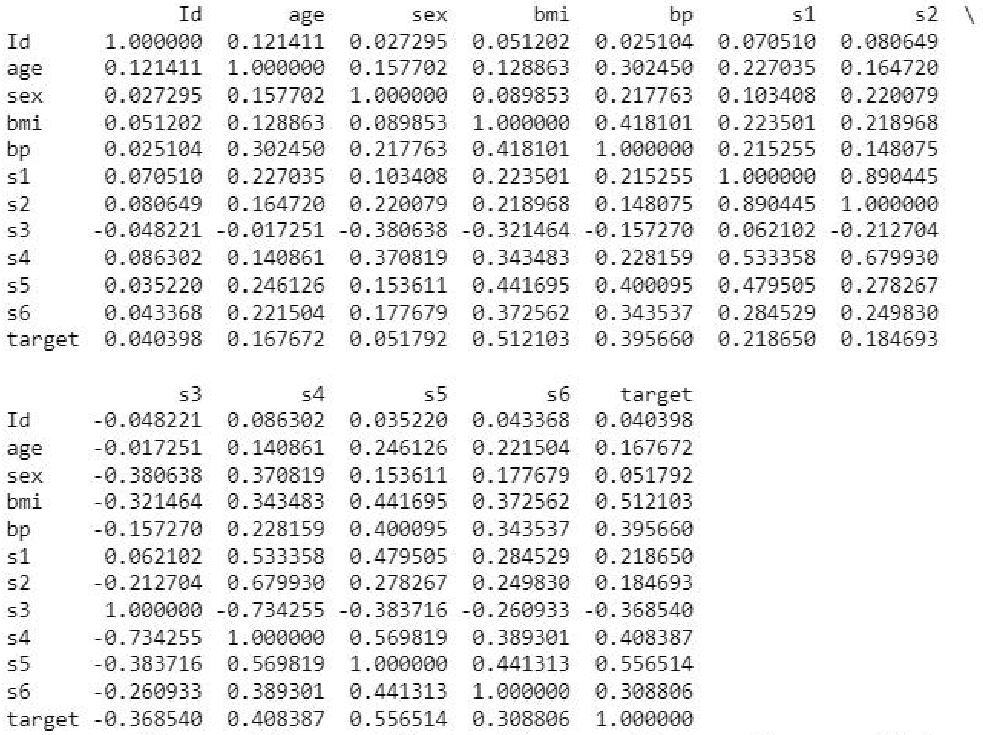
Report
Programming Assignment 1

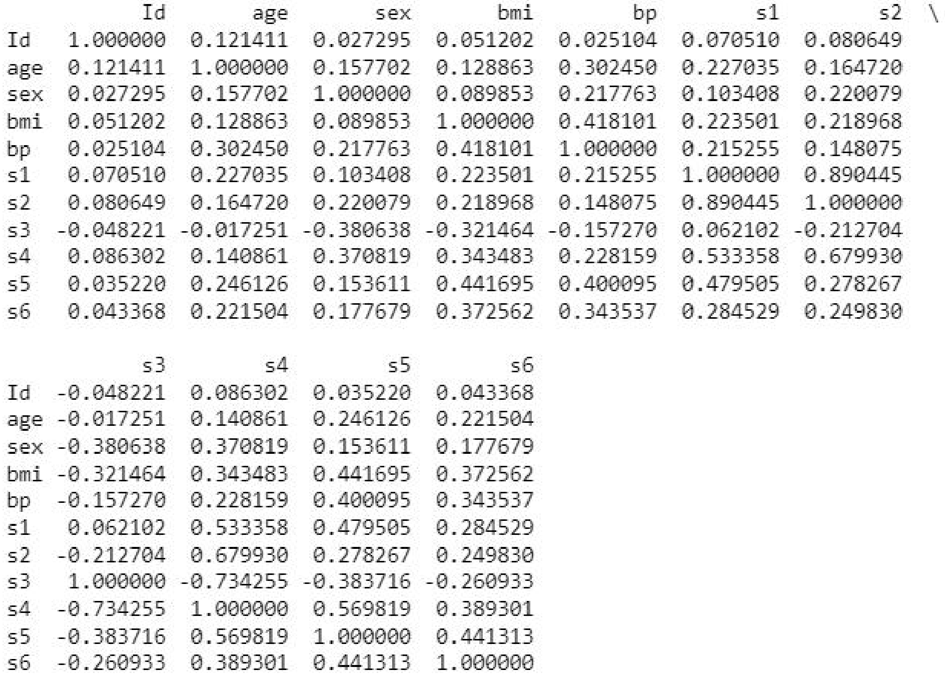
Xiangyu Gao

Answer 1)

1.1 Correlation matrix

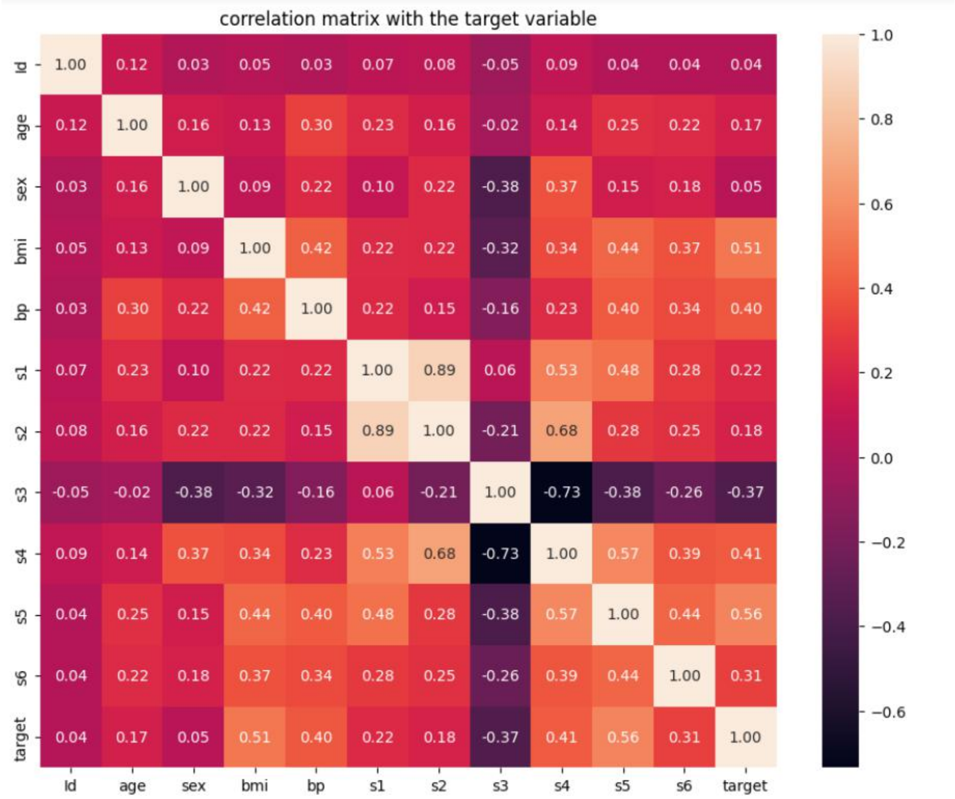


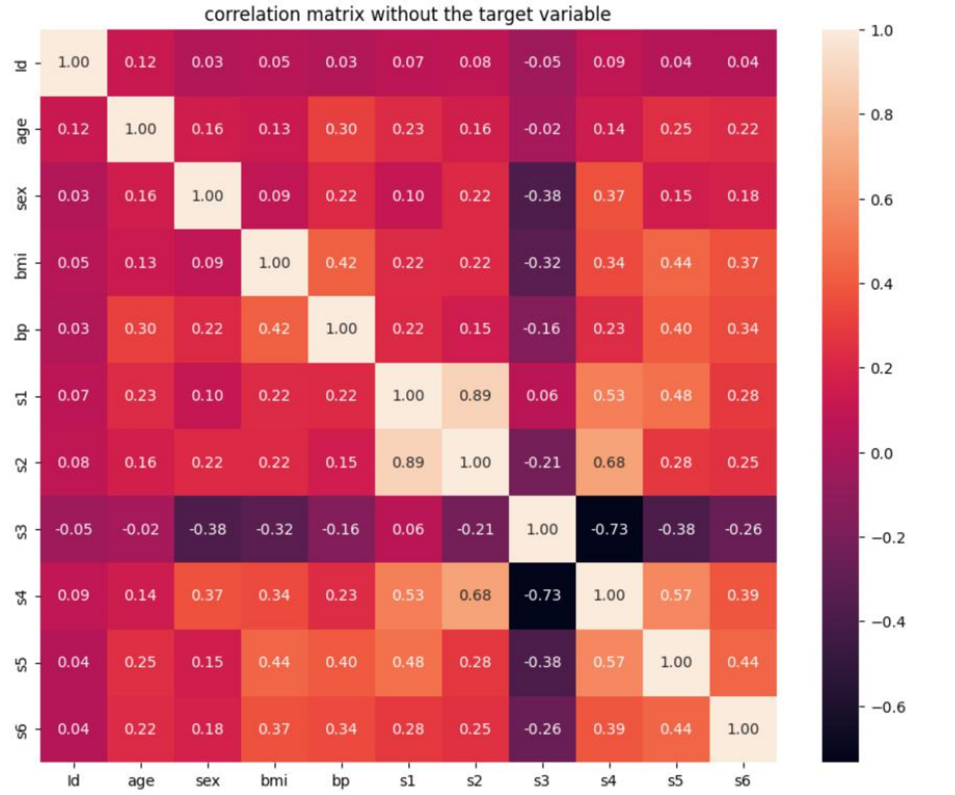
```
              Id       age       sex       bmi        bp        s1        s2  \
Id      1.000000  0.121411  0.027295  0.051202  0.025104  0.070510  0.080649
age     0.121411  1.000000  0.157702  0.128863  0.302450  0.227035  0.164720
sex     0.027295  0.157702  1.000000  0.089853  0.217763  0.103408  0.220079
bmi     0.051202  0.128863  0.089853  1.000000  0.418101  0.223501  0.218968
bp      0.025104  0.302450  0.217763  0.418101  1.000000  0.215255  0.148075
s1      0.070510  0.227035  0.103408  0.223501  0.215255  1.000000  0.890445
s2      0.080649  0.164720  0.220079  0.218968  0.148075  0.890445  1.000000
s3     -0.048221 -0.017251 -0.380638 -0.321464 -0.157270  0.062102 -0.212704
s4      0.086302  0.140861  0.370819  0.343483  0.228159  0.533358  0.679930
s5      0.035220  0.246126  0.153611  0.441695  0.400095  0.479505  0.278267
s6      0.043368  0.221504  0.177679  0.372562  0.343537  0.284529  0.249830
target  0.040398  0.167672  0.051792  0.512103  0.395660  0.218650  0.184693

              s3        s4        s5        s6    target
Id     -0.048221  0.086302  0.035220  0.043368  0.040398
age    -0.017251  0.140861  0.246126  0.221504  0.167672
sex    -0.380638  0.370819  0.153611  0.177679  0.051792
bmi    -0.321464  0.343483  0.441695  0.372562  0.512103
bp     -0.157270  0.228159  0.400095  0.343537  0.395660
s1      0.062102  0.533358  0.479505  0.284529  0.218650
s2     -0.212704  0.679930  0.278267  0.249830  0.184693
s3      1.000000 -0.734255 -0.383716 -0.260933 -0.368540
s4     -0.734255  1.000000  0.569819  0.389301  0.408387
s5     -0.383716  0.569819  1.000000  0.441313  0.556514
s6     -0.260933  0.389301  0.441313  1.000000  0.308806
target -0.368540  0.408387  0.556514  0.308806  1.000000
```



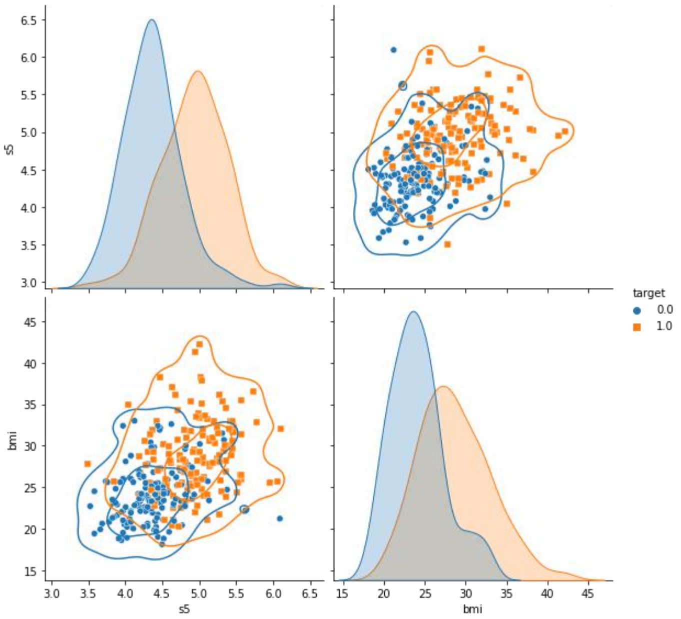
```
           Id       age       sex       bmi        bp        s1        s2  \
Id   1.000000  0.121411  0.027295  0.051202  0.025104  0.070510  0.080649
age  0.121411  1.000000  0.157702  0.128863  0.302450  0.227035  0.164720
sex  0.027295  0.157702  1.000000  0.089853  0.217763  0.103408  0.220079
bmi  0.051202  0.128863  0.089853  1.000000  0.418101  0.223501  0.218968
bp   0.025104  0.302450  0.217763  0.418101  1.000000  0.215255  0.148075
s1   0.070510  0.227035  0.103408  0.223501  0.215255  1.000000  0.890445
s2   0.080649  0.164720  0.220079  0.218968  0.148075  0.890445  1.000000
s3  -0.048221 -0.017251 -0.380638 -0.321464 -0.157270  0.062102 -0.212704
s4   0.086302  0.140861  0.370819  0.343483  0.228159  0.533358  0.679930
s5   0.035220  0.246126  0.153611  0.441695  0.400095  0.479505  0.278267
s6   0.043368  0.221504  0.177679  0.372562  0.343537  0.284529  0.249830

           s3        s4        s5        s6
Id  -0.048221  0.086302  0.035220  0.043368
age -0.017251  0.140861  0.246126  0.221504
sex -0.380638  0.370819  0.153611  0.177679
bmi -0.321464  0.343483  0.441695  0.372562
bp  -0.157270  0.228159  0.400095  0.343537
s1   0.062102  0.533358  0.479505  0.284529
s2  -0.212704  0.679930  0.278267  0.249830
s3   1.000000 -0.734255 -0.383716 -0.260933
s4  -0.734255  1.000000  0.569819  0.389301
s5  -0.383716  0.569819  1.000000  0.441313
s6  -0.260933  0.389301  0.441313  1.000000
```

1.2 Heat map


correlation matrix with the target variable
Id 1.00 0.12 0.03 0.05 0.03 0.07 0.08 -0.05 0.09 0.04 0.04 0.04
age 0.12 1.00 0.16 0.13 0.30 0.23 0.16 -0.02 0.14 0.25 0.22 0.17
sex 0.03 0.16 1.00 0.09 0.22 0.10 0.22 -0.38 0.37 0.15 0.18 0.05
bmi 0.05 0.13 0.09 1.00 0.42 0.22 0.22 -0.32 0.34 0.44 0.37 0.51
bp 0.03 0.30 0.22 0.42 1.00 0.22 0.15 -0.16 0.23 0.40 0.34 0.40
s1 0.07 0.23 0.10 0.22 0.22 1.00 0.89 0.06 0.53 0.48 0.28 0.22
s2 0.08 0.16 0.22 0.22 0.15 0.89 1.00 -0.21 0.68 0.28 0.25 0.18
s3 -0.05 -0.02 -0.38 -0.32 -0.16 0.06 -0.21 1.00 -0.73 -0.38 -0.26 -0.37
s4 0.09 0.14 0.37 0.34 0.23 0.53 0.68 -0.73 1.00 0.57 0.39 0.41
s5 0.04 0.25 0.15 0.44 0.40 0.48 0.28 -0.38 0.57 1.00 0.44 0.56
s6 0.04 0.22 0.18 0.37 0.34 0.28 0.25 -0.26 0.39 0.44 1.00 0.31
target 0.04 0.17 0.05 0.51 0.40 0.22 0.18 -0.37 0.41 0.56 0.31 1.00
Id age sex bmi bp s1 s2 s3 s4 s5 s6 target
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4 −0.6


correlation matrix without the target variable
Id 1.00 0.12 0.03 0.05 0.03 0.07 0.08 -0.05 0.09 0.04 0.04
age 0.12 1.00 0.16 0.13 0.30 0.23 0.16 -0.02 0.14 0.25 0.22
sex 0.03 0.16 1.00 0.09 0.22 0.10 0.22 -0.38 0.37 0.15 0.18
bmi 0.05 0.13 0.09 1.00 0.42 0.22 0.22 -0.32 0.34 0.44 0.37
bp 0.03 0.30 0.22 0.42 1.00 0.22 0.15 -0.16 0.23 0.40 0.34
s1 0.07 0.23 0.10 0.22 0.22 1.00 0.89 0.06 0.53 0.48 0.28
s2 0.08 0.16 0.22 0.22 0.15 0.89 1.00 -0.21 0.68 0.28 0.25
s3 -0.05 -0.02 -0.38 -0.32 -0.16 0.06 -0.21 1.00 -0.73 -0.38 -0.26
s4 0.09 0.14 0.37 0.34 0.23 0.53 0.68 -0.73 1.00 0.57 0.39
s5 0.04 0.25 0.15 0.44 0.40 0.48 0.28 -0.38 0.57 1.00 0.44
s6 0.04 0.22 0.18 0.37 0.34 0.28 0.25 -0.26 0.39 0.44 1.00
Id age sex bmi bp s1 s2 s3 s4 s5 s6
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4 −0.6

1.3 Scatter plots



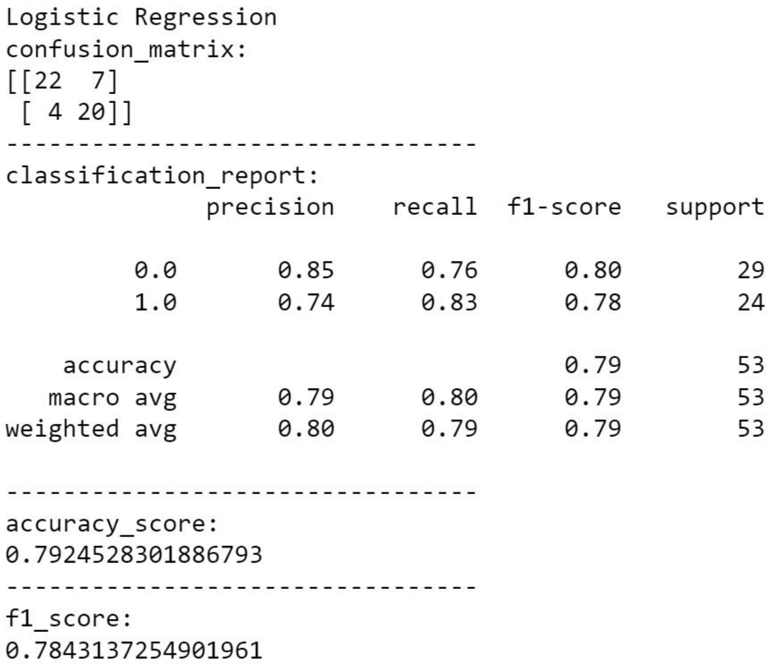
1.4 Discussion

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.

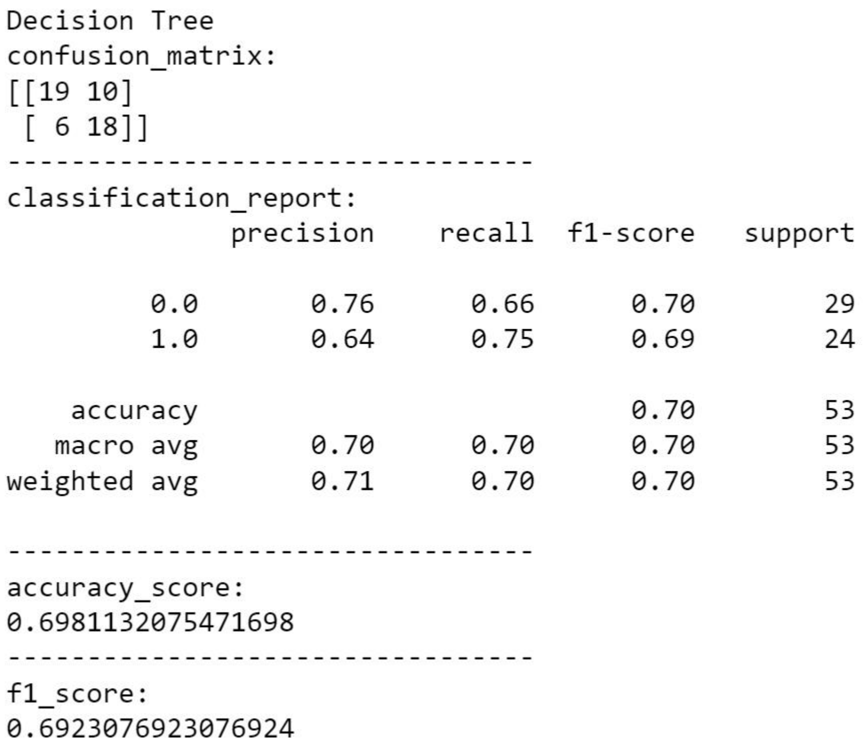
Answer 2)

2.1 Logistic regression



```
Logistic Regression
confusion_matrix:
[[22  7]
 [ 4 20]]
---------------------------------
classification_report:
              precision    recall  f1-score   support

         0.0       0.85      0.76      0.80        29
         1.0       0.74      0.83      0.78        24

    accuracy                           0.79        53
   macro avg       0.79      0.80      0.79        53
weighted avg       0.80      0.79      0.79        53

---------------------------------
accuracy_score:
0.7924528301886793
---------------------------------
f1_score:
0.7843137254901961
```

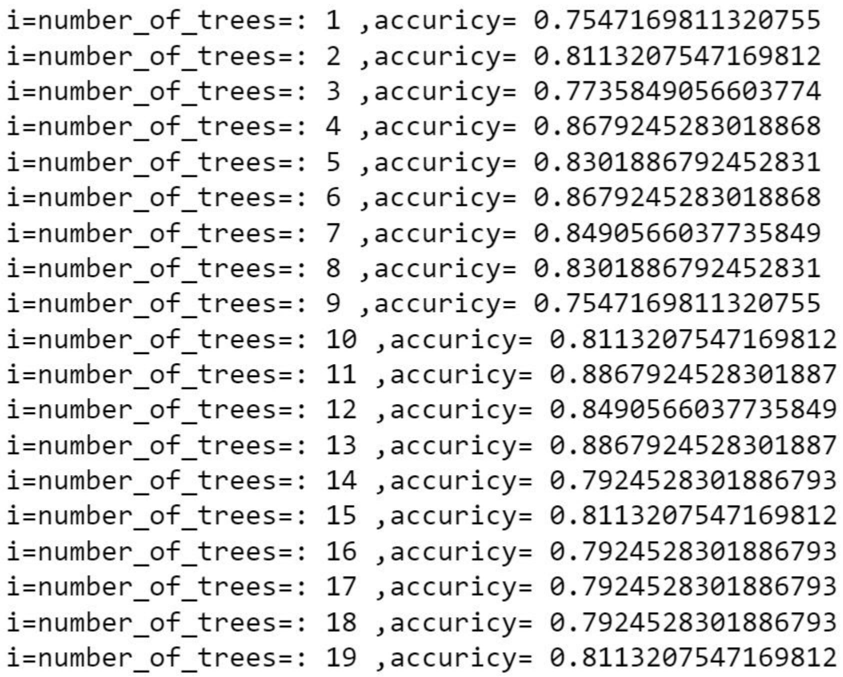
2.2 Decision Tree



```
Decision Tree
confusion_matrix:
[[19 10]
 [ 6 18]]
---------------------------------
classification_report:
              precision    recall  f1-score   support

         0.0       0.76      0.66      0.70        29
         1.0       0.64      0.75      0.69        24

    accuracy                           0.70        53
   macro avg       0.70      0.70      0.70        53
weighted avg       0.71      0.70      0.70        53

---------------------------------
accuracy_score:
0.6981132075471698
---------------------------------
f1_score:
0.6923076923076924
```

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.



```
i=number_of_trees=: 1 ,accuricy= 0.7547169811320755
i=number_of_trees=: 2 ,accuricy= 0.8113207547169812
i=number_of_trees=: 3 ,accuricy= 0.7735849056603774
i=number_of_trees=: 4 ,accuricy= 0.8679245283018868
i=number_of_trees=: 5 ,accuricy= 0.8301886792452831
i=number_of_trees=: 6 ,accuricy= 0.8679245283018868
i=number_of_trees=: 7 ,accuricy= 0.8490566037735849
i=number_of_trees=: 8 ,accuricy= 0.8301886792452831
i=number_of_trees=: 9 ,accuricy= 0.7547169811320755
i=number_of_trees=: 10 ,accuricy= 0.8113207547169812
i=number_of_trees=: 11 ,accuricy= 0.8867924528301887
i=number_of_trees=: 12 ,accuricy= 0.8490566037735849
i=number_of_trees=: 13 ,accuricy= 0.8867924528301887
i=number_of_trees=: 14 ,accuricy= 0.7924528301886793
i=number_of_trees=: 15 ,accuricy= 0.8113207547169812
i=number_of_trees=: 16 ,accuricy= 0.7924528301886793
i=number_of_trees=: 17 ,accuricy= 0.7924528301886793
i=number_of_trees=: 18 ,accuricy= 0.7924528301886793
i=number_of_trees=: 19 ,accuricy= 0.8113207547169812
```

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

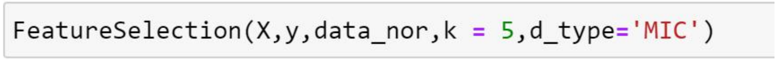
I found that default settings are useful and practical for the logistic regression model, so I turned down the tolerance for stopping criteria to a lower value to make the model more precise.

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.

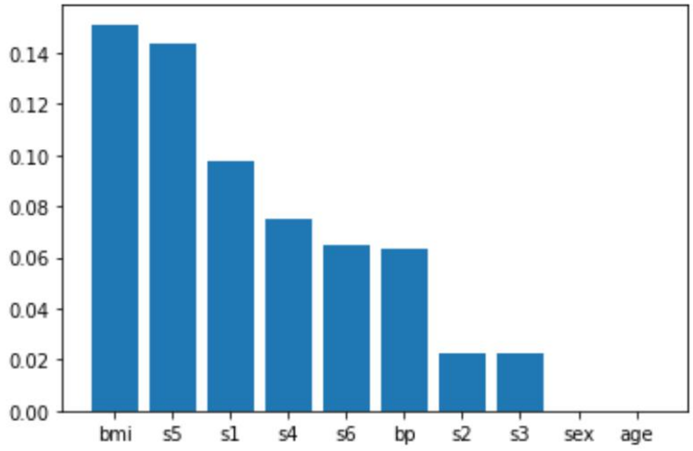
If the samples are weighted, it is easier to optimize the tree structure using a weight-based pre-pruning criterion, which ensures that leaf nodes contain at least a fraction of the sum of the sample weights

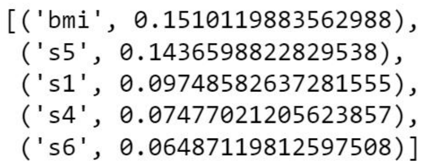
Answer 3)

3.1 Mutual information



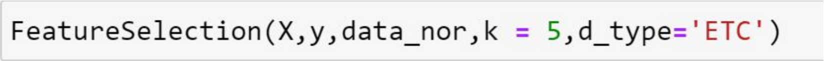
```
FeatureSelection(X,y,data_nor,k = 5,d_type='MIC')
```



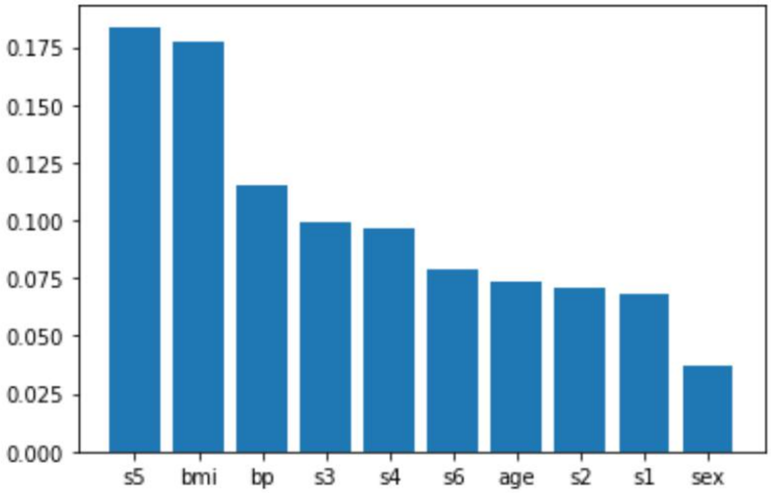


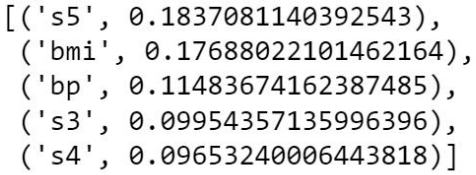
```
[('bmi', 0.1510119883562988),
 ('s5', 0.1436598822829538),
 ('s1', 0.09748582637281555),
 ('s4', 0.07477021205623857),
 ('s6', 0.06487119812597508)]
```

3.2 Extra-trees classifier



```
FeatureSelection(X,y,data_nor,k = 5,d_type='ETC')
```





```
[('s5', 0.1837081140392543),
 ('bmi', 0.17688022101462164),
 ('bp', 0.11483674162387485),
 ('s3', 0.09954357135996396),
 ('s4', 0.09653240006443818)]
```

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.

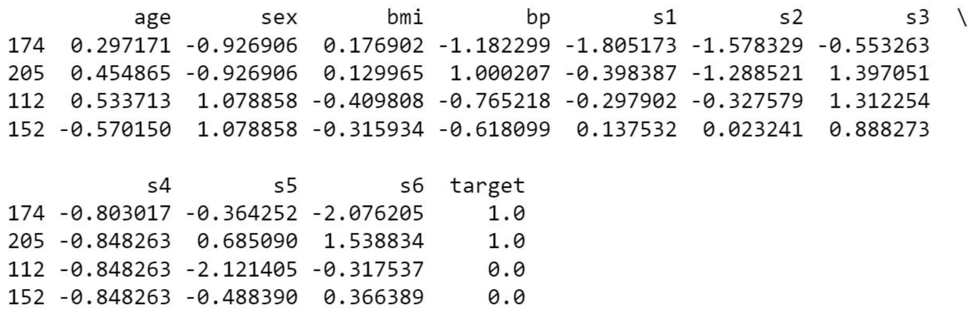
For mutual information, it is to select a feature subset that can provide as much "information" as possible to the category, so as to get more "information" about the category, and then provide help for classification. Univariate feature selection is to select the best features by a univariate based 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.

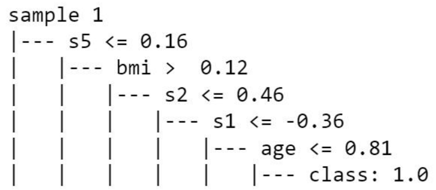
Answer 4)

4.1 Examples

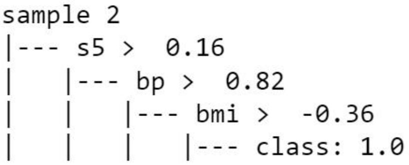


```
          age       sex       bmi        bp        s1        s2        s3  \
174  0.297171 -0.926906  0.176902 -1.182299 -1.805173 -1.578329 -0.553263
205  0.454865 -0.926906  0.129965  1.000207 -0.398387 -1.288521  1.397051
112  0.533713  1.078858 -0.409808 -0.765218 -0.297902 -0.327579  1.312254
152 -0.570150  1.078858 -0.315934 -0.618099  0.137532  0.023241  0.888273

           s4        s5        s6  target
174 -0.803017 -0.364252 -2.076205     1.0
205 -0.848263  0.685090  1.538834     1.0
112 -0.848263 -2.121405 -0.317537     0.0
152 -0.848263 -0.488390  0.366389     0.0
```

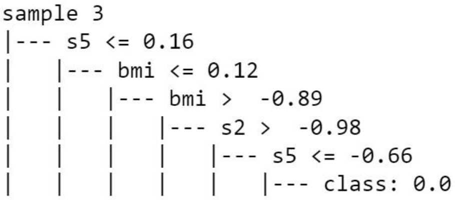
4.2 Results



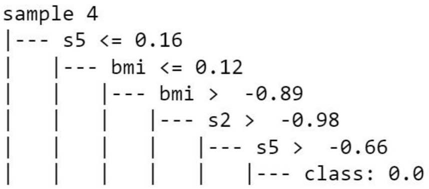
```
sample 1
|--- s5 <= 0.16
|   |--- bmi >  0.12
|   |   |--- s2 <= 0.46
|   |   |   |--- s1 <= -0.36
|   |   |   |   |--- age <= 0.81
|   |   |   |   |   |--- class: 1.0
```



```
sample 2
|--- s5 >  0.16
|   |--- bp >  0.82
|   |   |--- bmi >  -0.36
|   |   |   |--- class: 1.0
```



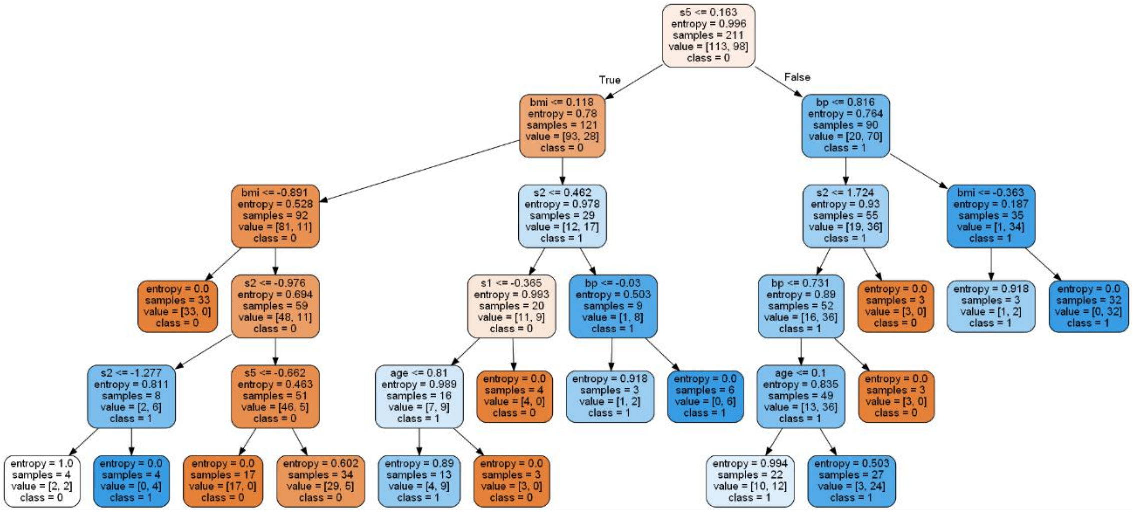
```
sample 3
|--- s5 <= 0.16
|   |--- bmi <= 0.12
|   |   |--- bmi >  -0.89
|   |   |   |--- s2 >  -0.98
|   |   |   |   |--- s5 <= -0.66
|   |   |   |   |   |--- class: 0.0
```



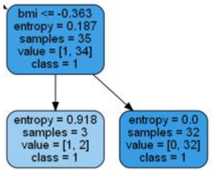
```
sample 4
|--- s5 <= 0.16
|   |--- bmi <= 0.12
|   |   |--- bmi >  -0.89
|   |   |   |--- s2 >  -0.98
|   |   |   |   |--- s5 >  -0.66
|   |   |   |   |   |--- class: 0.0
```

4.3 Discussion

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



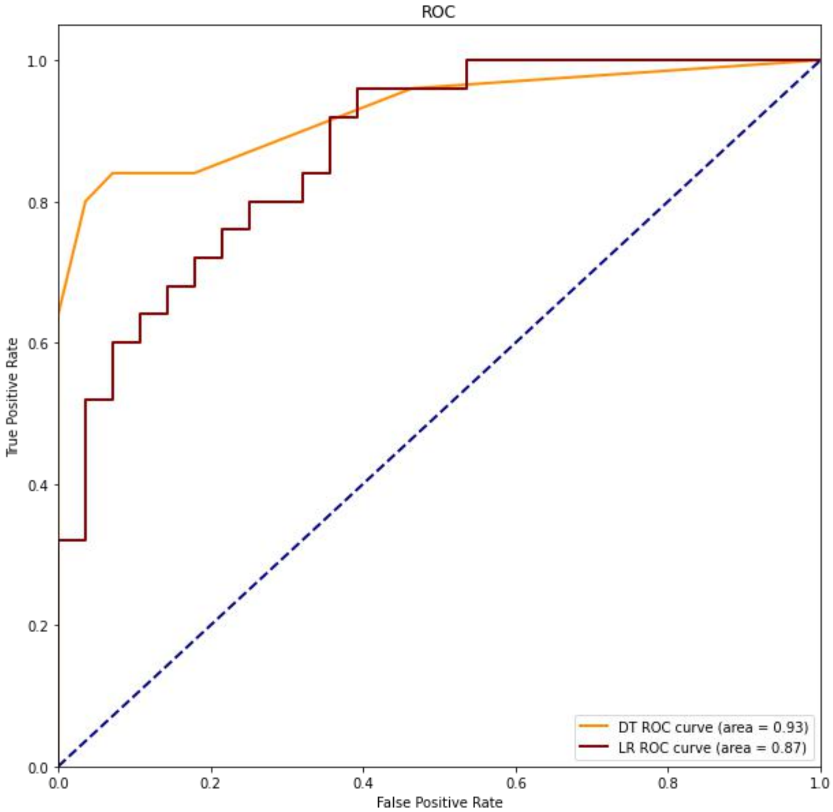
There are some duplicate classifications under the node, like

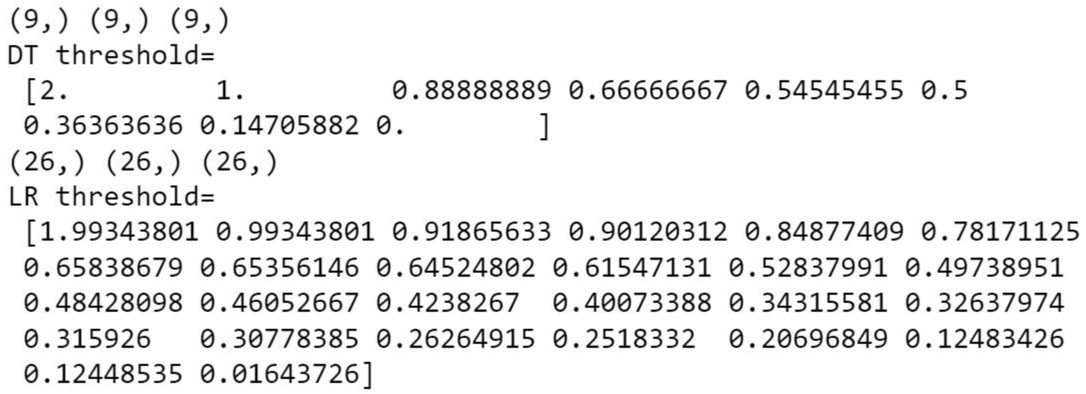


All of the classification results are class 1, which is an unnecessary classification of the tree and further pruning.

Answer 5)

5.1 Plot





```
(9,) (9,) (9,)
DT threshold=
 [2.         1.         0.88888889 0.66666667 0.54545455 0.5
 0.36363636 0.14705882 0.        ]
(26,) (26,) (26,)
LR threshold=
 [1.99343801 0.99343801 0.91865633 0.90120312 0.84877409 0.78171125
 0.65838679 0.65356146 0.64524802 0.61547131 0.52837991 0.49738951
 0.48428098 0.46052667 0.4238267  0.40073388 0.34315581 0.32637974
 0.315926   0.30778385 0.26264915 0.2518332  0.20696849 0.12483426
 0.12448535 0.01643726]
```

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.