HOMEWORK 4
BIOSTATISTICS 835
DUE OCTOBER 29TH, 2023

For this homework we're going to focus on the Indian Liver Patient Dataset. The data that we'll use here is different then what we used in our examples. You can read about the details of the predictors and outcome here.

Be sure to note that any patient whose age exceeded 89 is listed as being of age "90", for your functions below you will simply recode such people's ages as 92 (numeric). The complete data contains 583 observations, 416 patients diagnosed with liver disease and 167 patients without liver disease.

The data `ILPD.csv` is available on GitHub. You will use this data to fit all of your models. There is another dataset `ILPD_test.csv` which contains 125 observations that I removed (at random) from the complete dataset. This dataset is not available on the website.

You will answer the following questions based on this data. I recommend reading an understanding all questions before commencing with your analyses.

1. Fit a CART to this data. Prune the tree accordingly (i.e., using CV). What appears to be the most important factors in your final model?

2. Fit an SVM to the data. Choose the shape parameter of the kernal (or the two kernal parameters depending on what kernal you use) and the cost according to CV.

3. Fit a random forest model to the data, appropriately using CV to tune the parameters. What three variables seem to matter the most?

4. Fit a GBM model to the data, appropriately using CV to tune the parameters. Draw partial dependence plots for three variables you found to be the most important in the previous question.

5. What model do you think works the best? For your best model, estimate what the value of $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ would be for a new test data set with size N. Give a standard error for this value and a 95% confidence interval, i.e., using $est \pm 1.96SE$.

6. For this question, turn in a separate R program `lastname_firstname_Q5HW4.R` which:

   - Uses the data `ILPD.csv` to train the final version of the model from question 5. That is, it should not run CV, just your final model with your final tuning parameters and the same seed (if applicable). It should use the entire `ILPD.csv` dataset.
   - Reads in the data `ILPD_test.csv` and predicts for the model you felt worked the best. You will not have this dataset but I will.
   - Creates a data.frame called `lastname.firstname_error` with the **mean squared error** from the test data from your best model. This data should have 3 columns in the following order:

- `name`: which is your name in the format '`lastname.firstname`',
- `best.model`: character variable with the name of the best model (`CART`, `SVM`, `RF`, or `GBM`), and
- `est.error`: the test error.

I will run your program on my computer. It must run with zero error messages and produce the desired data frame.