**Erdos group project modelling approach**

Frank Seidl, Luke Kiernan, Keavin Moore, Nicholas Barvinok, Noah Rahman

* Based on our recent exploratory data analysis, we find correlations in the GEDI-L2A dataset between various features.
* In particular, we find that some/all of the 101 “relative height” metrics --- which describe the tree canopy height as well as its vertical structure --- can somewhat predict water persistence, tree cover, and urban proportion.
* We would like to predict “land usage” as a classification problem; we can distinguish between urban and vegetated regions using the GEDI data and combining it with land usage data for New York state.
* If needed during model training, we can reduce the dimensionality/scale some features, and create synthetic features by combining one or more features in the L2A dataset.
* Although our dataset is mainly comprised of numbers, we can make this a classification problem through inclusion of a second “land usage” dataset, from the MRLC-WMS webpage for the contiguous United States.
* We will focus on regression models and classifiers, specifically k-nearest neighbors and a random forest model. We can compare these models seek optimal hyperparameters for each corresponding model.
* We can perform cluster analysis on the many “relative\_height” metrics, i.e., a k-nearest neighbors model.
* We can cross-validate our model selections to aid in choosing the best predictive model of “land usage”, i.e., urban or non-urban.
* We can assess the predictive power of our model during cross-validation, especially its accuracy score (ideal for classification problems), to inform our model selection.
* Our entire modelling approach will likely be an iterative process and will involve cross-validation of various models using various optimal hyperparameters.
* Further to the above, we could test logistic regression. If our models are overfit due to the lack of distinct correlation between features, we could attempt bagging and pasting.
* Although complicated, principle component analysis may prove to be invaluable in this exercise: since the 101 “relative height” features provide a map of tree canopy cover, we could reduce the number to the most relevant features for model training.
* PCA could be especially useful since our exploratory data analysis doesn’t show any clear relations (currently --- but this was only using rh\_100) other than those stated above.
* A convoluted neural network may be useful as well, if we need to continue testing different modelling approaches --- since we have many features, a neural network predicting various hyperparameters and transforming the data through a convoluted neural network process may prove to be very useful.
* We may avoid the convoluted neural network if we are satisfied with previous models and their predictions.