ML ASSIGNMENT 6

Comparison between Classification Trees, SVM and Logistic Regression

The posted article by Yu et al utilized NHANES data from 1999-2004 to predict diabetes and pre-diabetes using Support Vector Machines. You will conduct a similar analysis using data within the NHANES package in R. For this exercise, you will try to predict Diabetes using similar (although not all) variables. The available data is also slightly different, so you likely won't get the same answers.

REMINDER: Look at the frequency of your outcome variable to check for balance

For this assignment, you will:

1. Restrict the NHANES data to the list of 11 variables below. Partition the data into training and testing using a 70/30 split.

"Age", "Race1", "Education", "HHIncome", "Weight", "Height", "Pulse", "Diabetes", "BMI", "PhysActive", "Smoke100"

2. Construct three prediction models to predict diabetes using the 11 features from NHANES. You will use the following three algorithms to create your prediction models:

a) Classification Tree

b) Support Vector Classifier (i.e. Support Vector Machine with a linear classifier)

c) Logistic regression.

3. You will optimize each model using cross-validation to choose hyperparameters in the training data and then compare performance across models.

4. Select a "optimal" model and calculate final evaluation metrics in the test set.

5. List and describe at least two limitations/considerations of the model generated by this analysis. Limitations can be analytical or they can be considerations that need to be made regarding how the model would be applied in practice.