Classification Using Machine Learning and GEOBIA

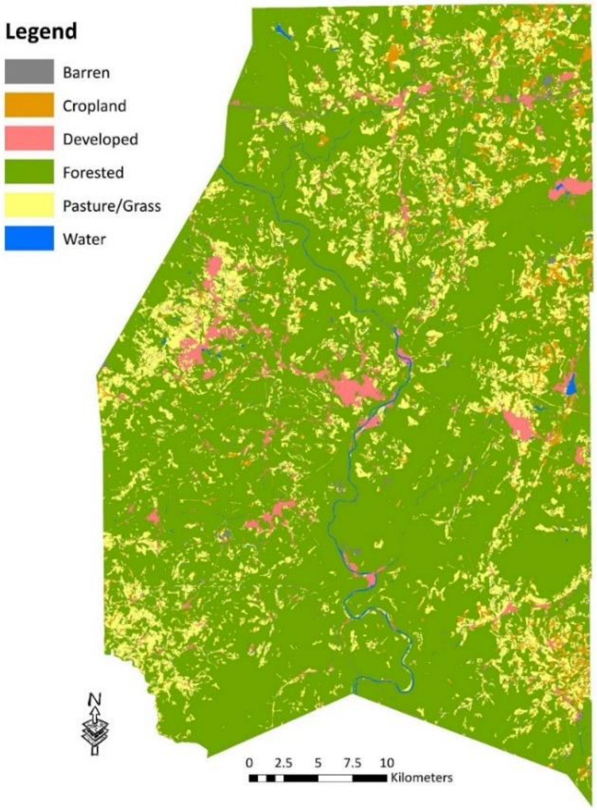
Your results should be delivered as an HTML webpage generated using R Markdown. Make sure to include all code and results along with the answers to the questions. Provide text to describe your methods and results. This should read like the Methods and Results sections of a paper.

Grading Criteria

- Correctness and completeness of code (16 Points)
- Description of process and results (12 Points)
- Answer to questions (8 Points)
- Webpage formatting (4 Points)

In this assignment, you will perform a classification of land cover for image objects using the random forests (RF) and support vector machines (SVM) algorithms. You have been provided with training data (**training.csv**) and validation data (**validation.csv**) as CSV files. You will also predict out to all of the image objects in the study area extent (Preston County, WV). You will also experiment with the impact of feature selection and training data balancing on the model accuracy.

The first column in the tables ("class") represents the land cover classes. Six classes are differentiated: forested, pasture/grass, barren, cropland, developed, and water. The remaining columns are the predictor variables calculated for each image object using the eCognition software. Note that I am not having you work with the original spatial vector data as the files are very large. However, results from R could be written to file and then appended to the image objects using table joins in a GIS software to generate a map or spatial output.



- Read in the data and packages.
- Use **caret** to produce RF and SVM models. Using 5-fold cross validation, optimize relative to Kappa, test 10 values for the hyperparameters using the tuneLength argument, and center and scale the data. Use the results to predict to the validation data and obtain confusion matrices and additional assessment metrics.
- Repeat the process with the same settings. However, this time implement variable selection using recursive feature elimination with random forests as implemented in **caret** based on 5-fold cross validation.
- Repeat the experiment with the same settings as the first set of models. However, this time use training data balancing with the up-sampling method.
- Write code to predict to all of the image objects.
- Answer the following questions.

Q1: What is the overall accuracy and Kappa statistic for the RF model without feature selection or training data balancing?

Q2: What is the overall accuracy and Kappa statistic for the SVM model without feature selection or training data balancing?

Q3: Which algorithm provides the best classification performance without feature selection or training data balancing?

Q4: Based on the error matrix, which classes have the lowest accuracy? Which are most commonly confused?

Q5: Did feature selection improve the classification accuracy for RF and/or SVM? Did it improve the classification accuracy of any of the classes specifically?

Q6: Did training data balancing improve the classification accuracy for RF and/or SVM? Did it improve the classification accuracy of any of the classes specifically?