Heart Disease Prediction

In this project, we will use R to analyse a dataset. The data is contained in the file heart.train.ass3.2019.csv. In this dataset, each observation represents a patient at a hospital that reported showing signs of possible heart disease. The outcome is presence of heart disease (HD), or not, so this is a classification problem. The predictors are summarised in Table 1 (overleaf). We are interested in learning a model that can predict heart disease from these measurements.

1.1)

fitting a decision tree to the data using the tree package. Use cross-validation with 10 folds and 1000 repetitions to select an appropriate size tree. What variables have been used in the best tree? How many leaves (terminal nodes) does the best tree have?

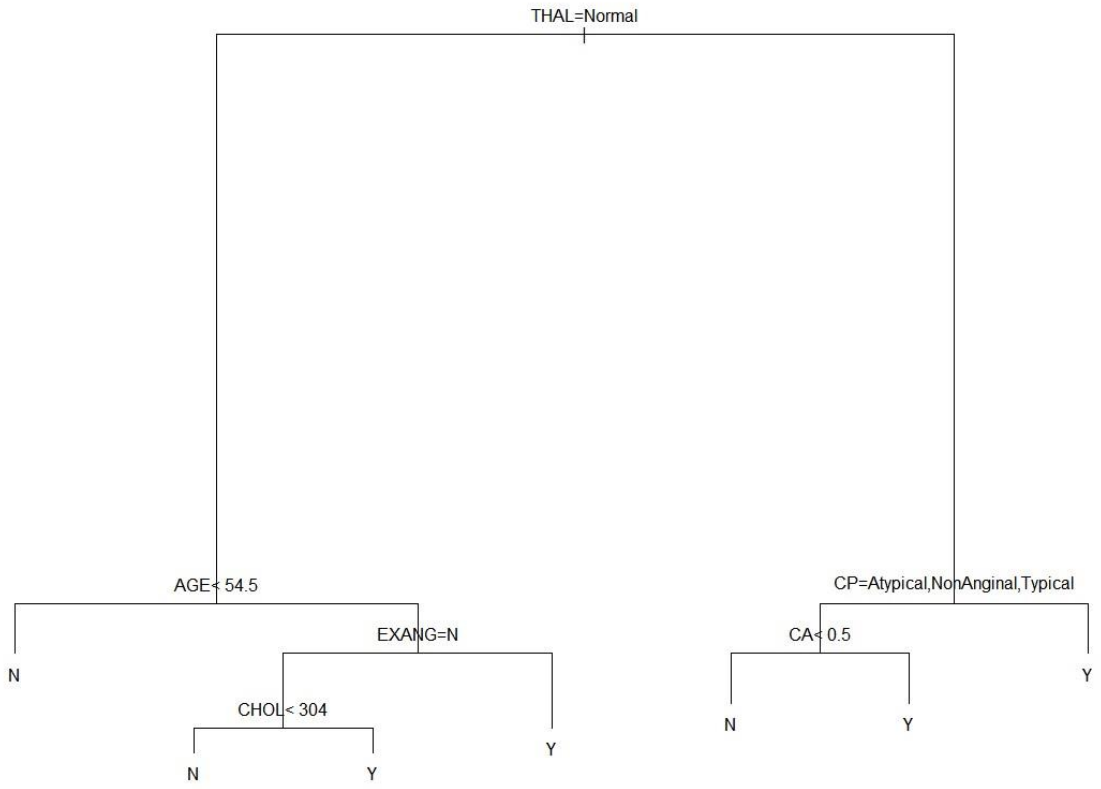
The variable that has been used in the best tree are:

THAL
AGE
EXANG
CHOL
CP
CA

There are 7 leaves in the best tee.

1.2)

Plot the tree found by CV and discuss what it tells you about the relationship between the predictors and heart disease.



In the decision plotted above, there are 5 variables including THAL, AGE, CP, CA, CHOL and EXANG. The tree is distributed according to these variables as mentioned above and in the tree. All of six variable leads to 'Yes' and 'No' category for having a heart-disease.

According to the tree, it interprets that if you have Normal THAL (Thallium scanning results) and are younger than the age of 54.5, there would be less probability of having a presence of heart disease in that particular patient.

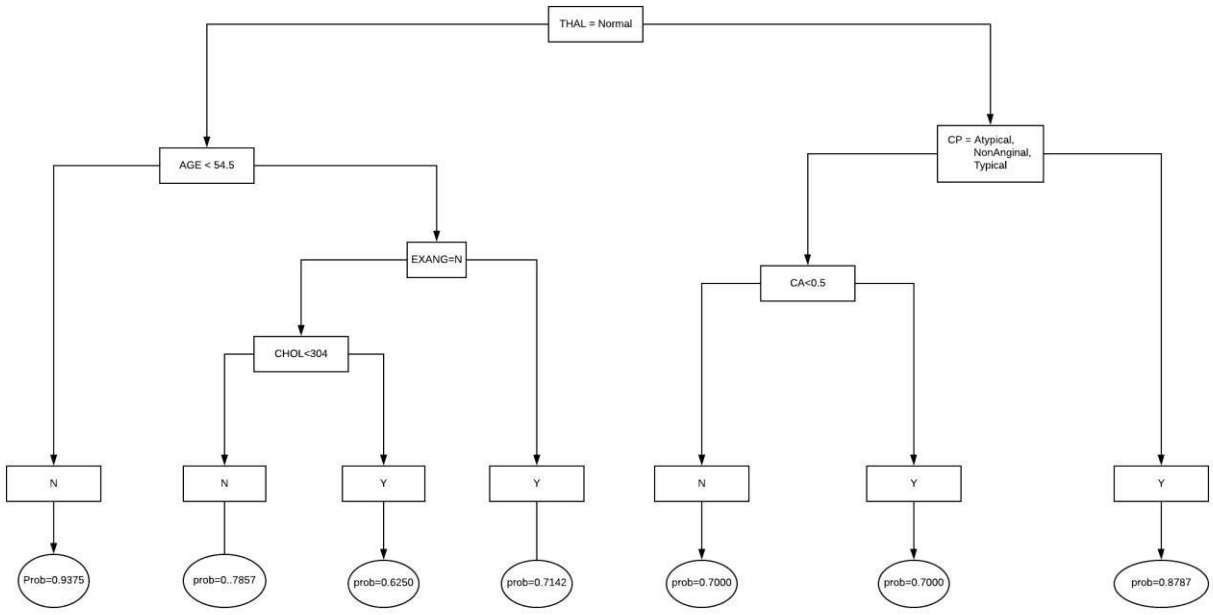
However, In the case of Thallium scanning results being normal, if you are older than or equal to the age of 54.5, then it will depend on EXANG variable (Exercise induced angina] to see if you have a heart disease or not. If you have Exercise induced angina meaning your EXANG value is 'yes', then you are likely to have heart-disease. But, if EXANG value is 'No', and your CHOL value is less than 304 mg/dl, then there would be less probability of having a presence of heart-disease found but if CHOL value is more than 304 mg/dl, you are having high chances of having heart-disease.

In the other main category of heart-disease, if your THAL value is not normal (meaning either you have Fixed fluid transfer effect or reverse fixed fluid transfer effect) and if you have your CP (chest pain type) being Asymptomatic, you will have high chance of having heart-disease.

While, in case of your THAL not normal and your CP (chest pain type) being Typical or Atypical or NonAnginal, your CA variable (Number of major vessels colored by flourosopy) being less than 0.5 will lead to less chance of having a heart-disease but CA being >=0.5 will give high chances of having a heart disease.

1.3)

For classification problems, the rpart package only labels the leaves with the most likely class. However, if you examine the tree structure in its textural representation on the console, you can determine the probabilities of having heart disease.



1.4) According to your tree, which predictor combination results in the highest probability of having heart-disease?

As it can be seen in the tree designed above, the more likely chance to get a heart disease meaning there will be the highest probability of heart-disease in the case where your Thallium scanning results are not normal, and your chest pain type is Asymptomatic. In that case, there is 87.87% probability that you will have heart-disease.

Hence, the combination of predictor for having the highest probability of heart-disease is: THAL variable not being normal and CP is not any of Atypical, Typical, NonAnginal.

1.5)

We will also fit a logistic regression model to the data. Use the glm () function to fit a logistic regression model to the heart data and use stepwise selection with the BIC score to prune the model. What variables does the final model include, and how do they compare with the variables used by the tree estimated by CV? Which predictor is the most important in the logistic regression?

There are five variables categorized into total of 11 variables that the final model that BIC returns.
The first five variables are: CP, OLDPEAK, EXANG, CA, THAL.

These variables are categorized into: CP into CPAsymptomatic, CPAtypical, CPNonAnginal, CPTypical
EXANG into EXANGN, EXANGY,
OLDPEAK,
CA,
THAL into THALFixed.Defect,
THALNormal, THALReversible.Defect

As comparing these variables by BIC model with the ones returned by decision tree, it can be seen that the pruned logistic regression model included the variable OLDPEAK, while the tree did not include it, and instead included AGE and CHOL variables. There are also THAL, CA, CP and EXANG which appear in both models.

The most important predictor among all predicted by BIC is CA as it has the slope nearly 1 (closest than Any other) which makes it really very important to predict Heart-disease.

1.6)

Write down the regression equation for the logistic regression model you found using stepwise selection.

The regression equation:

```
[HD] = -1.2172653 + 2.6324294*CPAsymptomatic + 1.7530971*CPAtypical
+ 1.2307159*CPNonAnginal - 1.3348209*EXANGN + 0.5519443*OLDPEAK
+ 0.8486996*CA - 0.8878904*THALFixed.Defect - 1.8851410*THALNormal
```

1.7)

Using the my. pred.stats() function contained in the file studio11.prediction.stats(), compute the prediction statistics for both the tree and the stepwise logistic regression model on the heart data. Contrast and compare the two models in terms of the various prediction statistics? Would one potentially be preferable to the other as a diagnostic test?

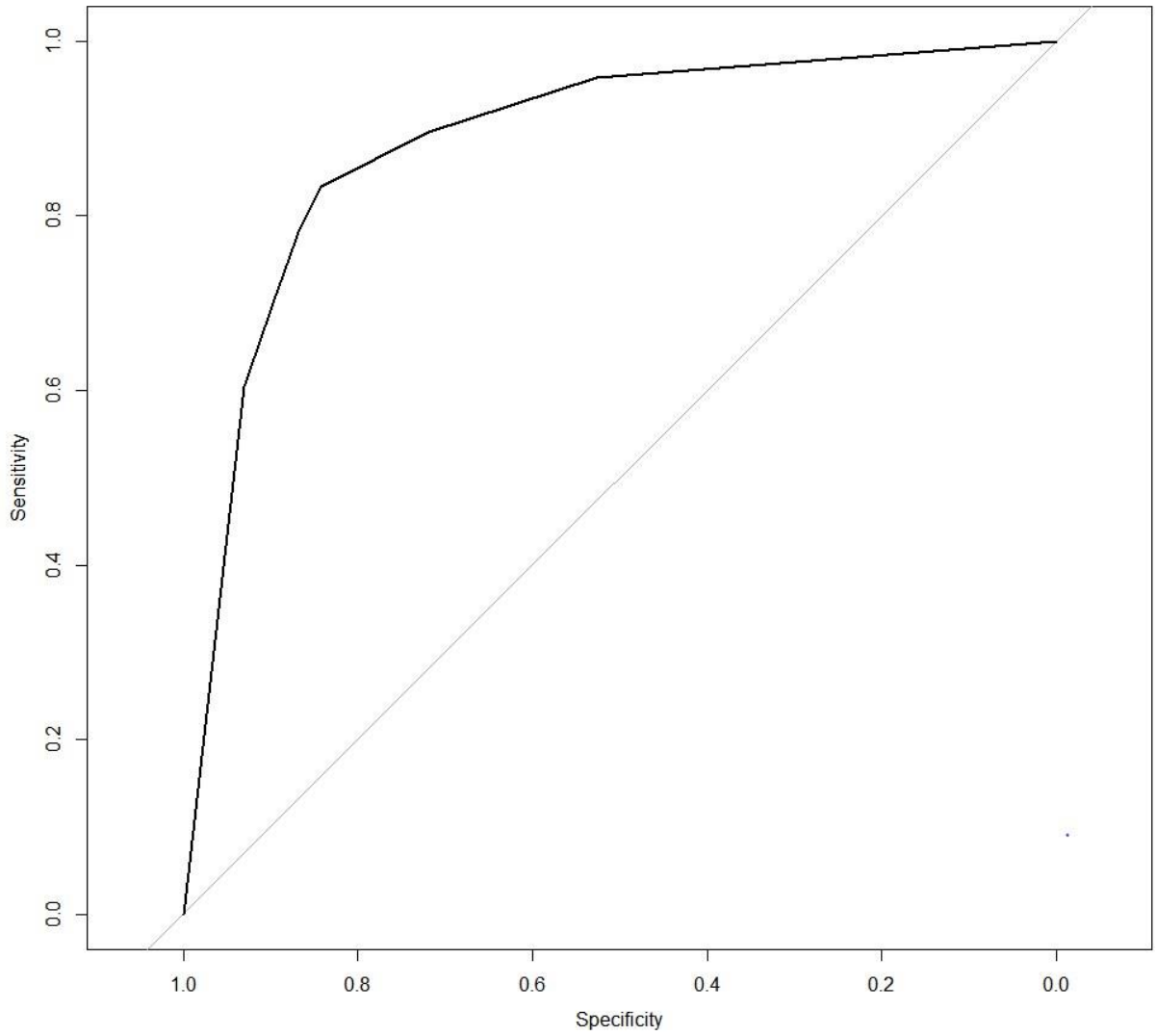
In terms of classification accuracy, the BIC model would correctly guess whether a person has heart disease more frequently than the best tree model. The BIC logistic model would be expected to be able to correctly classify people with heart disease more frequently than the best tree model in terms of

sensitivity, while the best tree model has higher specificity than the BIC model and would be able to correctly classify people with no heart diseases more frequently. The BIC model also seems to have a higher area-under-curve and less logarithmic loss as compared to the best tree model, which seems to indicate that the BIC model would be able to perform correct classification more frequently, while the best tree model would be able to predict future data about people with heart diseases with a slightly higher accuracy as it got higher specificity value.

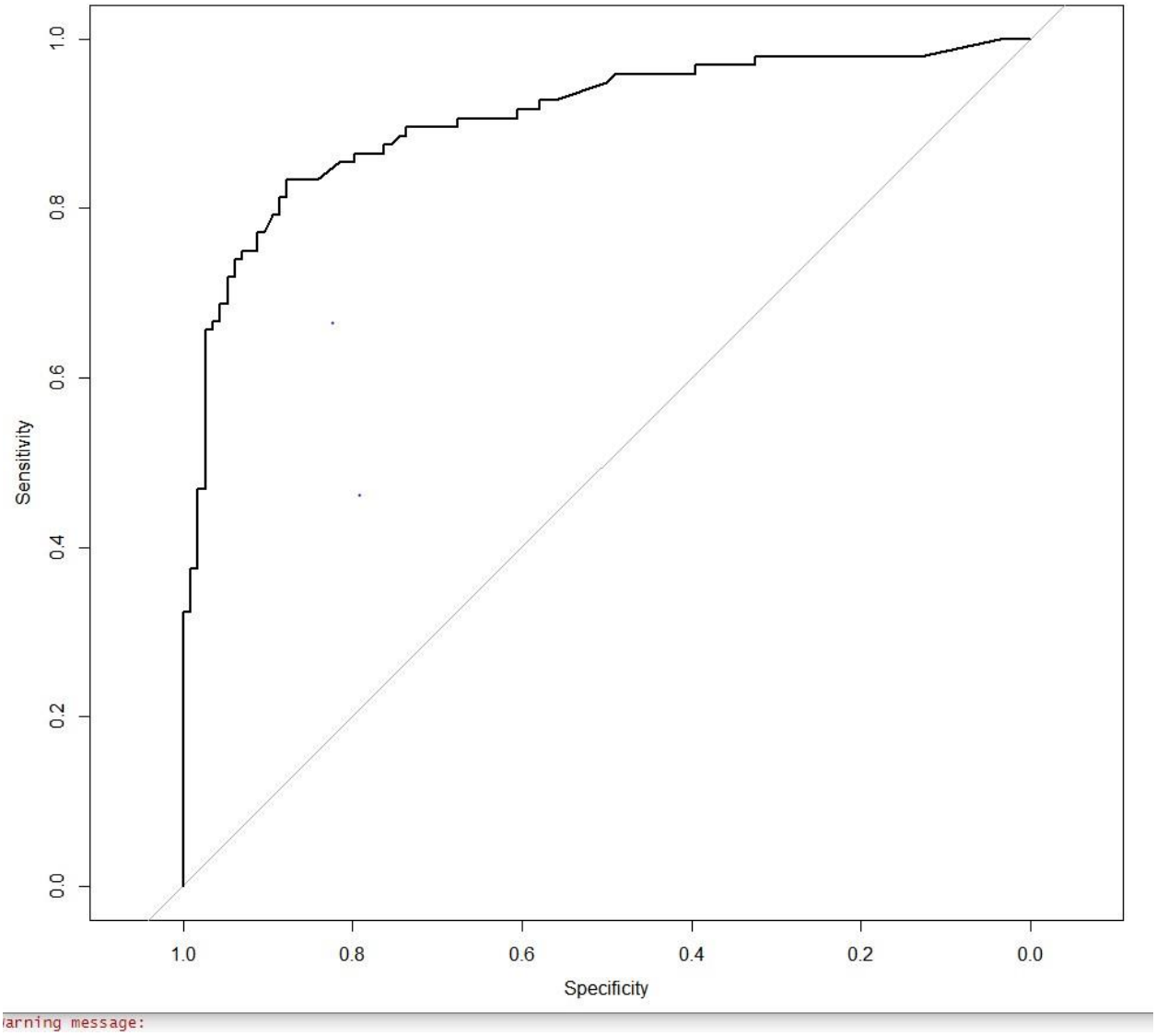
Given the information and analysis above, the BIC model would be preferable as a diagnostic test, as it would be able to predict future data regarding people with heart diseases more accurately, and thus allow us to determine factors that correlate to the symptoms or cause of heart diseases.

Furthermore, it would also be able to classify people with heart disease correctly as it has a high areaunder-curve and sensitivity, which would be appropriate for a diagnostic test where we screen patients for symptoms to decide on treatment methods.

For decision tree:



For Logistic regression model:



1.8)

The file heart.test.ass3.2019.csv contains the data on a further n = 66 individuals. Calculate the odds of having heart disease for the patient in the 45th row of this new dataset. The odds should be calculated for both: (a) the tree model found using cross-validation; and (b) the stepwise logistic regression model. How do the predicted odds for the two models compare?

The predicted odds for the tree model found using cross validation is: 2.33 [from probability: 0.7000]

It can be found by going down the tree [From: root node -> To the right category (as THAL= Reversible.Defect) -> To the left (CP=NonAnginal) -> To the left (as CA<0.5)

1.9)

For the logistic regression model using the predictors selected by BIC in Question 1.6, use the bootstrap procedure (use at least 5, 000 bootstrap replications) to find a confidence interval for the probability of having heart disease for patient in the 45th row in the test data. Use the bca option when computing this confidence interval. Discuss this confidence interval in comparison to the predicted probabilities of having heart disease for both the logistic regression model and the tree model.

As the bootstrap procedure used in R script, the confidence interval for the probability of having heart-disease for patient in the 45^{th} row in the test data is:

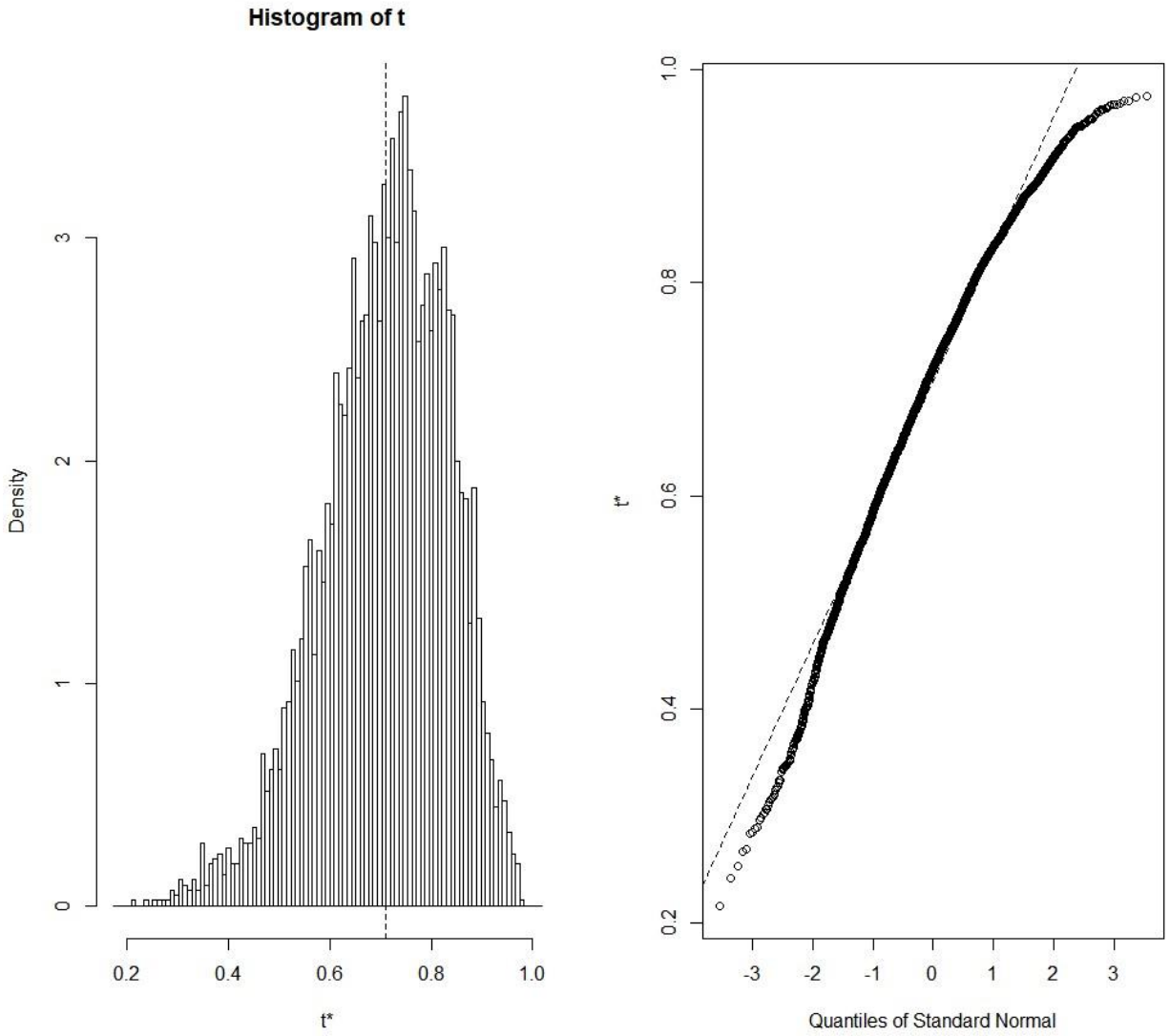
(0.4049, 0.8986) with a 95% confidence.

Comparing this confidence interval for the models computed previously,

When compared to the predicted odds for the BIC model and the best tree model, which are 0.7102 and 0.7000, we can see that the predicted odds do fall into the calculated confidence interval, which means that we can be quite sure that the odds for heart disease with the given predictors from 45^{th} row patient does indeed fall in the interval above. [There is a chance that the confidence interval will be changed for next time it runs as bootstrap runs for 5000 which makes it a little deviational]

1.10)

Finally, use the bootstrap to compute a 95% confidence interval for the classification accuracy of the logistic regression model using the predictors selected by BIC. Plot the bootstrap results. Comment on the histogram of the bootstrapped classification accuracies. Compare the bootstrap 95% confidence interval for the classification accuracy against the actual classification accuracy you obtain on the testing data heart.test.ass3.2019.csv using the model you learned in 1.6.



The classification accuracy was found 0.8428 while the confidence interval is found for classification accuracy is (0.75, 0.87) with 95% being confident. And classification accuracy does certainly contain the classification accuracy which is in the range.

When compared to the classification accuracy for the BIC, we can see that the predicted classification accuracy does fall into the calculated confidence interval, which means that we can be quite sure that the classification accuracy for heart disease with the given data does indeed fall in the interval above. And the classification seems to be accurate.