1. Ionosphere Data:

Load the `mlbench` package and load in the `Ionosphere` dataset from this package. Use help and str to understand the data that was collected on radar data at hospital in Labrador.

set.seed(153)

(a) Have a look at the data, are the any variables to remove before running any models?

(b) Create a training and test dataset (70:30)

Decision Trees:

(c) Create a decision tree for the train data with Class as the response and all of the other variables bar the variables discussed in (a) as predictors. Is this a classification tree or a regression tree?

(d) Create a diagram of the decision tree created in (c). Interpret the tree diagram. Is this a useful visualisation for the data?

(e) Using print function or otherwise, answer the following about the decision tree created in part (c):

How many terminal nodes are there?

What is the minimum number of observations in these terminal nodes?

(f) Use the predict function to find the predicted values for the test dataset. Create a confusion matrix. What is the accuracy of this model?

(g) Create a ROC plot to show the sensitivity vs specificity of the model. Find the Area Under the Curve (AUC). Interpret this.

Ensemble techniques Trees:

(h) Use the bagging function on the training data to predict Class. What are the important variables used in this technique.

(i) Use the predict function to find the predicted values for the test dataset using the model in (h). Create a confusion matrix. What is the accuracy of this model?

(j) Use the random forest technique on the training data to predict Class. What are the important variables used in this technique.

(k) Use the plot function on your output of the randomForest function in (j). What does it tell you?

(l) Predict the response for the test set and create the confusion matrix and calculate the accuracy. How does it compare with the model obtained for bagging in part (h)?

(m) Perform boosting on the training data to predict Class. What are the important variables used in this technique.

(n) Predict the response for the test set and create a confusion matrix. What is the accuracy of this model? How does it compare with the bagging and the random forest models?

(o) Create a ROC plot to show the sensitivity vs specificity of all the models from the previous section. Find the Area Under the Curve (AUC) for each. Interpret this.

Support Vector Machines:

(p) Have a look at the data briefly, do you think a linear or radial kernel is more appropriate given the visualizations.

(q) Use tune.svm to select the best hyperparameter values for the svm model with the kernel selected in part q. Run this model on training dataset

(r) Predict the response for the test set and create a confusion matrix. What is the accuracy of this model?

Overall:

(s) Based on all the models performed here and the different measures of performance, which of these techniques would you recommend, giving reasons.