Module 5 - Assignment

Start Assignment

- Due Sunday by 11:59pm
- Points 60
- Submitting a file upload

This assignment requires you to construct a machine learning classifier, specifically a type of decision tree.

You may use either R or Python as you prefer. Some starter scripts are provided in R below. There are a large number of online tutorials, open source scripts, and packages. Open source packages often lack support and change over time, so mileage may vary, but these scripts are more stable than social media scraping.

We will use a dataset or red and white wines to construct a machine learning classifier to predict the type of wine. Each wine in the dataset has a "quality" score ranging from 0 to 10. There are 11 variables associated with each wine and used to predict quality.

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8. Density
9. pH
10. Sulfates
11. Alcohol

Data is provided in two *.csv files one for red and one for white wines:

- **red.csv (https://jhu.instructure.com/courses/83281/files/11341294?wrap=1) ↓ (https://jhu.instructure.com/courses/83281/files/11341294/download?download_frd=1)**
- **white.csv (https://jhu.instructure.com/courses/83281/files/11341283?wrap=1) ↓ (https://jhu.instructure.com/courses/83281/files/11341283/download?download_frd=1)**

A starter R script is provided: **R_basic_ML.R (https://jhu.instructure.com/courses/83281/files/11341290?wrap=1) ↓ (https://jhu.instructure.com/courses/83281/files/11341290/download?download_frd=1)**

You assignment submission will involve several tasks:

- Task 1: Report the mean and standard deviation of the white wine quality score.

- Task 2: Plot a histogram of white wine pH.
- Task 3: Plot a scatter plot of red and white wine alcohol.
- Task 4: Construct a logistic regression, machine learning classifier, to predict wine color. Report which factors are significant in predicting wine color (p-value < 0.05).
- Task 5: Construct a random forest, machine learning classifier, to predict wine color. Use cross-fold validation techniques. Report your highest accuracy and Kappa.
- Task 6: Construct a CART, machine learning classifier, to predict wine color. Report performance. **NOTE:** There are many different CART models. Pick one and explain which one you are using. **HINT:** If using the starter R script, different models for the caret package can be found here, **available models. (http://topepo.github.io/caret/available-models.html)**
- Task 7: Write a paragraph to compare and contrast different models.

NOTE: Ensure that you include all code in a single HTML, PDF, or R Script.