Project 6 – Classification

In this project, you will build a classifier to predict the presence or absence of algae in water.

Due Date: December 5, 2019, by midnight.

Datasets: *algae_blooms.csv*.

Problem Description:

The ability to monitor and perform early forecasts of various river algae blooms is crucial to control the ecological harm they can cause. The dataset which is used to train the learning model consists of:

- chemical properties of various water samples of European rivers
- the quantity of seven algae in each of the samples, and
- the characteristics of the collection process for each sample.

What is the data science motivation for such a model? After all, we can simply analyze water samples to determine if various harmful algae are present or absent. The answer is simple: chemical monitoring is cheap and easy to automate, whereas biological analysis of samples is expensive and slow. Another answer is that analyzing the samples for harmful content does not provide a better understanding of algae drivers: it just tells us which samples contain algae.

Tasks:

1. Load the data and summarize/visualize it: you will be tasked with predicting the presence/absence of algae a1 and a2 (for tasks 1 and 2, you may use the algae bloom notebook as a starting point).

2. Clean the data, select reasonable features for the prediction task, and impute missing values.

3. Remove 20% of the observations and save them to a validation set.

4. Create a training/testing pair on the remaining 80% of the observations and train a
    a. decision tree
    b. naïve Bayes classifier
    c. artificial neural network
    d. logistic regression model, and
    e. support vector machine

    to predict the presence/absence of algaes a1 and a2. Evaluate the performance of each model. Which models performs best on your training/testing pair?

5. Repeat step 4 on at least 20 new training/testing pairs. Evaluate the performance of each model and save to a file.

6. Combine (?) all the models obtained in step 5 to make a prediction for the readings in the validation set.

7. Repeat steps 4-6 with CART models, linear regression models, and MARS models to predict the levels of algae a1 and a2.