Specialist Masters Programme

Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: (Surname, first name) 1. Kartikey Bhalla 2. Shashank Gola 3. Mohemmed Faisal	4. Preeti Pothireddy 5. Aparna Viswanathan 6. **7** **GROUP NUMBER:**
MSc in: Business Analytics	
Module Code: SMM 636	
Module Title: Machine Learning	
Lecturer: Dr Rui Zhu	**Submission Date: 3rd March**

Declaration:

By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission: **Final Mark:**

Total Number of Words: 540

Q1. The task of the dataset:

Overall, the heart dataset provides extensive information about 303 individuals across 13 different attributes. The primary purpose of the data set is to classify individuals with AHD (Acute Heart Disease), based on factors such as age, sex, heart rate, cholesterol, fasting blood sugar and much more. Why is this important? Globally, heart diseases are one of the primary causes of death. Hence, its early detection and management is quite important and we believe machine learning can be of great help.

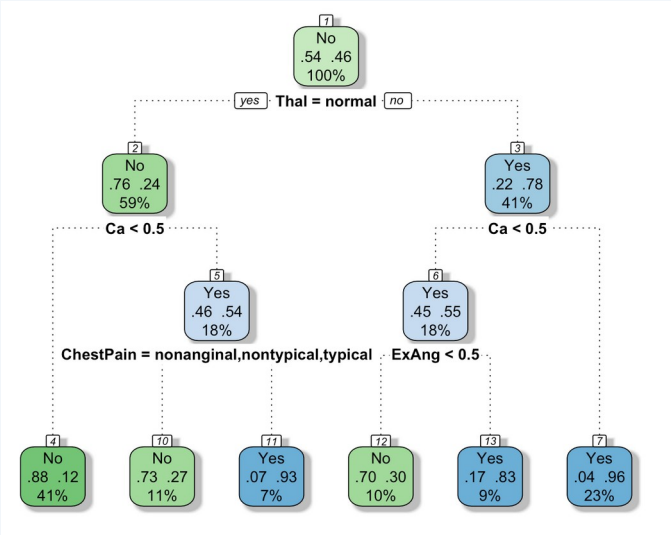
Q2. How you design the app and how it works:

Broadly, the shiny app has been divided into three tabs: (a) Decision Tree (b) Random Forest and (c) Additional Information. To prevent ambiguity, information about the primary purpose of the heart data set and its attributes are clearly communicated to the user across the app. In the first two tabs representing the two classifiers (i.e decision tree and random forest), users can choose the desired features (by selecting the checkboxes), and make changes to the hyper parameters (by moving the sliders). Once selected, you simply click 'create model' to generate your model, and click on 'test model' to test it. Training and testing results are immediately made available to assess the model accuracy. In addition, the confusion matrix summarizes the performance of the classifiers. While the decision tree is visualized for the decision tree classifier, the same is clearly not feasible for a random forest classifier. Hence, a variable importance plot is displayed, helping the users identify variables with higher importance. Finally, we also included an additional information tab (displaying youtube videos and key words) for users seeking additional resources to better understand the two classifiers.

Q3. Analysis of the results obtained from the two classifiers:

Since the user chooses the hyper parameters, our decision tree and random forest classifier results are not static. Hence, we arrived at a decision tree and random forest by taking all the parameters, using cross-validation and fixing a random seed. We were able to arrive at an accuracy of 72.64% in the case of a decision tree and 81.44% in the case of a random forest classifier.

Decision Tree Result analysis:



- Here, Thal (Thalessemia) has been identified as our root node. From the decision tree, we could identify three possibilities of being classified with AHD:

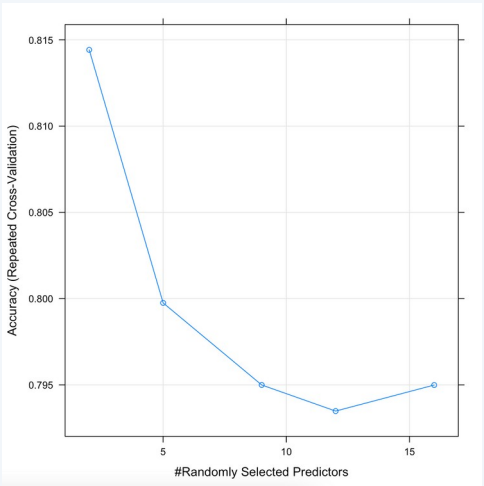
  (a) If Thalessemia levels are not normal and the number of major vessels (Ca) are less than 0.5.

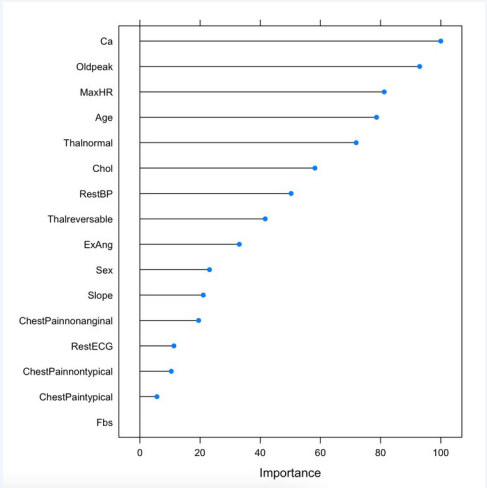
(b) If Thalessemia levels are not normal, number of major vessels (Ca) are less than 0.5 and ExAng (Exercise Induced Agina) is less than 0.5

(c) While the Thalessemia levels are normal, number of major vessels (Ca) are less than 0.5 and Chest pain= nonanginal, nontypical, typical.

Random Forest Result Analysis:

- The Variable Importance Plot identified Ca (I.e. number of blood vessels) and Oldpeak (i.e. ST depression induced by exercise relative to rest.) as the two most relevant/important attributes in our decision making process, while Fbs (Fasting blood sugar) is comparatively less important.
- Also, our accuracy vs number-of-predictors plot (attached below) identified that selecting two features in our random forest would give the highest accuracy.





References:

- https://mastering-shiny.org/action-layout.html
- https://towardsdatascience.com/building-your-first-shiny-app-in-r-82c7d1f5f309
- https://foggalong.shinyapps.io/sutton-dt/
- https://hlynur.shinyapps.io/tidy_penguins/
- https://ourcodingclub.github.io/tutorials/shiny/