CAB330 Data& Web-Analytics

Case Study 1

Donghyeon Kim n9967273
Jack Teys n9996532
Vladislav Kireyev n9897810

Case Study Scenario

In Michigan, the secondary education consists of 6 years of schooling, preceding 6 years of primary school education. Most of the students join the public and free education system. There are several courses (e.g. Sciences and Technologies, Visual Arts) that share core subjects such as the Spanish Language and Mathematics. A 20-point grading scale is used, where 0 is the lowest grade and 20 is the perfect score. Real values are permitted in the grade. During the school year, students are evaluated in three periods and the last evaluation (G3) corresponds to the final grade(PASS/FAIL). This study will consider data collected during the 2016- 2017 school year from two public schools, from Detroit Central High School(DCHS) and Troy High School(THS). Using surveys student's details about these attributes were collected such as mother's education, family income, social/emotional attributes (e.g. alcohol consumption) and school related (e.g. number of past class failures) variables that were expected to affect student performance. The questionnaire was reviewed by school professionals and tested on a small set of 15 students in order to get a feedback.

The department of education would like to identify which among these high school students will pass or fail. You have been hired as a data analyst consultant by this department. Your task is to inform decision makers the (characteristics of) secondary students using their past school grades (first and second periods), demographic, social and other school related data.

Case Study Dataset

The data set STUDENT contains 1044 observations and 35 variables. Variables are described in Table 1. You would note that some information is presented in multiple ways. This is an example of the presence of redundant variables in a dataset.

The following information would assist you in assigning the variables roles.

- There are three target variables namely, G1, G2 and G3, with different types. Choose the target that suits best according to the given task.
- Identify if the variable is an input variable or a supplementary variable.
- Data transformation is required for a few input variables to get improved accuracy.

Table 1: List of Variables

Attribute	**Description**
Id	student's id
InitialName	student's initial
School	student's school name
Sex	student's sex
Age	student's age
Address	student's home address type
Famsize	family size(≤ 3 or > 3)
Pstatus	parent's cohabitation status (living together or apart)
	mother's education(0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary

Medu	education or 4 – higher education)
Fedu	father’s education(0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Mjob	mother’s job
Fjob	father’s job
Reason	reason to choose this school
guardian	student’s guardian
traveltime	home to school travel time (1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour)
studytime	weekly study time (1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
Failures	number of past class failures(n if 1 ≤ n < 3, else 4)
schoolsup	extra educational school support (yes or no)

Famsup	family educational support (yes or no)
Paid	extra paid classes (yes or no)
activities	extra-curricular activities (yes or no)
Nursery	attended nursery school (yes or no
Higher	wants to take higher education (yes or no)
Internet	Internet access at home (yes or no)
romantic	with a romantic relationship (yes or no)
Famrel	quality of family relationships (1 – very bad to 5 – excellent)
freetime	free time after school (1 – very low to 5 – very high)
Gout	going out with friends (1 – very low to 5 – very high)
Dalc	workday alcohol consumption (1 – very low to 5 – very high)
Walc	weekend alcohol consumption (1 – very low to 5 – very high)
Health	current health status (1 – very bad to 5 – very good)
absences	number of school absences (0 to 75)
G1	first period grade (0 to 20)
G2	second period grade (0 to 20)
G3	Final result(PASS/FAIL)

Case Study Tasks

Your task is to build various predictive models such as decision tree, regression function, and neural network on this data set and compare them. Results inferred by these models should inform decision makers the (characteristics of) students performance.

Set up a new project for this task with **DMProj1** as the Python file and **STUDENT** as the dataset. Include various models in this source file. Name all the models meaningfully.

Task 1. Data Selection and Distribution. (4 marks)

1. What is the proportion of students who will pass?

The proportion of the students that will pass is **661/383** or **1.73**

2. Did you have to fix any data quality problems? Detail them.

Apply imputation method(s) to the variable(s) that need it. List the variables that needed it. Justify your choice of imputation if needed.

In the STUDENT dataset variables “school”, “age”, “G1” and “G2” had missing values. These missing values were eliminated by imputation using .fillna() function. For the variable “school” NaN values were replaced with the previous valid valid using ‘ffil’ method of .fillna() function. For the other columns requiring imputation NaN values were replaced with the mean of the corresponding variables using .mean() function. The mean was also rounded according to the the way values in the column were presented. Mean of the “age” was rounded to a full number, while “G1” and “G2” was rounded to one decimal place.

When a variable can only has two values, it is prefered that they are represented in a 0/1 binary fashion. It occurs in the columns: “address”, “famsize”, “Pstatus”, “schoolsup”, “famsup”, “paid”, “activities”, “nursery”, “higher”, “internet”, “romantic”, “G3”. The values of these variables were replaced with ones and zeros using .map() function.

3. What variables did you include in the analysis and what were their roles and measurement level set? Justify your choice.

traveltime	**Role:** Independent variable **Measurement level:** Qualitative – Ordinal (1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour) + Travel time may affect student’s academic result as travel time increases, their study time might be affected.
studytime	**Role:** Independent variable **Measurement level**: Quantitative – Ordinal (1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours) + Lack of study time can reflect the laziness or low motivation of the student, not having enough study time could result to failure.
failures	**Role:** Independent variable **Measurement level:** Interval (n if 1 ≤ n < 3, else 4) + Past failure history might indicate, that the student could have a higher chance of failing compared to other students that haven’t failed.
schoolsup	**Role:** Independent variable **Measurement level**: Qualitative – Asymmetric Binary (yes or no) + If students take extra educational support, It could mean they really want to pass, lowering the chance of failing.
famsup	**Role:** Independent variable **Measurement level**: Qualitative – Asymmetric Binary (yes or no) + Similar to school support but parents probably knows the student better personally, giving unique learning support.

Paid	**Role:** Independent variable **Measurement level**: Qualitative – Asymmetric Binary (yes or no) + More professional support compared to ‘schoolsup’ and ‘familysup’
Activities	**Role:** Independent variable **Measurement level**: Qualitative – Asymmetric Binary (yes or no) + Shows more participation in willing to learn.
Higher	**Role:** Independent variable **Measurement level**: Qualitative – Asymmetric Binary (yes or no) + Has a goal / motivation / more willingness to pass the subjects in school, yes == higher chance of pass.
Internet	**Role:** Independent variable **Measurement level**: Qualitative – Asymmetric Binary (yes or no) + Most of the learning resources are obtained from the internet, no == loss of valuable resources -> higher fail rate.
Famrel	**Role**: Independent variable **Measurement level**: Quantitative – Interval (1 – very bad to 5 – very good) + Bad family relationship may mentally affect the student’s ability to study.
Freetime	**Role:** Independent variable **Measurement level**: Quantitative – Interval (1 – very low to 5 – very high) + Free time combined with study time could reveal the student’s willingness to pass. (high free time + low study time = lazy or low free time + low study time = busy)
Goout	**Role:** Independent variable **Measurement level**: Quantitative – Interval

	(1 – very low to 5 – very high) + Higher ‘going out with freidns’ with low study time, could indicate, the student is spending too much time playing rather than studying.
Dalc	**Role:** Independent variable **Measurement level**: Quantitative – Interval (1 – very low to 5 – very high) + High alcohol consumption rate could affect the student’s ability to study / focus.
Walc	**Role:** Independent variable **Measurement level**: Quantitative – Interval (1 – very low to 5 – very high) + High alcohol consumption rate could affect the student’s ability to study / focus.
Health	**Role:** Independent variable **Measurement level**: Quantitative – Interval (1 – very bad to 5 – very good) + Poor health might mean, student might have special conditions that could prevent student from studying / focusing.
Absence	**Role:** Independent variable **Measurement level**: Quantitative – Interval (0 to 75) + high absence may show student’s low motivation to learn and/or behind schedule of learning materials, leading to higher fail rate.
G3	**Role**: Dependent variable **Measurement level**: Qualitative – Asymmetric binary (Pass or Fail) + This is the main variable which we want to predict the value of provided with given independent values.

4. What distribution scheme did you use? What “data partitioning allocation” did you set? Explain your selection. (Hint: Take the lead from Week 2 lecture on data distribution)

For the distribution of testing and training data, we used a **70/30** distribution scheme to allow for an adequate allocation of testing data, while at the same time not limiting the training data or training the model using the entire data set.

Task 2. Predictive Modeling Using Decision Trees (4 marks)

1. Build a decision tree using the default setting. Examine the tree results and answer the followings:

   a. What is classification accuracy on training and test datasets?

i. The decision tree managed a training accuracy of 0.993 and testing accuracy of 0.597.
ii. Classification accuracy

	precision	recall	f1-score	support
0	0.46	0.53	0.49	96
1	0.70	0.64	0.67	165
avg/total	0.61	0.60	0.60	261

   b. List the decision rules.

The decision tree uses the 'failure' variable as part of it's first decision in which it uses "True" and "False" for a conditional "<= 0.5". The two decision rules branching from the 'failure' decision include the use of "absences <= 23.0" and "higher <= 0.5".

   c. What are the 5 important variables in building the tree?

Failures, absences, freetime, studytime, and famrel

   d. Report if you see any evidence of model overfitting.

The model had a training accuracy of 0.993 and test accuracy of 0.597. This would indicate that it overfits the dataset rather badly.

2. Build another decision tree tuned with GridSearchCV. Examine the tree results.

    a. What is classification accuracy on training and test datasets?

i. The GridSearchCV decision tree had a better test/training accuracy of 0.743 and 0.747 for the ii. training and test accuracies.
ii. Classification accuracy:

	precision	recall	f1-score	support
0	0.77	0.45	0.57	96
1	0.74	0.92	0.82	165
avg/total	0.75	0.75	0.73	261

    b. What are the parameters used? Explain your decision.

The three important parameters used for the GSCV were:

- criterion: 'gini'
- max_depth: range of 2 to 7
- min_samples_leaf: range of 20 to 60 with interval of 10

    c. What are the optimal parameters for this decision tree?

The three optimal paramaters used for the GSCV were:

a. criterion: 'gini'
b. max_depth: 3
c. min_samples_leaf: 20

    d. Which variable is used for the first split? What are the competing splits for this first split?

The 'failures' variable is used for the first split, for this split. The failure value was split at 0.5 (between 0 and 1).

    e. What are the 5 important variables in building the tree?

Failures, higher, absences, famsup, and health are the top five variables used in the decision tree.

f. Report if you see any evidence of model overfitting.

The model has a rather close score for training/testing accuracies, and as such doesn't over fit.

3. What is the significant difference do you see between these two decision tree models? How do they compare performance-wise? Explain why those changes may have happened.

A large difference concerns the difference in accuracy scores (decision tree: 0.395913155, CV: 0.00383141762). The GridSearchCV also eliminated a large number of non-essential data and left the model with the four remaining variables: failures, higher, absences, and famsup.

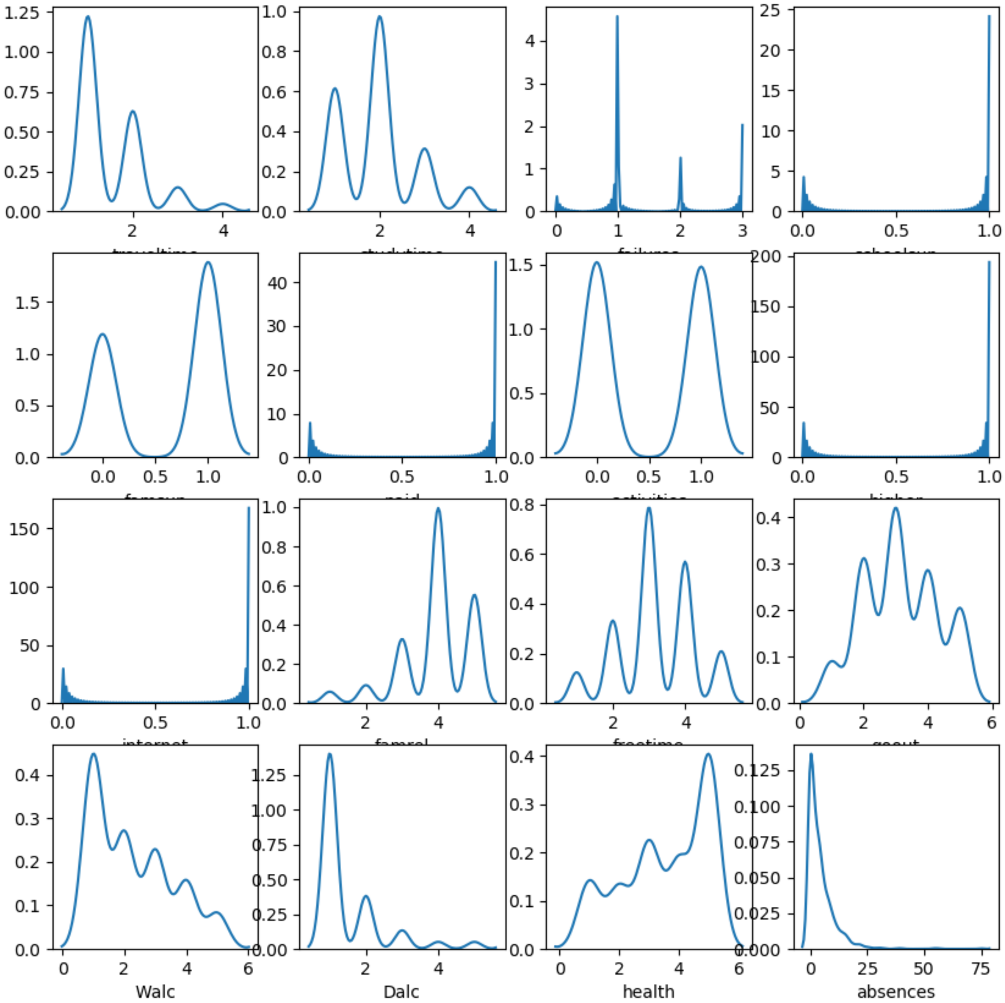
4. From the better model, can you identify which students to target for further consultation? Can you provide some descriptive summary of those students?

Students that don't intent on pursuing a higher education and have an absence count above nine would be the primary targets for further consultation as displayed by the decision tree, students who have these traits have the highest G3 failure rate (77 samples).

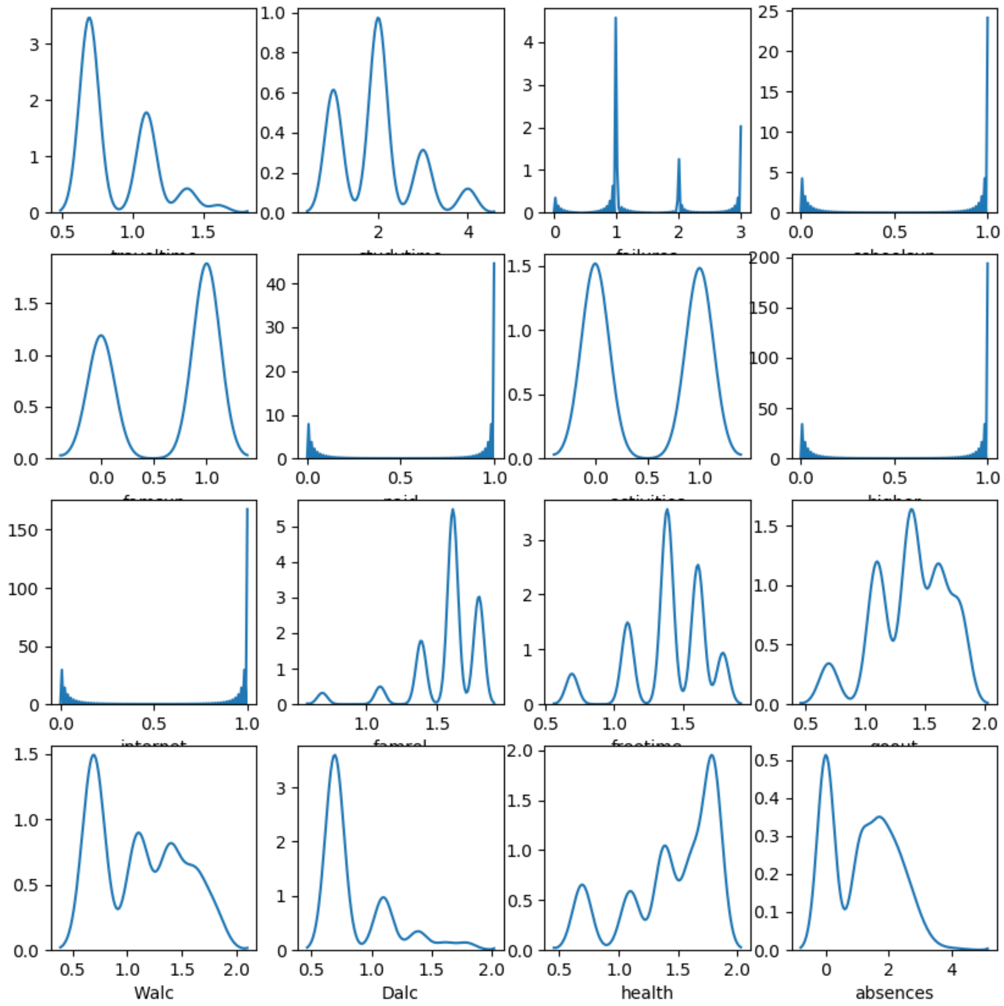
Task 3. Predictive Modeling Using Regression (5.5 marks)

1. In preparation for regression, apply transformation method(s) to the variable(s) that need it. List the variables that needed it.

After plotting the distribution plots (Figure 1) of the columns, used in the model, with custom plotCol() function, it was seen that variables 'traveltime', 'famrel', 'freetime', 'goout', 'Walc', 'Dalc', 'health' and 'absences' had distribution skewed toward minimum and maximum values. Because of this, it was decided to transform them, which resulted in these columns being better distributed, as seen in the produced distribution plots (Figure 2).



(Figure 1)



(Figure2)

2. Build a regression model using the default regression method with all inputs. Once you done it, build another one and tune it using GridSearchCV. Answer the followings:
    a. Report which variables are included in the regression model.

Failures, activities, Walc, Dalc, goout, higher, health, studytime, freetime, absences, famrel, paid, internet, traveltime, famsup and schoolsup variables were included into the regression model.

b. Report the top-5 important variables (in the order) in the model.

Failures, higher, Walk, traveltime and studytime are the top-5 variables in the model. Higher and studytime have positive coefficients, while failures, Walk and traveltime have negative coefficients.

c. Report any sign of overfitting.

Overfitting occurs when training accuracy of the model is higher than testing accuracy. The first model's train accuracy of 0.7465753424657534 is slightly lower than the test accuracy of 0.7484076433121019, which means that no overfitting has occurred. However, the overfitting can be observed in the model tuned with Search Grid CV, where the train accuracy 0.7465753424657534 is higher than testing accuracy 0.7292993630573248.

d. What are the parameters used? Explain your decision. What are the optimal parameters? Which regression function is being used?

Grid search CV was used to obtain an optimal parameters for the model. The parameter C, required for the regression was returned as 0.01. It did not produced any more improvement to the accuracy of the model. The Logistic regression was used here, because the target variable is of categorical, not continuous nature, which requires the prediction of the probability of occurrence of the value, not the value itself.

e. What is classification accuracy on training and test datasets?

The train accuracy of the first model was was equal to 0.7465753424657534 and the test accuracy was 0.7484076433121019, while the classification accuracy was as following:

	precision	recall	f-1 score	support
0	0.75	0.47	0.58	115
1	0.75	0.91	0.82	199
avg / total	0.75	0.75	0.73	314

The train accuracy of the model tuned with Grid Search CV was was equal to 0.7465753424657534 and the test accuracy was 0.7292993630573248, while the classification accuracy was as following:

	precision	recall	f-1 score	support
0	0.72	0.43	0.54	115
1	0.73	0.90	0.81	199
avg / total	0.73	0.73	0.71	314

3. Build another regression model using the subset of inputs selected by RFE and selection by model methods. Answer the followings:

    a. Report which variables are included in the regression model.

    After performing recursive feature elimination (RFE) the original number of features, which was 16, was decreased to 2. The variables, included in the produced regression model, were studytime and traveltime.

    b. Report the top-5 important variables (in the order) in the model.

    The most important variable in the model was traveltime with a negative coefficient, followed by studytime with a positive coefficient.

    c. Report any sign of overfitting.

    Overfitting has occurred in this model, because the train accuracy of 0.7452054794520548 is slightly higher than the test accuracy of 0.7420382165605095.

d. What is classification accuracy on training and test datasets?

The train accuracy was was equal to 0.7452054794520548 and the test accuracy was 0.7420382165605095, while the classification accuracy was as following:

	precision	recall	f-1 score	support
0	0.74	0.46	0.57	115
1	0.74	0.90	0.82	199
avg / total	0.74	0.74	0.72	314

4. Using the comparison statistics, which of the regression models appears to be better? Is there any difference between two models (i.e one with selected variables and another with all variables)? Explain why those changes may have happened.

From the analysis of the testing accuracy of the produced models it can be concluded that the model first model is the best, because it has the highest testing accuracy, even though tuning it with Grid Search CV has not improved it’s predictive capability. The model that used recursive feature elimination lost both training and testing accuracy in the comparison with the original model.

5. From the better model, can you identify which students to target? Can you provide some descriptive summary of those students?

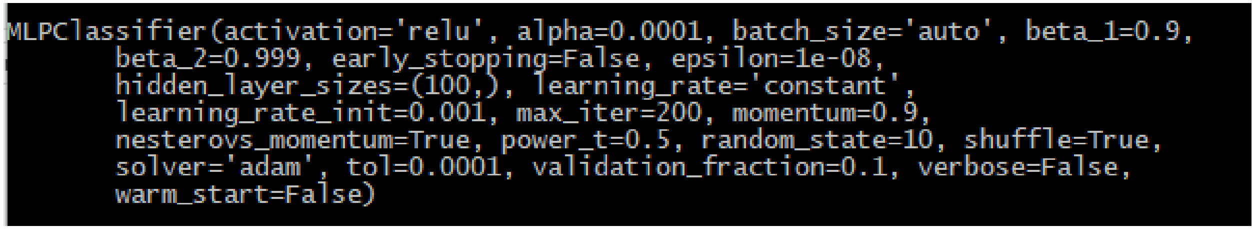
Failures is the top important variable in the model, however, it’s relation to the target variable is clear and does not provide any interesting insight of what affects the likelihood of student to pass or fail. Higher and Wal are the following two important variables, and they, on the other hand, could provide interesting results. Students who are interested in higher education seem to be more likely to pass, while those with higher weekend alcohol consumption are more likely to fail. These two variables could be thoroughly monitored to provide further insight between these relations.

Task 4. Predictive Modeling Using Neural Networks (5.5 marks)

1. Build a Neural Network model using the default setting. Answer the following:

a. What are the parameters used? Explain your decision. What is the optimal network architecture?

First model was built on default parameters. Such as



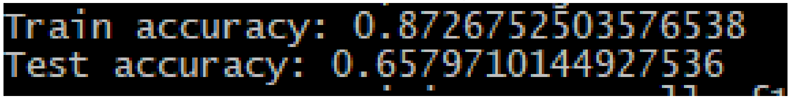
```
MLPClassifier(activation='relu', alpha=0.0001, batch_size='auto', beta_1=0.9,
       beta_2=0.999, early_stopping=False, epsilon=1e-08,
       hidden_layer_sizes=(100,), learning_rate='constant',
       learning_rate_init=0.001, max_iter=200, momentum=0.9,
       nesterovs_momentum=True, power_t=0.5, random_state=10, shuffle=True,
       solver='adam', tol=0.0001, validation_fraction=0.1, verbose=False,
       warm_start=False)
```

b. How many iterations are needed to train this network?

Since the model was built and tested on the default **MLPClassifier** settings, ***200 epochs (iterations)*** was used to train this network

c. Do you see any sign of over-fitting?

There is a **major overfitting** issue in this model. As seen in the picture below, there is a significant accuracy difference of around 0.22. When the train accuracy gives distinct difference compared to the test accuracy, it is most likely saying, the learning algorithm has failed to generalise on the data inputs and has found relationships between input and target variable, outputting high training accuracy, however, when it came to test data (which does not display its target values), it will perform poorly.



```
Train accuracy: 0.8726752503576538
Test accuracy: 0.6579710144927536
```

d. Did the training process converge and resulted in the best model?

As this is the first model produced with default hyperparameters, we cannot conclude this is the best model. However, from what we can see from the Train / Test accuracy values, this model is poorly performing, due to overfitting.

e. What is classification accuracy on training and test datasets?

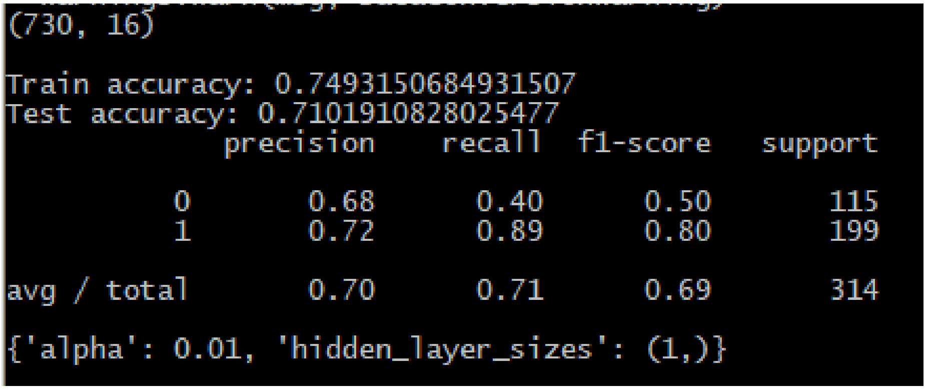
The classification accuracy on training and test datasets are;
+ Train accuracy: 0.8808219178082192
+ Test Accuracy: 0.6624203821656051

	precision	recall	f1-score	support
0	0.54	0.50	0.52	115
1	0.72	0.76	0.74	199
avg / total	0.66	0.66	0.66	314

2. Refine this network by tuning it with GridSearchCV. Report the trained model, same as Task 4.1

a. What are the parameters used? Explain your decision. What is the optimal network architecture?

Using the GridSearchCV library, **“max_iter(max iterations or epoch)”**, **“hidden_layer_sizes (number of neurons in each hidden layer)”** and **“alpha (regularsation parameter)”** was tuned to optimse the neural network’s model.
Its optimal values were listed below (also shown in photo below)
+ max_iter = 300+
+ hidden_layer_sizes = 1
+ alpha = 0.01
It is analysed that, less complex models (smaller feature sets) tend to generalise on this dataset, therefore, hidden layer sizes were chosen as 1, which it was preformed best at.
‘Alpha’ is a learning rate for gradient descent algorithm (process of going back and forward and modifying the weights). Larger the alpha, bigger steps and faster it trains and vice versa. GridSearhCV has returned 0.01 as the optimal value for ‘alpha’



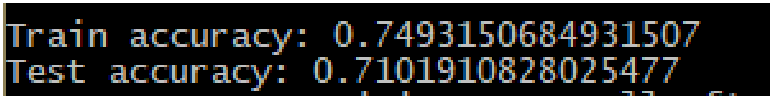
```
(730, 16)

Train accuracy: 0.7493150684931507
Test accuracy: 0.7101910828025477
             precision    recall  f1-score   support

          0       0.68      0.40      0.50       115
          1       0.72      0.89      0.80       199

avg / total       0.70      0.71      0.69       314

{'alpha': 0.01, 'hidden_layer_sizes': (1,)}
```

b. How many iterations are needed to train this network?

As shown in the picture above, approximately **300 epochs** (iterations) are required to optimally train this network.

c. Do you see any sign of over-fitting?

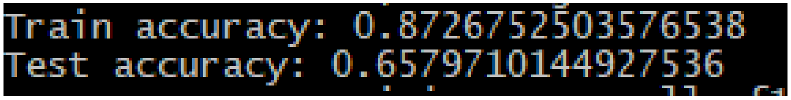
As seen in the picture below, there is a slight sign of over-fitting.



```
Train accuracy: 0.7493150684931507
Test accuracy: 0.7101910828025477
```

d. Did the training process converge and resulted in the best model?

Compared to the previous model (picture below), there is a remarkable improvement in the test accuracy. Even though Train accuracy has 'reduced', big value difference between train and test accuracy signifies overfitting of the models. Therefore, current model tuned with GridSearchCV is significantly better compared to the default model.



```
Train accuracy: 0.8726752503576538
Test accuracy: 0.6579710144927536
```

e. What is classification accuracy on training and test datasets?

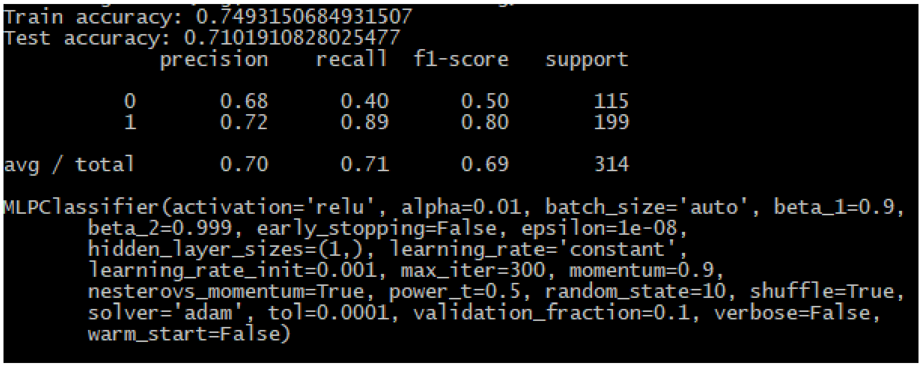
The classification accuracy on training and test datasets are;
+ Train accuracy: 0.7493150684931507
+ Test Accuracy: 0.7101910828025477

	precision	**recall**	**f1-score**	**support**
0	0.68	0.40	0.50	115
1	0.72	0.89	0.80	199
avg / total	0.70	0.71	0.69	314

3. Build another Neural Network model with inputs selected from RFE with regression (use the best model generated in Task 3) and selection with decision tree (use the best model from Task 2). Answer the following:

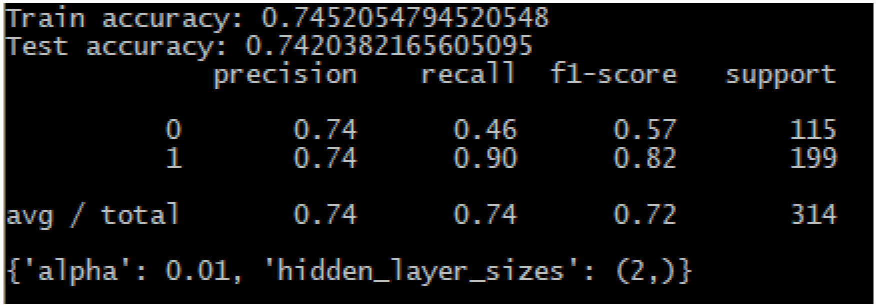
a. Did feature selection help here? Any change in the network architecture? What inputs are being used as the network input?

In the model tuned with GridSearchCV, there was a big improvement from the default model with a slight sign of overfitting. Using RFE, the decision tree's feature selection did help slightly, however, using RFE with regression displayed better results.
Overall, using Recursive Feature Elimination has improved accuracy notably and most importantly, RFE with Logistic Regression shows close to no overfitting with well improved accuracy.
The difference in neural network architecture in all three models was 'hidden_layer_sizes' or number of inputs in the hidden node. GridSearchCV showed optimal performance with at 1, RFE with Logistic Regression at 2, and RFE Feature selection with Decision tree at 3.



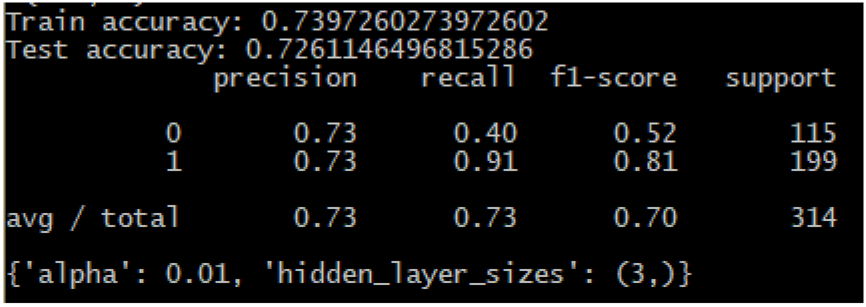
```
Train accuracy: 0.7493150684931507
Test accuracy: 0.7101910828025477
             precision    recall  f1-score   support

          0       0.68      0.40      0.50       115
          1       0.72      0.89      0.80       199

avg / total       0.70      0.71      0.69       314

MLPClassifier(activation='relu', alpha=0.01, batch_size='auto', beta_1=0.9,
       beta_2=0.999, early_stopping=False, epsilon=1e-08,
       hidden_layer_sizes=(1,), learning_rate='constant',
       learning_rate_init=0.001, max_iter=300, momentum=0.9,
       nesterovs_momentum=True, power_t=0.5, random_state=10, shuffle=True,
       solver='adam', tol=0.0001, validation_fraction=0.1, verbose=False,
       warm_start=False)
```

GridSearchCV only



```
Train accuracy: 0.7452054794520548
Test accuracy: 0.7420382165605095
             precision    recall  f1-score   support

          0       0.74      0.46      0.57       115
          1       0.74      0.90      0.82       199

avg / total       0.74      0.74      0.72       314

{'alpha': 0.01, 'hidden_layer_sizes': (2,)}
```

Logistic Regression RFE



```
Train accuracy: 0.7397260273972602
Test accuracy: 0.7261146496815286
             precision    recall  f1-score   support

          0       0.73      0.40      0.52       115
          1       0.73      0.91      0.81       199

avg / total       0.73      0.73      0.70       314

{'alpha': 0.01, 'hidden_layer_sizes': (3,)}
```

Decision Tree RFE

b. What is classification accuracy on training and test datasets? Is there any improvement in the outcome?

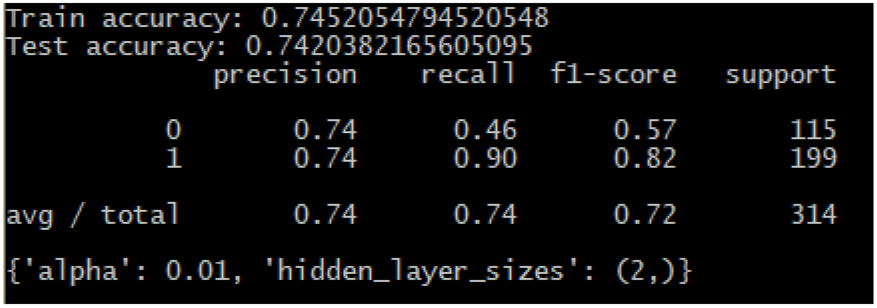
As explained above, there was a notable improvement on accuracy compared to the last model tuned with GridSearchCV with a difference of 0.03, and compared to the default model, a significant difference of 0.08.

c. How many iterations are now needed to train this network?

800 iterations are now required to train the network.

d. Do you see any sign of over-fitting?

As seen in the picture below, the accuracy score displays no overfitting signs from this model.



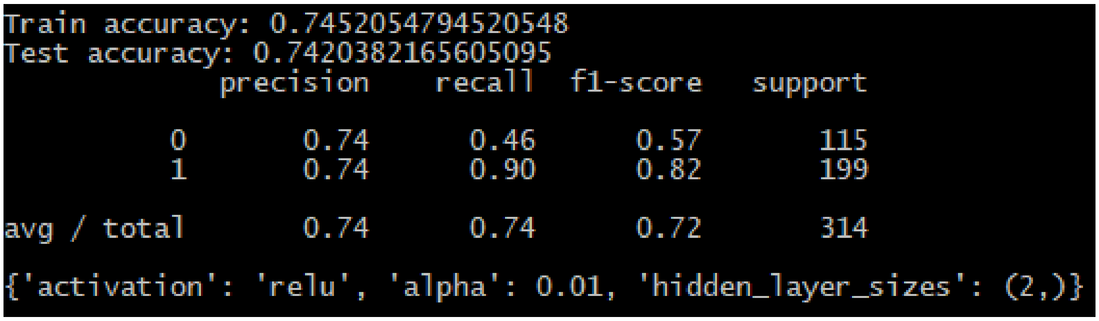
```
Train accuracy: 0.7452054794520548
Test accuracy: 0.7420382165605095
             precision    recall  f1-score   support

          0       0.74      0.46      0.57       115
          1       0.74      0.90      0.82       199

avg / total       0.74      0.74      0.72       314

{'alpha': 0.01, 'hidden_layer_sizes': (2,)}
```

e. Did the training process converge and resulted in the best model?

Combining GridSearchCV and dimensionality reduction with recursive feature elimination using logistic regression has significantly improved its accuracy and removed overfitting. Therefore, it can be concluded that this model is the best model out of previous methods.

f. Finally, see whether the change in network architecture can further improve the performance, use GridSearchCV to tune the network. Report if there was any improvement.

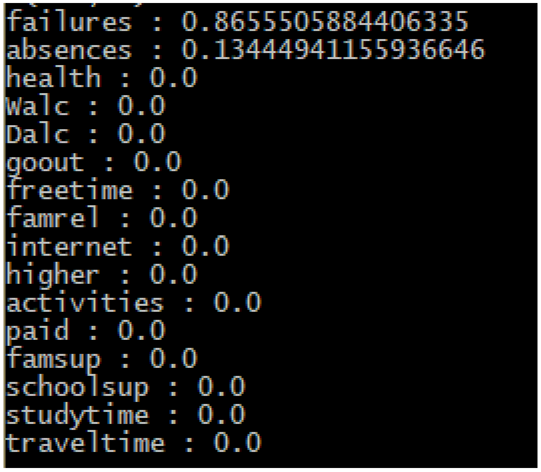
Best performing model was attempted to tune ‘activation’ using the parameter ‘relu(current version)’, ‘identity’, ‘tanh’ and ‘logistic’ for further optimization.
The model’s accuracy showed no improvements



```
Train accuracy: 0.7452054794520548
Test accuracy: 0.7420382165605095
             precision    recall  f1-score   support

          0       0.74      0.46      0.57       115
          1       0.74      0.90      0.82       199

avg / total       0.74      0.74      0.72       314

{'activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (2,)}
```

4. Using the comparison methods, which of the models (i.e one with selected variables and another with all variables) appears to be better? From the better model, can you identify which students to target? Can you provide some descriptive summary of those students?

Using the accuracy comparison method (first picture below), it is seen that GridSearchCV refined logarithmic regression with selected variables using Recursive Feature Elimination applied to the Neural networks classifier worked the best.
From this model, using the 'feature importance' function, it has displayed in descending order of the input variables that were deemed the most important during the decision process, it is seen that two input variable, 'failures' and 'absences' were the only variables that were used in the classifier and 'failures' with the importance of 0.87 and 'absences' with the importance of 0.13. From this feature importance analysis and this optimal classification model, we can conclude students with **previous failure history** and also students with **high absence history** are most likely to receive an overall grade of **fail**.



```
failures : 0.8655505884406335
absences : 0.13444941155936646
health : 0.0
Walc : 0.0
Dalc : 0.0
goout : 0.0
freetime : 0.0
famrel : 0.0
internet : 0.0
higher : 0.0
activities : 0.0
paid : 0.0
famsup : 0.0
schoolsup : 0.0
studytime : 0.0
traveltime : 0.0
```

Task 5. Comparing Predictive Models (4 marks)

1. Use the comparison methods to compare the best decision tree model, the best regression model and the best neural network model.

   a. Discuss the findings led by (a) ROC Chart and Index; (b) Accuracy Score; (c) Classification Report.

   From ROC index score and the ROC chart, it is seen that Logistic Regression model performs best under the varied threshold values, followed by Neural Networks and Decision Tree, as larger curve compared to other models display best performing model overall (or the biggest ROC index value) by having higher true positive rate and lowest false positive rate.

   However, from the comparison of the accuracy value from the three best models of decision tree, linear regression and neural networks, it is shown that Decision tree has the highest test accuracy value of 0.747, followed by linear regression of 0.742 and finally, neural networks of 0.735. As seen by the values, there are not much accuracy difference between each models.

   From the Classification report, we can see all three of the model has returned similar precision, recall and f1-score, varying from 0.74 – 0.75 (precision), 0.74 – 0.75 (recall) and 0.72 – 0.73 (f1-score). Precision defines how many of the times the model has successfully identified the relevant answer (in this case 0 or 1), recall defines, from those relevant answers identified in precision, how many were actually correct, and f1-score averages those two results out by finding the average of precision and recall.

   b. Do all the models agree on the students' characteristics? How do they vary?

   Decision tree suggests; Failures, higher, absences, famsup, and health are the top five characteristics that results in overall grade of fail.
   Which can be concluded into these categories where, students that don't intent on pursuing a higher education and have an absence count above nine would be the primary targets for further consultation as displayed by the decision tree, students who have these traits have the highest G3 failure rate (77 samples).

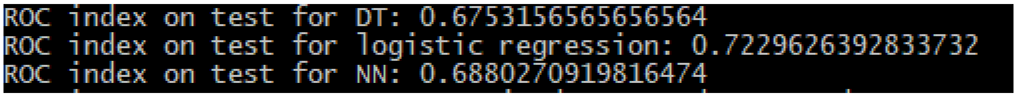
Linear regression suggests; failures as the top important variable in the model, however, it's relation to the target variable is clear and does not provide any interesting insight of what affects the likelihood of student to pass or fail. Higher and Wal are the following two important variables, and they, on the other hand, could provide interesting results. Students who are interested in higher education seem to be more likely to pass, while those with higher weekend alcohol consumption are more likely to fail. These two variables could be thoroughly monitored to provide further insight between these relations.

Finally, Neural networks suggests; failure as the most important factor of 0.87 and followed by absences of 0.133, these conclude, previous failure history and also students with high absence history are most likely to receive an overall grade of fail.

In conclusion, three models all agree on the students who had previous failure history and absence history to receive a final grade of fail.

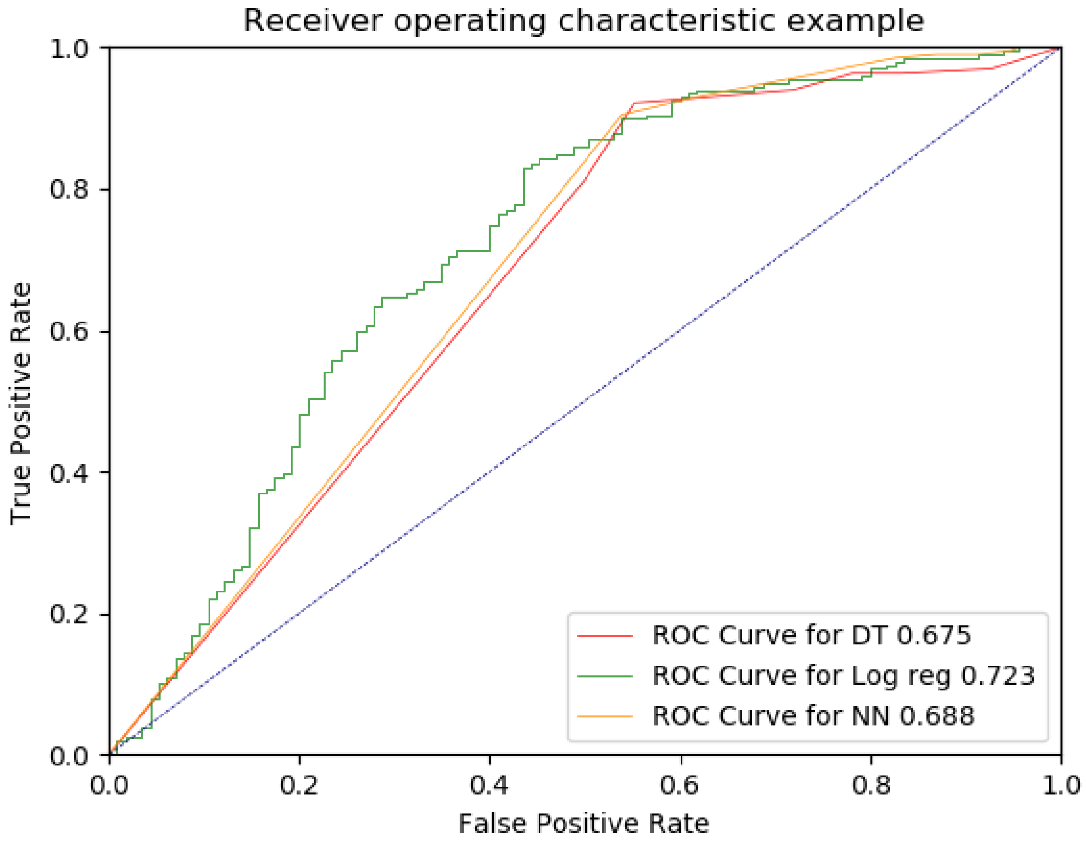
2. Summarise your findings and present the results in a table.

ROC Index Scores

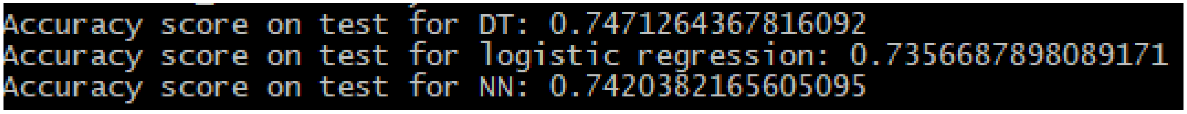


```
ROC index on test for DT: 0.6753156565656564
ROC index on test for logistic regression: 0.7229626392833732
ROC index on test for NN: 0.6880270919816474
```

ROC Curve

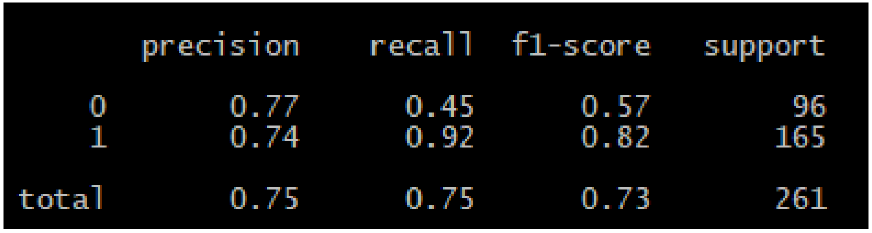


Accuracy Score



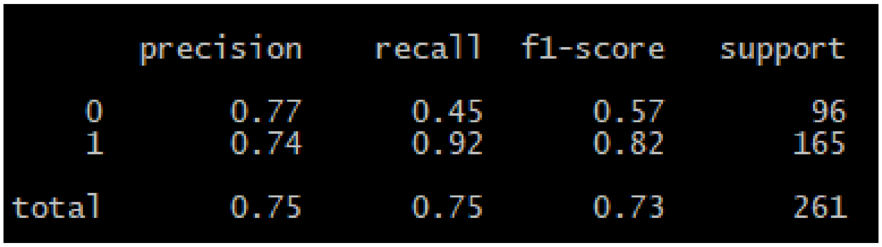
```
Accuracy score on test for DT: 0.7471264367816092
Accuracy score on test for logistic regression: 0.7356687898089171
Accuracy score on test for NN: 0.7420382165605095
```

Decision Tree's Classification Report



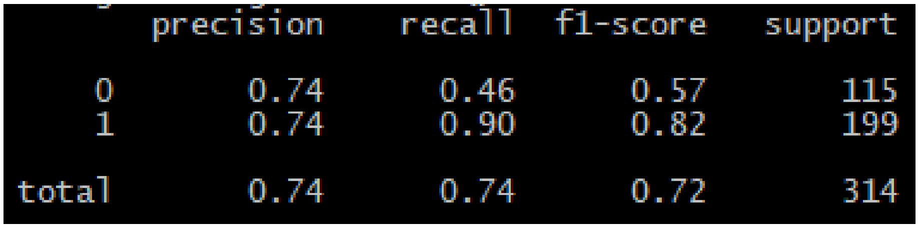
```
          precision    recall  f1-score   support

     0         0.77      0.45      0.57        96
     1         0.74      0.92      0.82       165

 total         0.75      0.75      0.73       261
```

Logistic Regression's Classification Report



```
       precision    recall  f1-score   support

   0       0.77      0.45      0.57        96
   1       0.74      0.92      0.82       165

total      0.75      0.75      0.73       261
```

Neural Network's Classification Report



```
       precision    recall  f1-score   support

   0       0.74      0.46      0.57       115
   1       0.74      0.90      0.82       199

total      0.74      0.74      0.72       314
```

3. Finally, based on all models and analysis, is there a particular model you will use in decision making? Justify your choice.

Based on all model's analysis, it is recommended that the decision makers utilise the linear regression's model to justify whether the targeted student is most likely to fail. While decision tree may have the highest accuracy value, it is only difference of 0.01 and from the ROC curve and the index score suggests, it has the highest true positive rate with lowest false positive rate, resulting in high precision overall.

4. How the outcome of this study can be used by decision makers? Can you summarise positives and negatives of each predictive modelling method based on this analysis?

This study can be utilised by decision makers, as it explains which students with certain characteristics or previous history to target. I recommend the decision maker to use Decision tree model and linear regression model when coming to final decisions. As Linear regression model will provide high assurance of accuracy and precision, where decision tree can be used to justify the reasoning behind the action of why the specific student was chosen. Reason why Neural Networks was dis-encouraged by the decision maker was, Neural networks excel in clustering problems (where it outputs multiple answers) and within this study, it has shown to perform poorer compared to the decision tree model (which is strong at explaining the reasoning behind the choice) and the linear regression model.

Assignment 1 Criteria Sheet:

Criteria	Comments and scoring
Non Submission of all components/ evidence of plagiarism	0
Has demonstrated a task with a working model with /without submission and demonstrates the ability to run the program and add some components. Questions were poorly answered.	1-5
Has demonstrated a task with a working model having a data source, and diagram with the substantial but incorrect implementation of at least one of the three components (predictive models). Questions were poorly answered.	6-9
Has implemented models for all three tasks (three data mining algorithms) with at least one being substantially correct. Shows some understanding of concepts with some success applying knowledge in basic questions	10-13
Has implemented models for all three tasks: Two of the three tasks are fundamentally correct, with substantially correct work flow diagrams which may contain minor errors. Response to questions shows a fundamental understanding of terms and concepts.	14-17
Has fundamentally correct implementation of all five tasks i.e. correct allocations of a target, rejections of variables according to instructions, running three models and comparing them. Includes a demonstration of the competent application of tools. Almost all questions have been reasonably answered. Demonstrate a strong understanding of the methods and terms including predictive mining, partitioning, imputation, comparison node, ensemble, misclassification, average squared error, sensitivity, specificity, lift, ROC chart, lift chart, support and confidence during written analyses. Some minor errors are allowed. Written application is required to be of reasonable standard.	18-20
Has implemented all of the requirements above with very few errors. A strong focus on the application on creative application of tools, and evaluation and interpretation of results is evident.	21-23
All of the criteria above are met; extensive model generation and analysis have been conducted to produce exceptional outcomes and have applied principles learnt in lectures to enhance the results.	24-25