is surprisingly the weakest of the features, with "G2" having the strongest correlation, and we will say that these features are significant enough to include. Scatterplots between "final grades" and each correlated attribute are created. We expect to use the first and second-period grades' significant association, combined with the lesser correlations, to improve our ability to predict final grades. Then, using the data at hand, we used the multiple linear regression code.



G2	0.905788
G1	0.801468
Medu	0.217147
age	-0.161747
Fedu	0.154527
goout	-0.132791
traveltime	-0.117659
studytime	0.099024

Figure: Features with the strongest correlation to G3

-> 1) We implemented a multivariate linear regression model with the following columns: 'G2', 'G1', 'Medu', 'age', 'Fedu', 'goout', 'traveltime', 'studytime'.

2) We then implemented another multivariate linear regression model with the following columns: 'G2', 'G1', 'Medu', 'Fedu' in hopes of simplifying any noise that the weaker correlated variables may have been creating.

3) Lastly, we put into place a decision tree classifier model with the same set of columns as the first linear regression: 'G2', 'G1', 'Medu', 'age', 'Fedu', 'goout', 'traveltime', 'studytime'

-> Model 1) performed well but after further testing with different variables, we came to the conclusion that removing some of the less influential variables would provide a positive outcome on performance. These were the first metrics of model 1):

```
MAE= 1.1521119899343129
MSE= 2.712909102960211
RMSE: 1.6470911034184512
```

Model 2) performed slightly better through the lens of mean absolute error while not sacrificing an insignificant amount of mean squared error and root-mean squared error scores and so we were pleased with this adaptation of the first model:

```
MAE= 1.1008403361344539
MSE= 2.7142857142857144
RMSE: 1.647508942095828
```

Model 3) used a decision tree classifier and achieved an accuracy score of 80.67%, we thought this is indisputably an acceptable model for predicting our target feature. In contrast, we believe its performance is below par with the second model and so it serves as a nice compliment to our findings:

```
Accuracy: 0.8067226890756303
```

In our third model, we used a decision tree classifier to do our classification since we had many features and it would make it easier to break it down into more manageable parts. First, we started with creating a new column with converted G3 into classes/grade ranges based on the Portuguese grading system. Then, we used the same features found from the first linear regression model for features that we will use to predict our target feature. Our classification model contained some incorrect classifications and this may be due to a small dataset. There were very few entries in the tail ends of the classification, especially for the 'Poor', 'Very good', and 'Excellent' classes. However, our predictions never strayed from being more than

one grade range away from the true class as seen in the confusion matrix in the next section. Our worst precision or predictor giving many false positives is in the ‘weak’ and ‘very good’ classes. And the model’s worst recall was also in the ‘very good’ class.

Visualisation techniques:

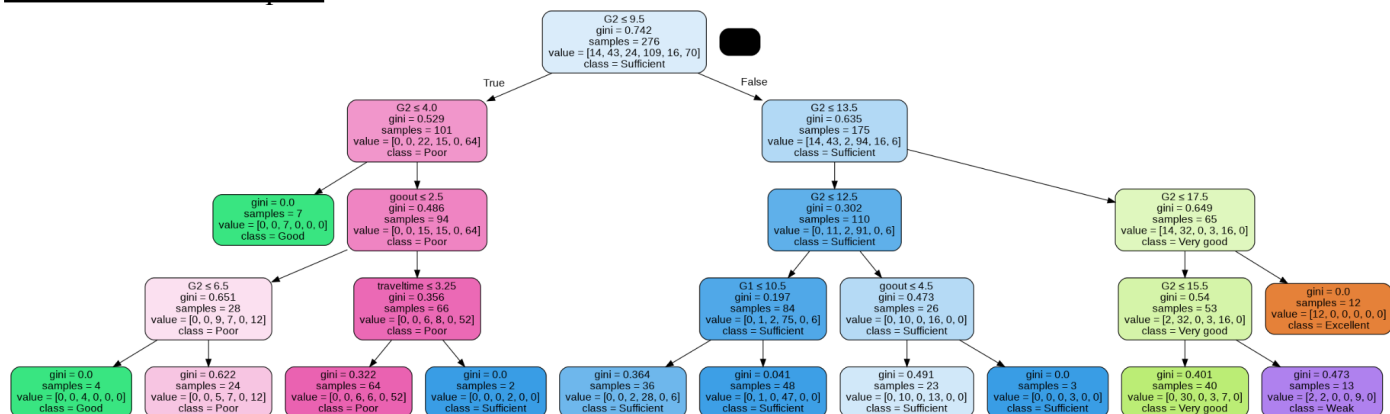


Figure: A decision tree illustrating all of the probable consequences. Value in each node shows how many values are in each class in that node (ie number of excellents in the first element of value, very good is the 2nd element, etc). Thus, at each leaf node we can see how many correct and incorrect predictions there are by looking at value.

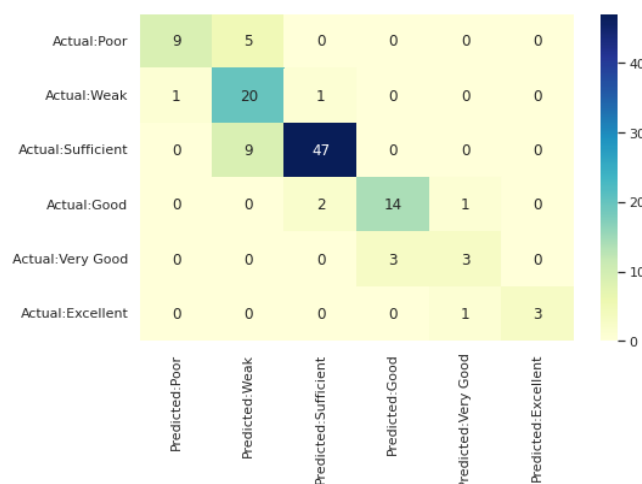


Figure: Confusion Matrix - It is a way to gauge how well a machine learning classification algorithm performs when the output can include two or more classes.

According to this figure, we can see that the numbers for the diagonal of the matrix are the true positives (ie the correct predictions). Additionally, there have been cases of incorrect classification, such as the model predicted weak for 9 entries but the actual class was sufficient, so all indices in the matrix not in the diagonal are the number of incorrect predictions.

	precision	recall	f1-score	support
Poor	0.90	0.64	0.75	14
Weak	0.59	0.91	0.71	22
Sufficient	0.94	0.84	0.89	56
Good	0.82	0.82	0.82	17
Very good	0.60	0.50	0.55	6
Excellent	1.00	0.75	0.86	4
accuracy			0.81	119
macro avg	0.81	0.74	0.76	119
weighted avg	0.84	0.81	0.81	119

Figure: Precision - Additionally known as positive predictive value. The ratio of correct positive predictions to the total predicted positives.

Recall - The ratio of correct positive predictions to the total positives examples.

F1-Score is a metric for the precision of a model on a dataset.

Support - Support counts the number of true values in each class.

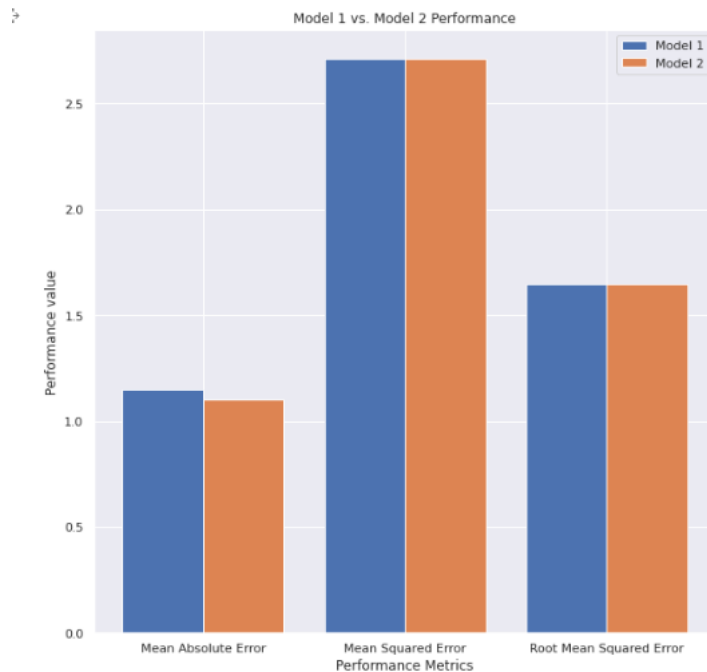


Figure: Performance evaluations of our two models were compared. As you can see, our results from both models are largely comparable with the exception of the category of mean absolute error, which shows some difference. This means that our models are producing reliable predictions.

Reference:

<http://archive.ics.uci.edu/ml/datasets/Student+Performance#>

Appendix:

	Age	Mothers Education	Fathers Education	Travel Time	Study Time	Go Out With Friends	First Period Grades	Second Period Grades	Actual Final Grades	Sklearn Final Grade Predictions
0	18.0	4	4.0	2.0	2.0	4	5	6.0	6	5
1	17.0	1	1.0	1.0	2.0	3	5	5.0	6	3
2	15.0	1	1.0	1.0	2.0	2	7	8.0	10	7
3	15.0	4	2.0	1.0	3.0	2	15	14.0	15	14
4	16.0	3	3.0	1.0	2.0	2	6	10.0	10	9
...
390	20.0	2	2.0	1.0	2.0	4	9	9.0	9	8
391	17.0	3	1.0	2.0	1.0	5	14	16.0	16	17
392	21.0	1	1.0	1.0	1.0	3	10	8.0	7	7
393	18.0	3	2.0	3.0	1.0	1	11	12.0	10	12
394	19.0	1	1.0	1.0	1.0	3	8	9.0	9	8

395 rows x 10 columns