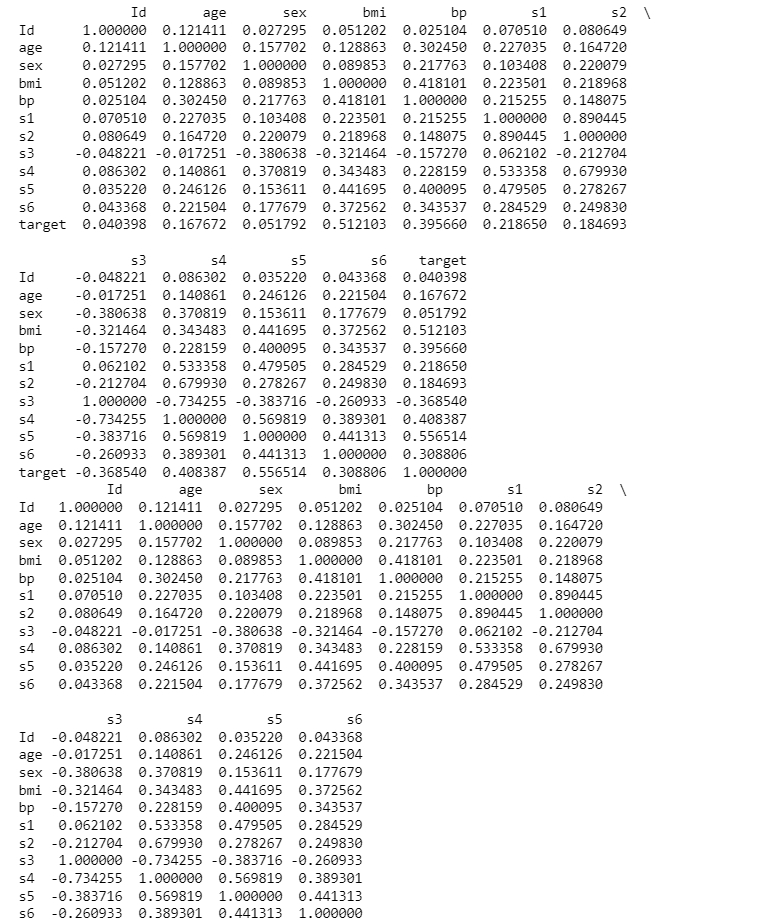
**Report**

**Programming Assignment 1**

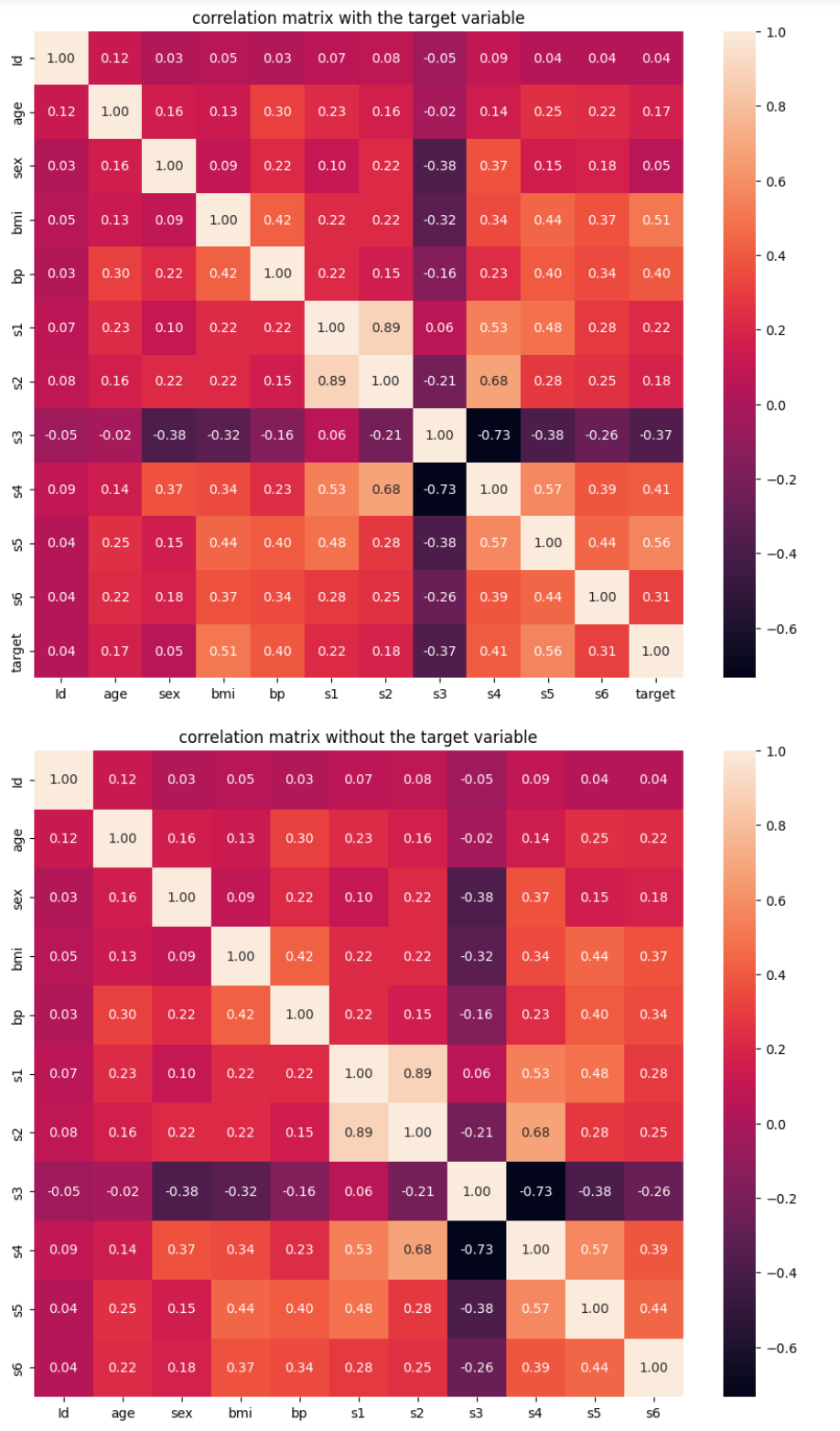
Xiangyu Gao

**Answer 1)**

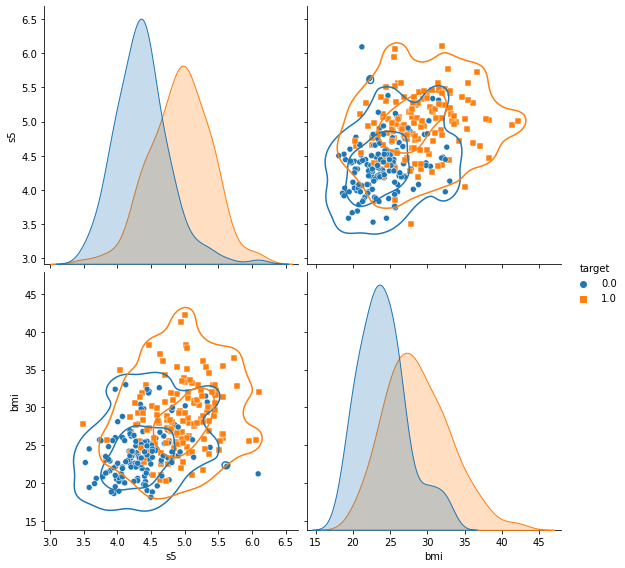
* 1. **Correlation matrix**



**1.2 Heat map**



**1.3 Scatter plots**



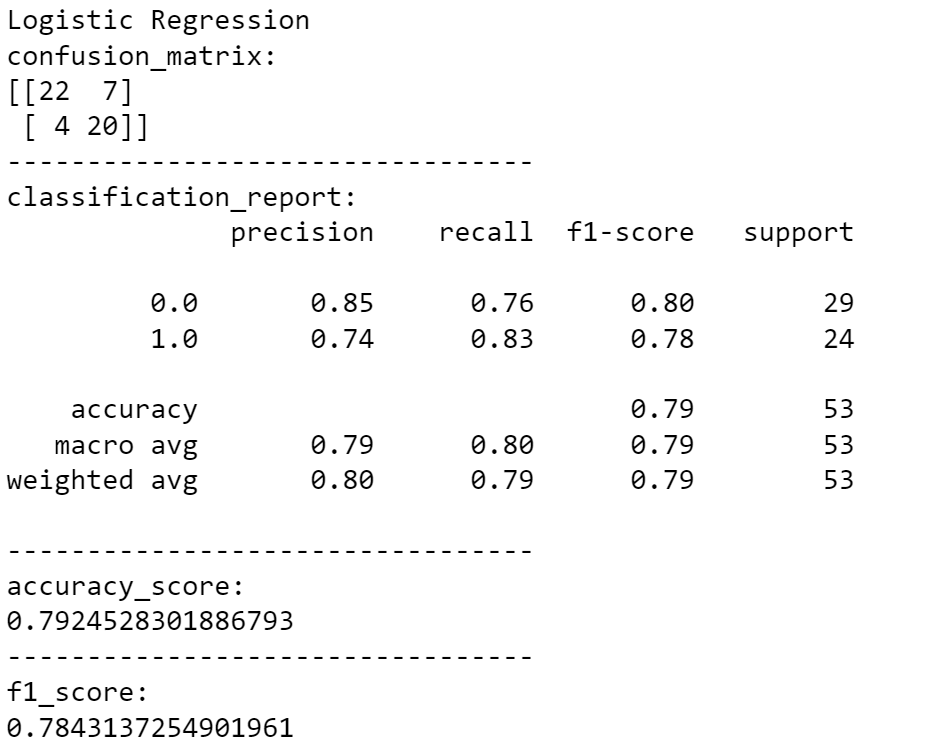
**1.4 Discussion**

From heat maps in section 1.2, it is obvious that some attributes have a positive correlation to the judgment of the target, like ‘s5’ for 0.56, ‘BMI’ for 0.51, and ‘s4’ for 0.41. On the contrary, some have a significantly negative correlation, like ‘s3’ for -0.37.

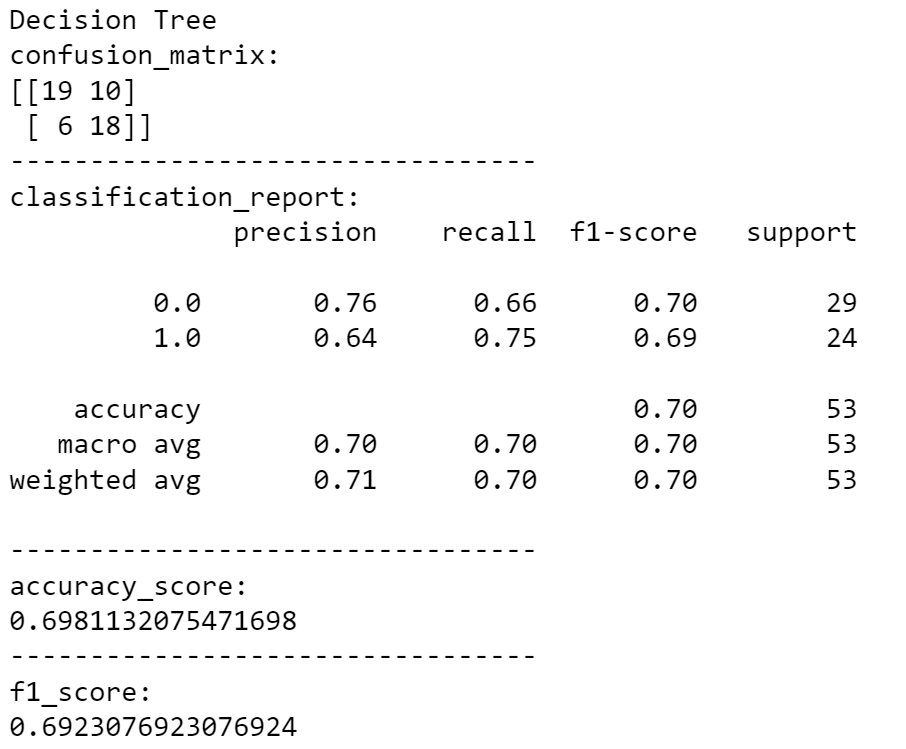
In the scatter plots in section 1.3, I chose the two most significant attributes of the examples, ‘BMI’ and ‘s5’, to make the scatter plots. We can see that the higher value, the more positive targets in the left lower graph.

**Answer 2)**

**2.1 Logistic regression**

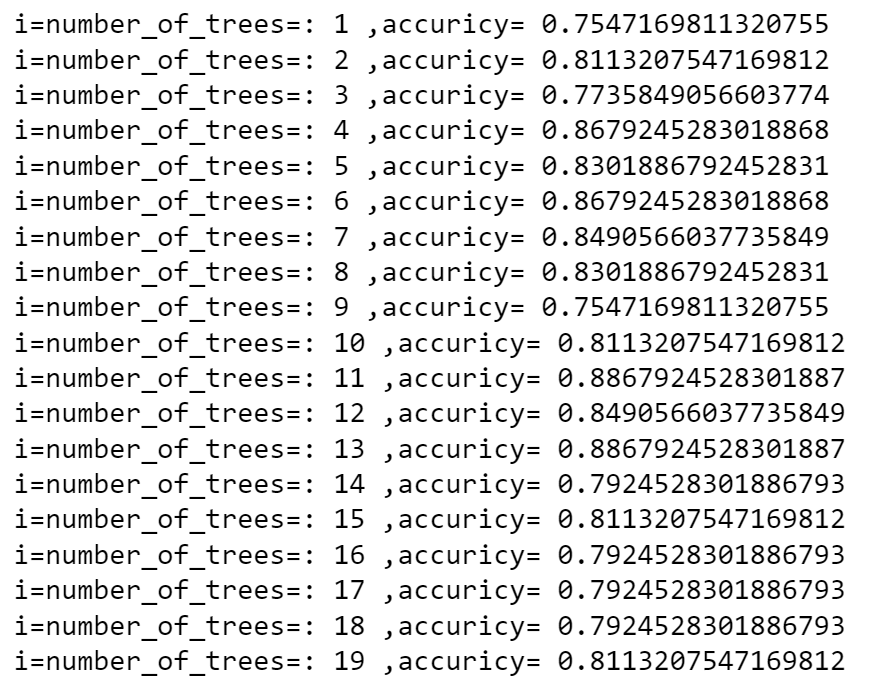


**2.2 Decision Tree**



**2.3 Random Forest**

Additionally, I have tried a random forest method to predict the examples, which is advanced and integrates the method of the decision tree. The essence of random forest is an ensemble learning algorithm. The method is to randomly generate multiple decision trees, and combine the classification results of the decision trees to determine the final classification result. This process is achieved by voting.



Randomness is embodied in: randomly selecting k subsets with replacement to train k trees, and randomly selecting subsets (data random + bifurcation random) when selecting the features of the bifurcation. The advantage of random forest is to eliminate the uncertainty caused by only using the classification tree once. Get the best results by tuning the parameters (number of trees).

It can be found that when the number of trees = 11, the accuracy rate is the highest, reaching about 88%.

**2.4 Discussion**

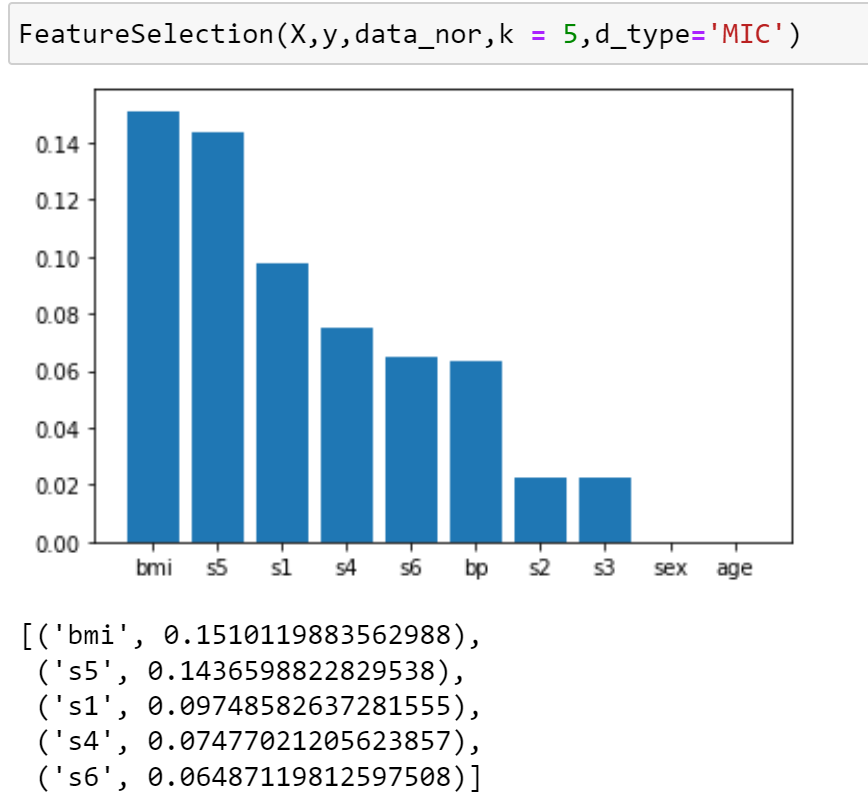
I found that default settings are useful and practical for the logistic regression model, so I turned down the tolerance for stopping criteria to a lower value to make the model more precise.

As for the decision tree, the impurity of the branch is measured by the information entropy. In each branch, the impurity is calculated for all the features, and the feature with the lowest impurity is selected for branching, and then the impurity is calculated for the remaining samples of the branch, and the impurity is continuously selected. The lowest characteristic, and so on. The tree stops growing until no features are available, or the overall impurity is optimal. The maximum depth is 5, which is definitely not higher the better, to prevent overfitting. The min\_weight\_fraction\_leaf -Weight-based pruning parameters- is 0.01. With weights, the sample size is no longer simply the number of records, but is affected by the weight of the input; min\_weight\_fraction\_leaf will be less biased towards the dominant class than the parameter min\_samples\_leaf.

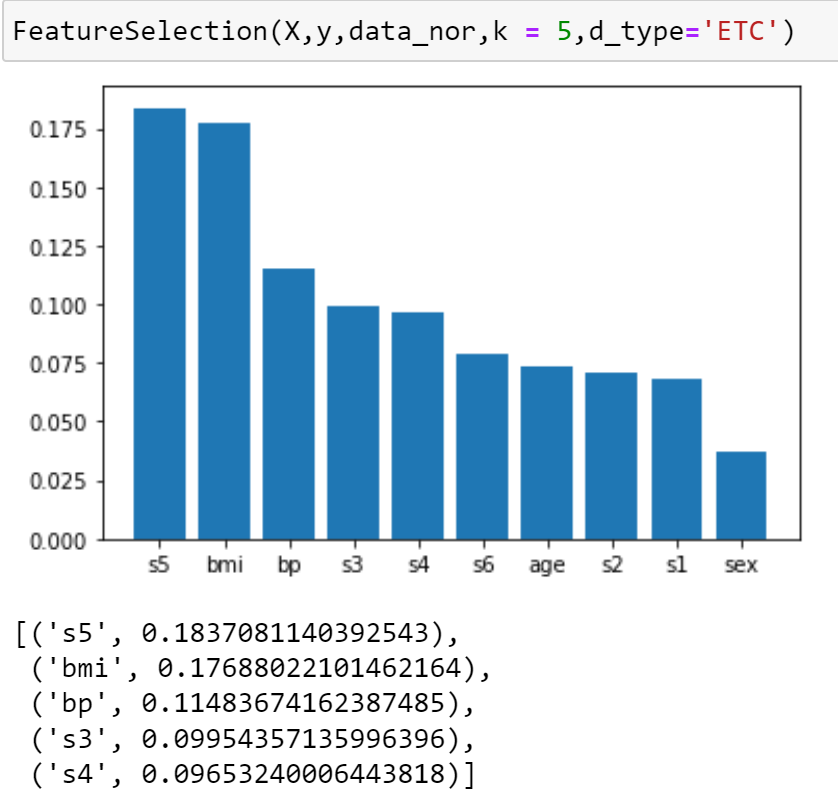
If the samples are weighted, it is easier to optimize the tree structure using a weight-based pre-pruning criterion, which ensures that leaf nodes contain at least a fraction of the sum of the sample weights

**Answer 3)**

**3.1 Mutual information**



**3.2 Extra-trees classifier**



**3.3 Discussion**

I applied two different methods to select features: mutual information and extra-trees classifier. In each method, different calculation ways cause different results. I got ‘BMI’, ‘s5’, ‘s1’, ‘s4’, and ‘s6’ as the top five attributes that are worth selecting by mutual information. By contrast, ‘s5’, ‘BMI’, ‘bp’, ‘s3, and ‘s4’ are the choices.

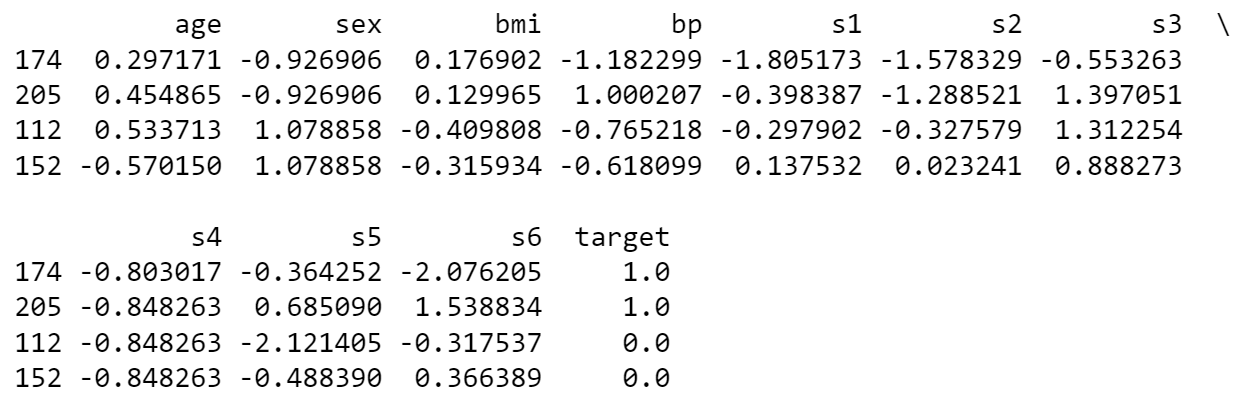
For mutual information, it is to select a feature subset that can provide as much "information" as possible to the category, so as to get more "information" about the category, and then provide help for classification. Univariate feature selection is to select the best features by a univariate based statistical test. It can be used as a preprocessing step for the estimator. The starting point of univariate feature selection is to individually detect the correlation between each feature and the target variable Y. The feature selection method for classification is chosen here.

The extra-trees classifier, this classifier implements a meta-estimator that fits many random decision trees (also called extra trees) on various subsamples of the dataset and uses averaging to improve prediction accuracy and control overfitting.

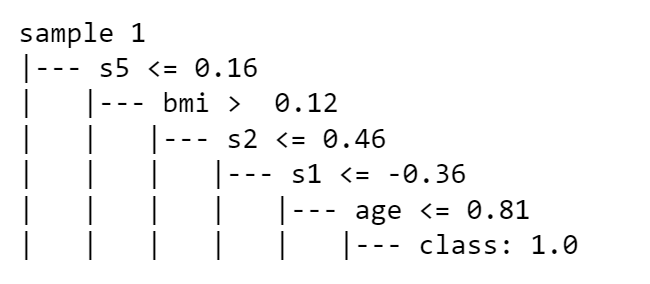
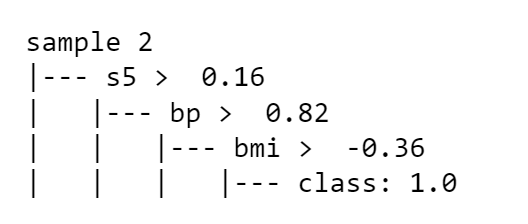
Basically yes. Between these two sets of data, the ranking of each data can basically be correlated one-to-one, but there will be some differences due to different methods.

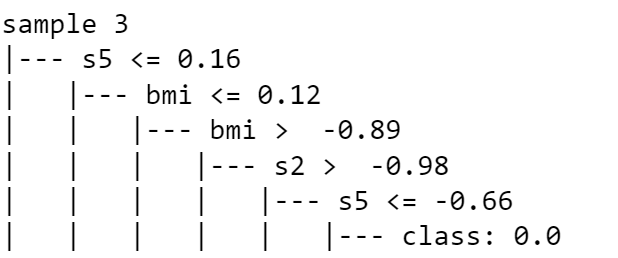
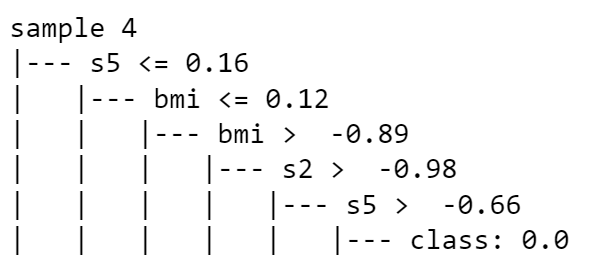
**Answer 4)**

**4.1 Examples**

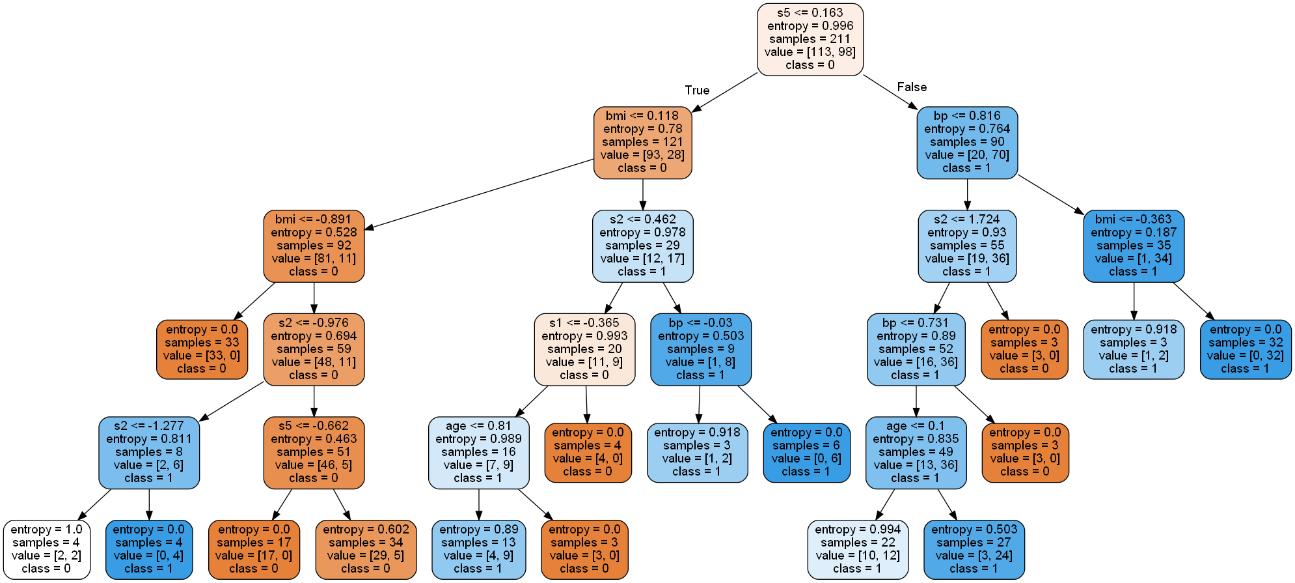


**4.2 Results**

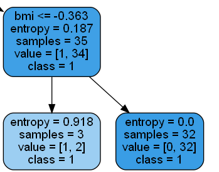
 

**4.3 Discussion**

Firstly, let us have a look at the whole structure of the decision tree:

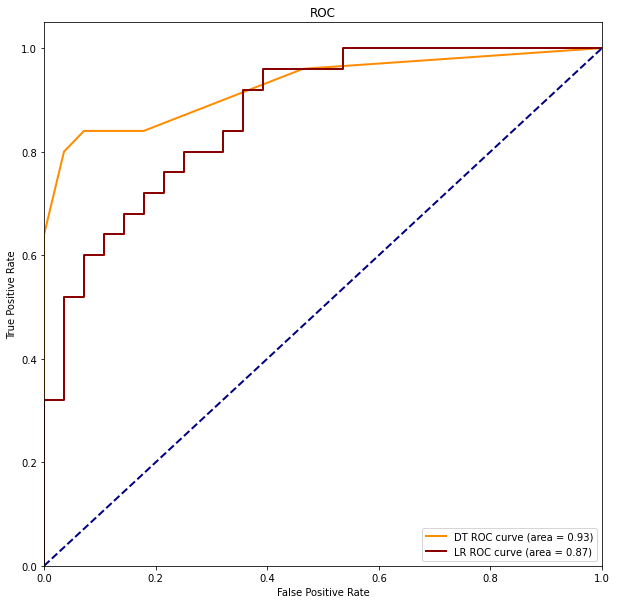
There are some duplicate classifications under the node, like

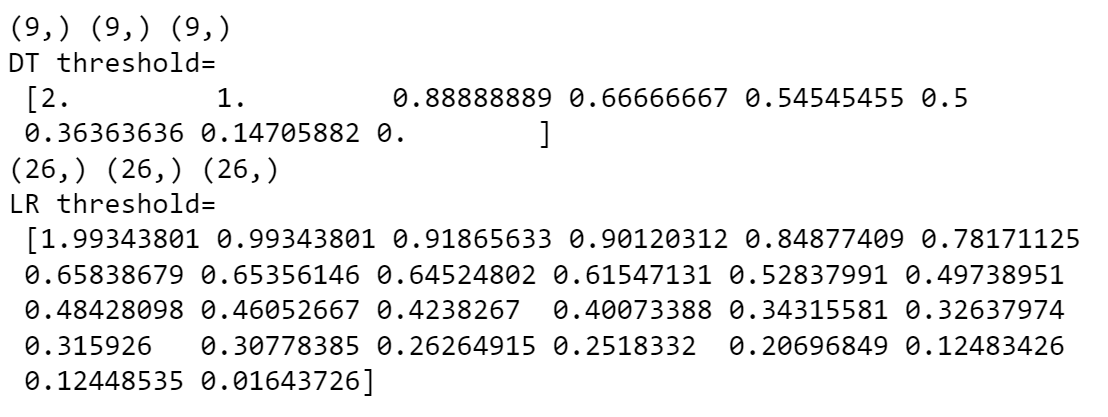


All of the classification results are class 1, which is an unnecessary classification of the tree and further pruning.

**Answer 5)**

**5.1 Plot**





**5.2 Discussion**

According to the plot, the decision tree has a better performance in the interval from 0.2 to 1, and the decision tree has a better performance in the interval from 0 to 0.2. Basically, the decision tree performs better, which has an AUC value of 0.93. By contrast, logistic regression just has 0.87 of AUC, which is lower than the decision tree.