Statistics for Soil Survey - Part 2 – Tree Based Models - May 2021

Before you begin, you must initialize you RStudio session. Set your working directory and library locations, if needed. Then load the required packages and set the random seed before proceeding with this exercise. Follow the script "Treemodels2021.R" to complete the exercise. In addition to the review questions listed below, please attach all screen shots and email the results to your mentor when the exercise is completed.

Section 1 – Exploratory data analysis on the West Virginia "soildata.csv" dataset

1. Describe the correlation matrix plot. Which environmental covariates are highly correlated? How does this inform your modeling approach?

Section 2 – Regression Tree model

2. How may terminal nodes are in the original model?
3. How many terminal nodes are in the pruned model?
4. What is the RMSE of the original model when comparing observed to predicted?
5. What is the RMSE of the pruned model when comparing observed to predicted?

Section 3 – Classification Tree model and model evaluation

6. For the gini based model, what is the overall accuracy? Kappa?
7. For the information gain based model, what is the overall accuracy? Kappa?
8. Which splitting criteria produced the best model?

Section 4 – Random Forest model for classification with parameter tuning using gradient descent

9. What is the out of bag error?
10. Which variable has the highest variable importance scores?
11. What are the values of the following parameters in your random forest model?
    a. Ntree
    b. Mtry
    c. Nodesize
12. How many models were trained using gradient descent?
13. What are the optimal parameters for the model?
    a. Ntree
    b. Mtry
    c. Nodesize
14. What was the change in OOB error by applying the grid search method for tuning?

Section 5 – Fast random forest with ranger

15. Which parameters are different when using ranger vs random forest?
16. How much faster is ranger for random forest modeling?
17. How much faster is ranger for random forest model tuning with gradient descent?

Section 6 – Random Forest prediction

18. Attach a copy of your NDVI and land cover maps to this doc.

19. Use the following color ramp provided to symbolize your landcover results