DS 861 | Final Exam

Description of the data:

The data you will use for this exam is the vaccine response.csv data set.

The data set consists of the clinical measurements (cytokines) from children of a variety of age (in days). Each cytokine’s name is in the first row of the data file. The goal is to use this cytokine

Information to predict whether a child belongs to either one of the three group: Low vaccine response (LVR), Neutral Vaccine Response( NVR), or High Vaccine Response (HVR). The group information is stored in the VR variable.

For this final exam, you are required to complete the following tasks (in order). Unless

specified, you are welcomed to use any functions or tools.

Specific tasks:

Here are the specific tasks you need to complete:

1. Preprocess data: There are some missing values in the data set. For each missing value, you can impute that value with the MINIMUM value of the corresponding column. Remove all children that are older than 120 days. Report the dimensions of the final data set (10 pts).

2. Inferences (10 pts)

Consider a binary classification problem with response variable of values LVR and HVR. Build a logistic regression model using all features of cytokines and all observations. Provide a list of significant variables and interpret them.

3.(30pts) Build a bagging tree, a random forest, and a boosting tree model to make predictions of response variables of all 3 values: LVR, NVR, and HVR. Perform prediction on the testing set. Report the accuracy score. Report the important variables for each model.

You are required to tune at least 2 hyperparameters for each model.

You are required to perform hyperparameters tuning using 5-fold cross validation.

Select your best model using the accuracy score.

4. PCA (20 pts)

(a) Reload the original data set, impute missing values, and drop the Age and VR variables. (5pts)

(b) Perform PCA and plot the scree plot. (5 pts)

(c) How many principal components do you need to cover 90% variability? (5pts)

(d) Plot the data set using the first two principal component PC1 and PC2. Color the data point using a color code of your choice (i.e LVR=red, NVR=green, HVR=blue). (5pts)

Extra credit problem (optional)

5. (20pts) (a) Consider a binary classification problem with response variable of values LVR and HVR. Build a lasso logistic regression model using all features of cytokines and all observations.

This includes all steps of training-validating-testing split, tune parameter…Is there any variable (cytokine) that stands out?

(b) Repeat the previous part but with response variable of values HVR and NVR.