Comparison of Machine Learning techniques For Land Cover / Land Use Classification of Landsat Data

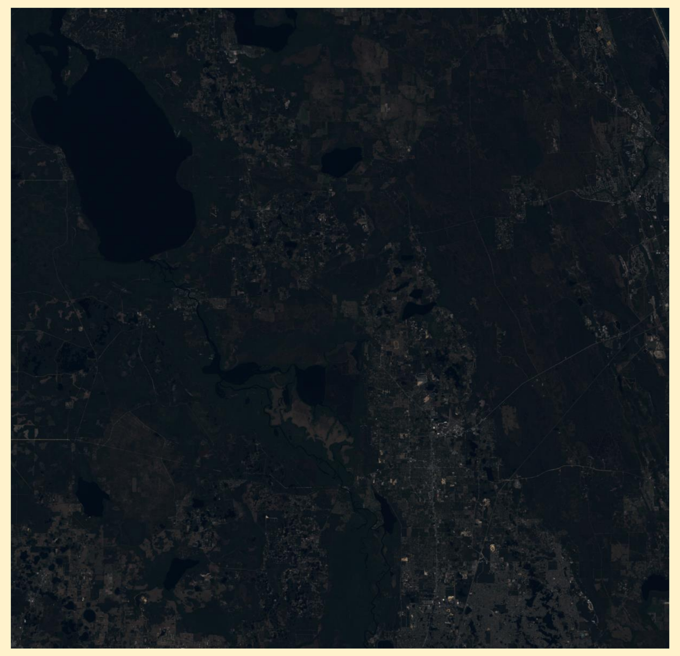
Giovany Addun, Conner McCloney, Ryley Rodriguez
Gianforte School of Computing
Montana State University

Overview

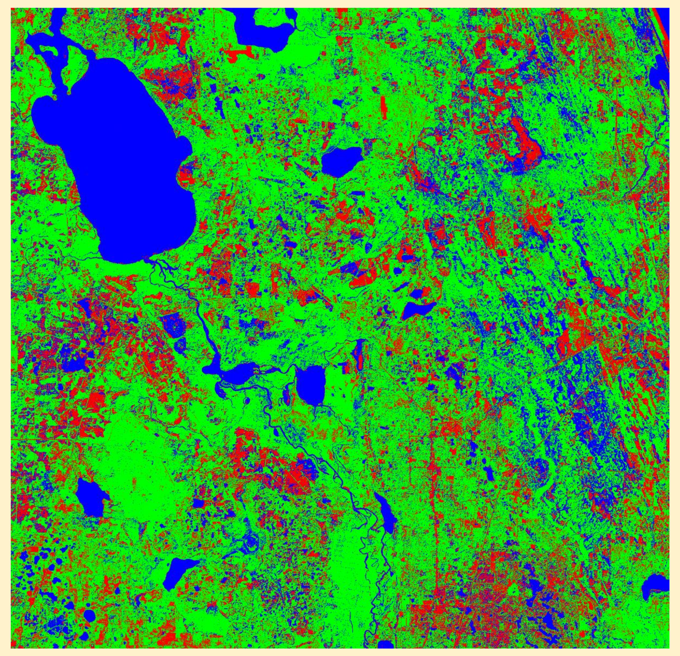
Land Cover and Land use (LCLU) classification is the process of determining what types of man-made and geological features cover an area on the surface of the Earth. Typically, machine learning techniques are used for LCLU classification. For our project we compared 5 different machine learning techniques for classification of Landsat satellite images

LCLU Classification

Original Landsat image



Classified image

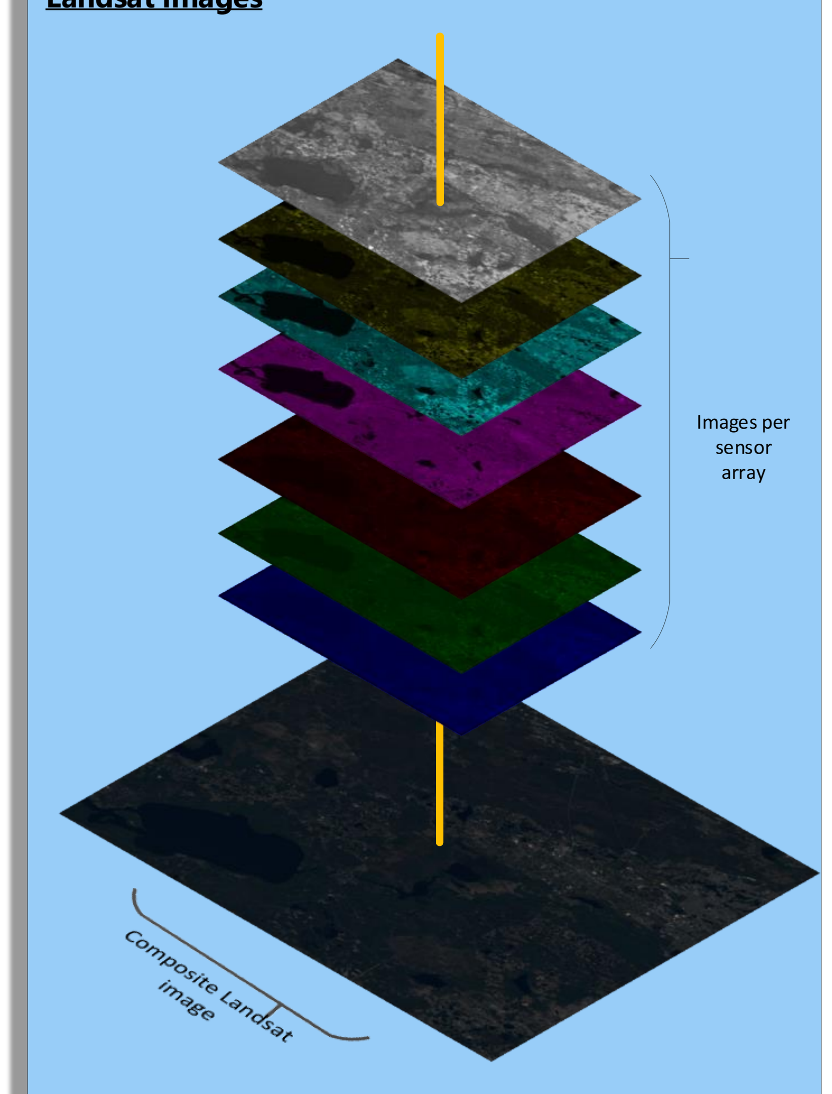


Satellite Imagery

Unlike most images which only capture the reflectance values for RED, GREEN and, BLUE, satellite images can capture electromagnetic radiation far beyond the spectrum of visible light. For this project we used data collected from Landsat satellites. There are several operational Landsat satellites, each with an array of sensors for detecting specific wavelengths of electromagnetic radiation.

For this project we used data from 7 sensors that detect electromagnetic radiation of wavelengths from 0.45μm to 2.29μm

Decomposed and Composite Landsat images



Methodology

For this project we compared the performance of 5 Machine learning (ML) models on 3 datasets, where each dataset consisted of data points collected from a Landsat image of an urban area, and datapoints consisted of sensor readings for the corresponding pixels in a Landsat image. The 5 models were compared on their ability to classify pixels in a Landsat image as 1 of 3 classes: {Developed, Undeveloped, Water} given the sensor readings of that pixel.

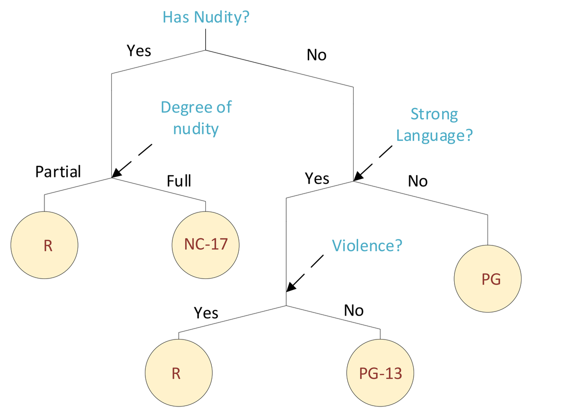
Models Compared

Decision Trees

RandomForest and Catboost Classifiers

Both Random forest and Catboost fall into the category of Decision Tree algorithms. Decision trees are among the most human-interpretable machine learning models.

Below is a decision tree for determining movie ratings.



Bayesian Methods

Naïve Bayes Classifier

Bayesian classifiers use Bayes' theorem, a statistical rule that describes how the likelihood that some event **A** occurs given that some event **B** has already occurred Is related to the likelihood that some event **B** occurs given that some event **A** has already occurred

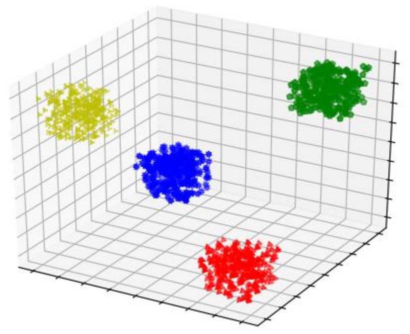
Bayesian classifiers use this rule to estimate the likelihood a **sample belongs to some class** given **the attributes of the sample** based on some correctly labeled training sample

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Nearest Neighbors

K-Neighbors classifier

K-Neighbors classifiers, like Random Forest classifiers are voting classifiers. When a new data point is given to the model, it looks at some set number (k) of the closest known data points. Each of the k neighbors "votes" on what the correct class label should be

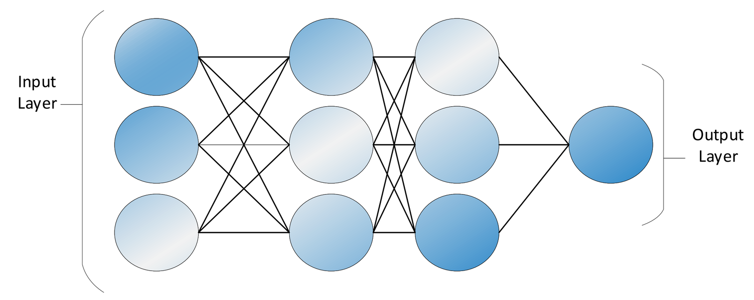


Neural Networks

Multi-Layer Perceptron

Neural networks come in many shapes and varieties, but the way they operate is generally the same. Some input is fed into the network through an "input layer" of nodes. This input travels along the connections of the network, having mathematical operations performed on it whenever it reaches a new node in the network, until it reaches the "output layer" of nodes

Multi-Layer Perceptron



Metrics used

Precision, recall and accuracy were used as metrics to compare the models. To generate testing samples by which we could compare the performance of the models, each model was run on each dataset 30 times. The metrics were then computed using 4-fold cross validation, where each of the 30 runs used a unique partition of training and testing data.

Methods of analysis

Due to the extreme variability in types of terrain that could be classified as {Developed, Undeveloped, Water}, we chose to compare the algorithms' performances over 3 different geographic locations with varied climate and geography.

Because of the nature of the data, we chose to use pairwise comparisons between each (city, algorithm)-pair, on each compared metric and also to use bootstrapping.

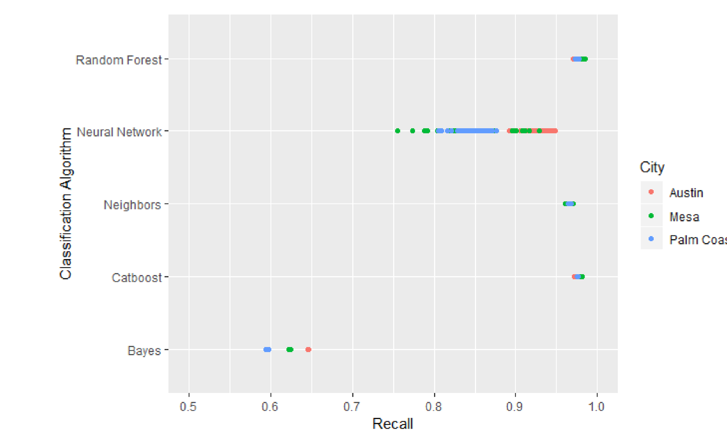
In order to test the significance of these comparisons, we tested the hypothesis that the true mean of the metric of interest for each of the compared pairs are equivalent, or:

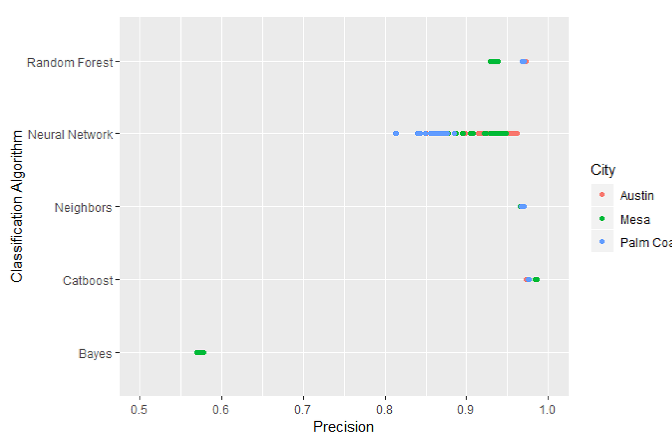
$H_0 : \mu_{i,j} = \mu_{k,l}$ where {i,k} are a given City, and {j,l} are a given algorithm, and where $i \neq k$ or $j \neq l$.

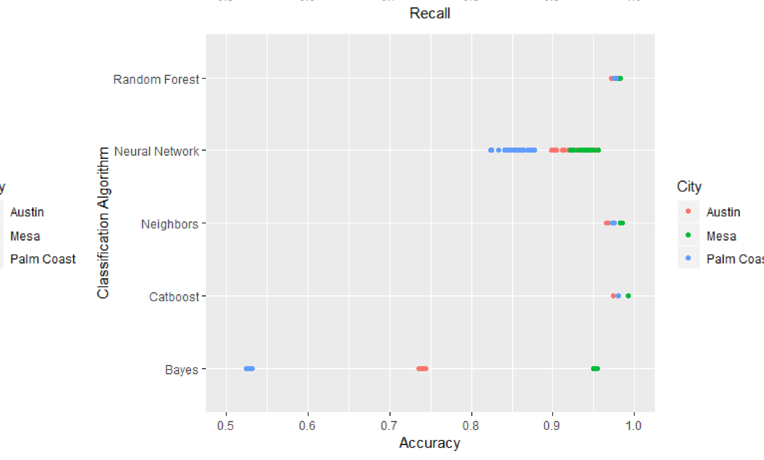
Comparison of performance

Figured below and to the right are plots for each of the metrics {Precision, Recall, Accuracy}

For each (city, algorithm)-pair, 30 independent trials were conducted. The performances for each trial are plotted







Conclusions

Out of the 415 pairwise comparisons performed, only 8 comparisons were not found to be statistically significant, with a significance level of: $\alpha < 0.05$

In terms of overall performance, the Catboost algorithm had the consistently highest metrics. However, our results indicate that there is a non-negligible relationship between performance and location where data was collected.