Question 2 (18 marks)

In this question we will analyse the data in heart.train.ass3.2020.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 3. We are interested in learning a model that can predict heart disease from these measurements. To answer this question you must:

- Provide an R script containing all the code you used to answer the questions. Please use comments to ensure that the code used to identify each question is **clearly identifiable**. Call this fn.sn.Q2.R, where "fn.sn" is your first name followed by your family name.
- Provide appropriate written answers to the questions, along with any graphs, in a report document.

When answering this question, you must use the rpart package that we used in Studio 9. The wrapper function for learning a tree using cross-validation that we used in Studio 9 is contained in the file wrappers.R. Don't forget to source this file to get access to the function.

1. Using the techniques you learned in Studio 9, fit a decision tree to the data using the tree package. Use cross-validation with 10 folds and 5,000 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? **[2 marks]**

2. Plot the tree found by CV, and discuss clearly and thoroughly in plain English what it tells you about the relationship between the predictors and heart disease. *(hint: you can use the* text(cv$best.tree,pretty=12) *function to add appropriate labels to the tree)*. **[3 marks]**

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 (see Question 2.3 from Studio 9 as a guide) in each leaf (terminal node). Take a screen-capture of the plot of the tree (don't forget to use the "zoom" button to get a larger image) or save it as an image using the "Export" button in R Studio.

   Then, use the information from the textual representation of the tree available at the console and annotate the tree in your favourite image editing software; next to all the leaves in the tree, add text giving the probability of contracting heart disease. Include this annotated image in your report file. **[2 marks]**

4. According to your tree, which predictor combination results in the highest probability of having heart-disease? **[1 mark]**

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? **[3 marks]**

6. Write down the regression equation for the logistic regression model you found using step-wise selection. **[1 mark]**

7. The file heart.test.ass3.2020.csv contains the data on a further $n^0 = 200$ individuals. Using the my.pred.stats() function contained in the file my.prediction.stats.R, compute the prediction statistics for both the tree and the step-wise logistic regression model on this test 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? Justify your answer. **[2 marks]**

8. Calculate the *odds* of having heart disease for the patient in the 69th row of the test dataset. The odds should be calculated for both:

(a) the tree model found using cross-validation; and (b) the

step-wise logistic regression model.

How do the predicted odds for the two models compare? **[2 marks]**

9. For the logistic regression model using the predictors selected by BIC in Question 2.6, 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 69th 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. **[2 marks]**

Variable name	Description	Values
AGE	Age of patient in years	29 – 77
SEX	Sex of patient	M = Male F = Female
CP	Chest pain type	Typical = Typical angina Atypical = Atypical angina NonAnginal = Non anginal pain Asymptomatic = Asymptomatic pain
TRESTBPS	Resting blood pressure (in *mmHg*)	94 – 200
CHOL	Serum cholesterol in *mg/dl*	126 – 564
FBS	Fasting blood sugar > 120*mg/dl* ?	<120 = No >120 = Yes
RESTECG	Resting electrocardiographic results	Normal = Normal ST.T.Wave = ST wave abnormality Hypertrophy = showing probable hypertrophy
THALACH	Maximum heart rate achieved	71 – 202
EXANG	Exercise induced angina?	N = No Y = Yes
OLDPEAK	Exercise induced ST depression relative to rest	0 – 6.2
SLOPE	Slope of the peak exercise ST segment	Up = Up-sloping Flat = Flat Down = Down-sloping
CA	Number of major vessels colored by flourosopy	0 – 3
THAL	Thallium scanning results	Normal = Normal Fixed.Defect = Fixed fluid transfer defect Reversible.Defect = Reversible fluid transfer defect
HD	Presence of heart disease	N = No Y = Yes

Table 3: Heart Disease Data Dictionary. ST depression refers to a particular type of feature in an electrocardiograph (ECG) signal during periods of exercise. Thallium scanning refers to the use of radioactive Thallium to check the fluid transfer capability of the heart.