Classification with caret

Your results should be delivered as an HTML webpage generated using R Markdown. Make sure to include all code and results along with the answers to the questions. Provide text to describe your methods and results. This should read like the Methods and Results sections of a paper.

Grading Criteria

- Correctness and completeness of code (16 Points)
- Description of process and results (12 Points)
- Answer to questions (4 Points)
- Webpage formatting (4 Points)
- Map output (4 Points)

Part 1: Classification of Wines

In the first part of this assignment, you will compare four machine learning algorithms (*k*-nearest neighbors (*k*-NN), decision trees (DTs), random forests (RF), and support vector machines (SVM)) for differentiating three different wines based on their chemical characteristics. These data are made available by the UCI Machine Learning Repository and are available at the following URL: https://archive.ics.uci.edu/ml/datasets/Wine. The data set provides examples of three different wines from the same region of Italy but from three different cultivars. There are 59 samples of Wine A, 71 of Wine B, and 48 of Wine C. Each record has 13 attributes:

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

Compare *k*-NN, DTs, RF, and SVM for classifying/differentiating these wines based on the 13 characteristics using the **caret** package.

- Split the original data (**wine_data.csv**) into training and validation sets using random sampling. The training and validation sets should each contain 50% of the samples per class.
- Optimize the algorithms using 5-fold cross validation and test 10 values for the hyperparameters using the tuneLength argument.
- Train each model. Center and scale the data and optimize relative to the Kappa metric.
- Predict the validation samples using each model.

- Use the confusionMatrix function from **caret** to obtain an error matrix and additional validation measures for each prediction.
- Obtain variable importance measures estimated with the RF model.
- Use your results to answer the following questions.

Q1: Which algorithm yielded the highest overall accuracy?

Q2: Which algorithm yielded the best Kappa statistic?

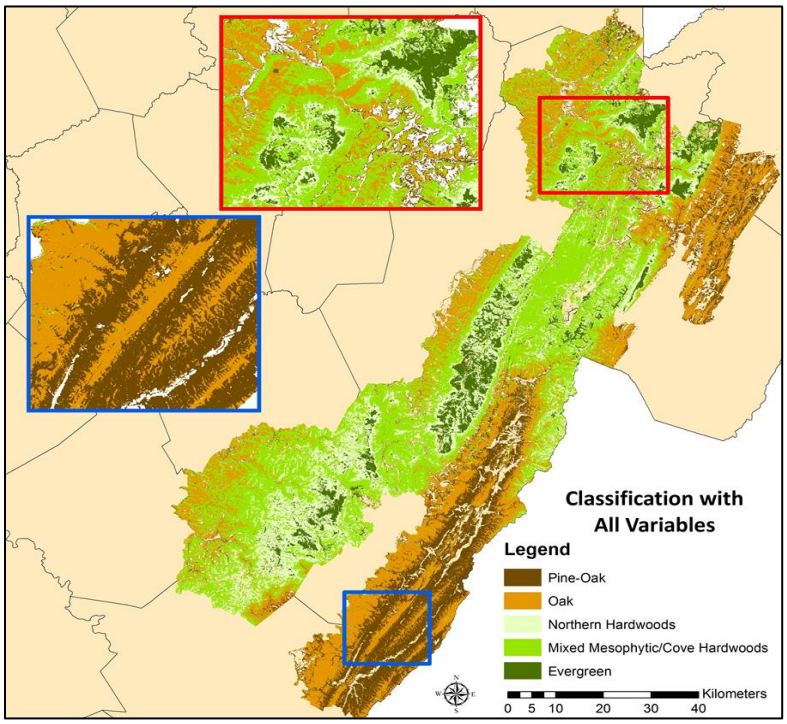
Q3: Based on the error matrices, what were the primary sources of confusion in this classification. Or, which wines were most confused?

Q4: Make a copy of the decision tree and provide it with your answers. What variables were used to split the data in the decision tree?

Q5: What variables were found to be most important based on the RF importance measures?

Part 2: Forest Type Classification

You will now classify forest cover types for a portion of the Monongahela National Forest in West Virginia. The forest types to be mapped include the following: Evergreen, Mixed Mesophytic/Cove Hardwood, Northern Hardwood, Oak, and Oak-Pine.



The following codes are used to differentiate the forest types:

Ever = Evergreen

Mix = Mixed Mesophytic/Cove Hardwood

North = Northern Hardwood

Oak = Oak

Poak = Oak-Pine

A total of 89 variables are provided to predict forest types. These predictors include spectral information derived from Landsat 5 Thematic Mapper (TM) imagery. A total of four dates of imagery are used: two from September, one from November, and one from April. The goal of using multiple dates is to potentially improve the differentiation of the forest types using seasonal changes (for example, differentiating evergreen and deciduous forests using a combination of leaf-on and leaf-off imagery).

A variety of terrain variables are also included in the analysis. These variables were derived from the 3-meter West Virginia Statewide Addressing and Mapping Board (WVSAMB) digital elevation model (DEM).

The **Band_Order.csv** file provides a description of the input data.

- Read in the training samples (**training.csv**), validation samples (**validation.csv**), predictor variables raster stack (**all.tif**), forest raster mask (**for_mask.tif**), and data description (**Band_Order.csv**) files.
- From the training set, randomly sample 100 samples of each forest type (this is a large dataset, so training on the full set will take too long).
- Train *k*-NN, DT, RF, and SVM models using **caret**. Optimize the models using 5-fold cross validation and the Kappa statistic. Test 10 values of the hyperparameters using tuneLength.
- Use the models to predict the validation samples and create confusion matrices and additional statistics.
- Use the "name" column of the **Band_Order** data to rename the raster bands in the raster stack. They are in the same order as the rows in the **Band_Order** table.
- Predict to the raster stack.
- Read the raster result back in and mask out all pixels that are not in forest extents.
- Make a map of the results using **tmap**. Assign labels and colors to each class. Provide a title and legend.
- Obtain variable importance measures from the random forest model.
- Answer the following questions:

Q6: Which algorithm yielded the highest overall accuracy?

Q7: Which algorithm yielded the best Kappa statistic?

Q8: Which forest types proved most difficult to map?

Q9: Based on the error matrices, what were the primary sources of confusion in this classification. Or, which forest types were most confused.

Q10: What variables were found to be most important based on the RF importance measures?