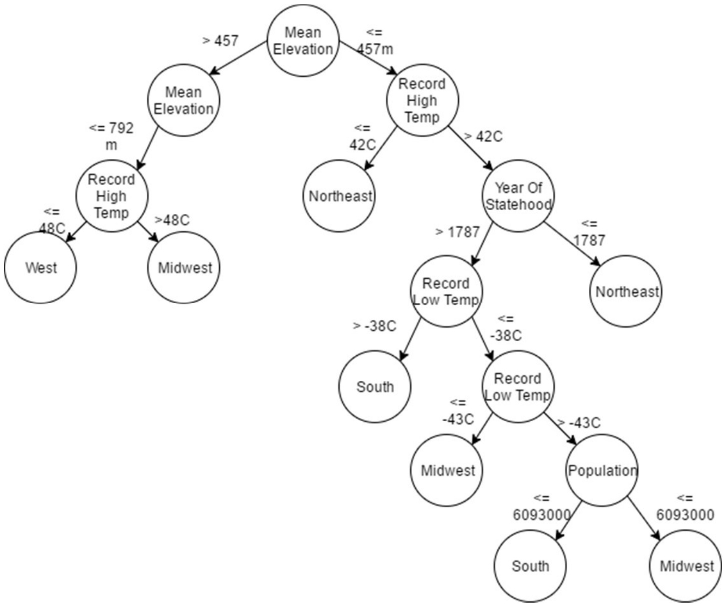
Brendan Krull

2/1/2017

CS 760 HW 1

For my dataset, I attempted to see if I could determine what region of the United States a state is in based on several attributes. The regions were defined based on the US Census Bureau, which splits up the United States into the West, Midwest, Northeast, and South. The attributes that I chose were GDP, population, population density, the year a state entered the united states, the mean elevation, per-capita income, total area, and the record high/low temperatures in each state.

After using the J48 method, I ended up with a pruned decision tree that relied on very few of my attributes:



Obviously, I expected there to be a lot of information gain from the elevation and temperature values, but I did not expect them to dominate the decision tree. Overall, the J48 method could correctly classify 66% of the states, and seemed to have the most confusion when classifying the states in the Midwest.

However, the 1-Nearest Neighbor method did not perform as well, only correctly classifying 54% of the states. Again, the method had a lot of trouble classifying the Midwestern states, as could be shown by the confusion matrix, but this method also had a lot of confusion with classifying states from the South, where the decision tree classified the south well.

The ROC Curves that follow help to paint a picture of which regions were easier to identify as the sample size increased.

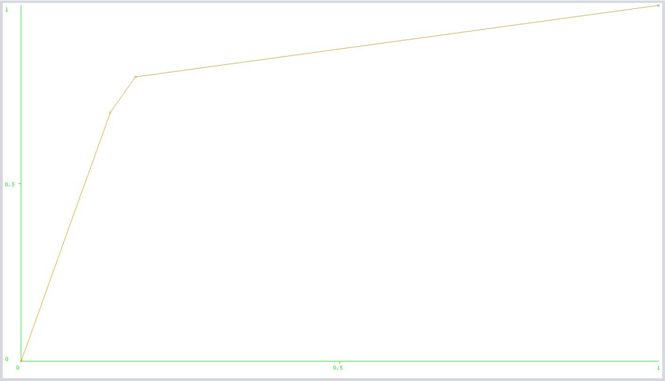


Figure 1: J48 ROC True Positive vs Sample Size for the Northeast Region

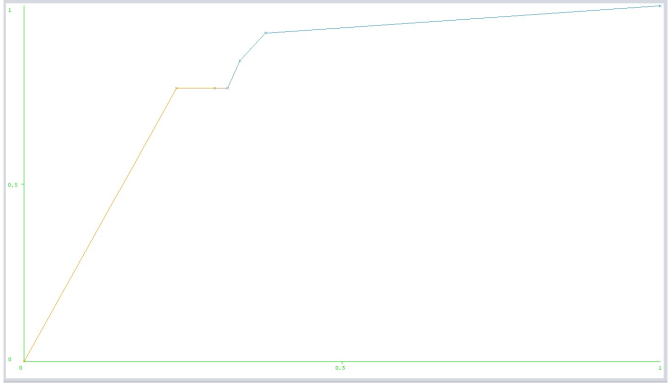


Figure 2 J48 ROC True Positive vs Sample Size for the West Region

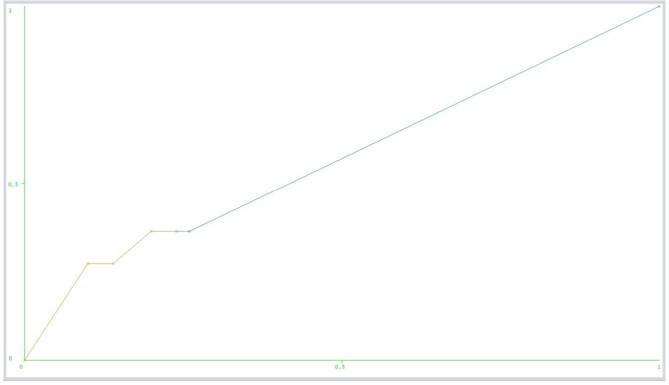


Figure 3 J48 ROC True Positive vs Sample Size for the Midwest Region

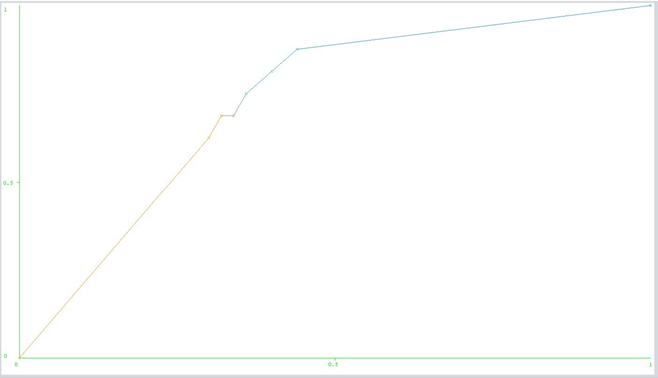


Figure 4 J48 ROC True Positive vs Sample Size for the Southern Region

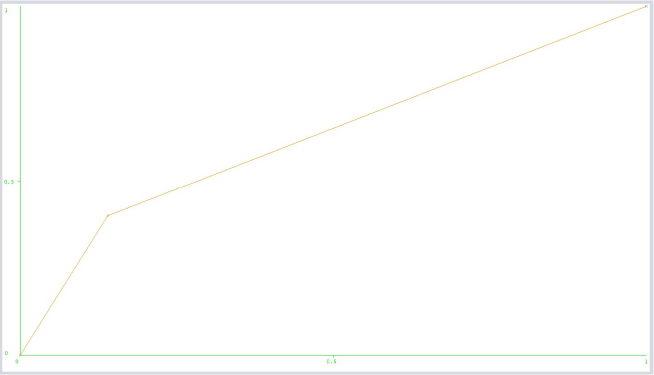


Figure 5 1NN ROC True Positive Vs Sample Size for the Northeast Region

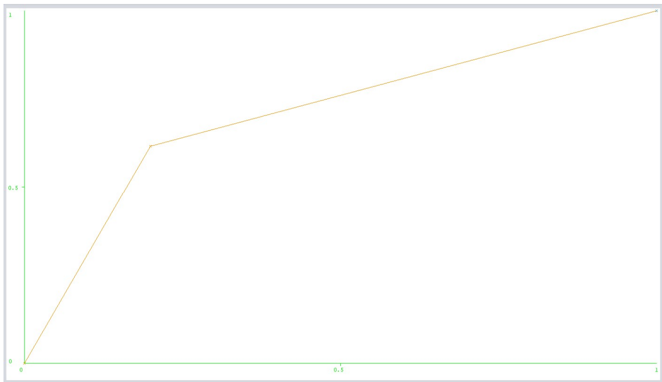


Figure 6 1NN ROC True Positive Vs Sample Size for the West Region

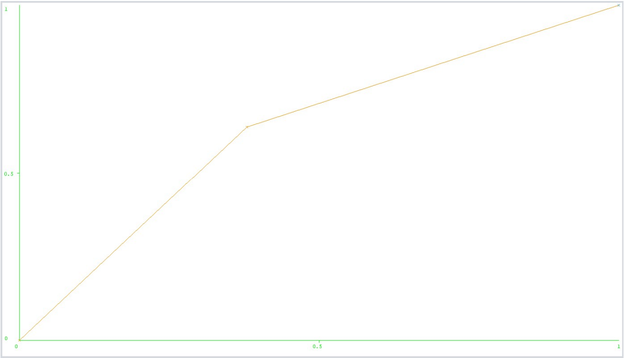


Figure 7 1NN ROC True Positive Vs Sample Size for the Midwest Region

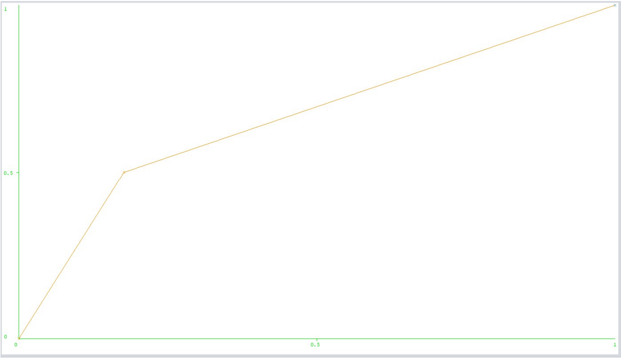


Figure 8 1NN ROC True Positive Vs Sample Size for the Southern Region