Datasets

I wanted to explore the differences between a medium-sized dataset and a large dataset when implementing the various machine learning algorithms.

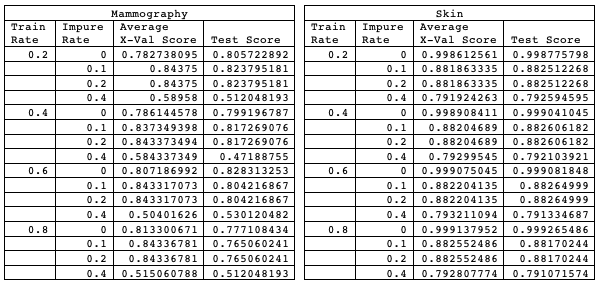
The first one chosen is based off of mammography data that is used to identify whether a tumour is malignant or benign. This dataset has 6 attributes with a binary output and consisted of 961 samples. I chose this dataset because of my interest in healthcare and because I believe the application of machine learning has the potential to play a massive role in improving patient care. The second sample chosen was of identifying whether a particular RGB value corresponds to a skin colour. This dataset only has 3 attributes and the output is also binary, however the number of samples is 245,057. My main reason for choosing this dataset was to contrast its large size to the mammography data. Its binary output also helps maintain consistency between the datasets.

Only the mammography data required pre-processing. It was reduced to 830 total samples due to missing values for some of the attributes. Some of the attributes, such as shape, were nominal and needed to be broken into multiple attributes to prevent some algorithms, like neural networks, from weighting one “better” than the other. I broke shape down into 4 mutually exclusive binary attributes: round, oval, lobular, and irregular. This way, “round” won’t be valued as a lower weight than “oval” if its index value was simply less in the mapping. I followed the same procedure for margin and density. Finally, in order to prevent the attributes from being far too specific, I transformed the age attribute into a decade attribute. This would place, for example, ages 60-69 into the 60 decade.

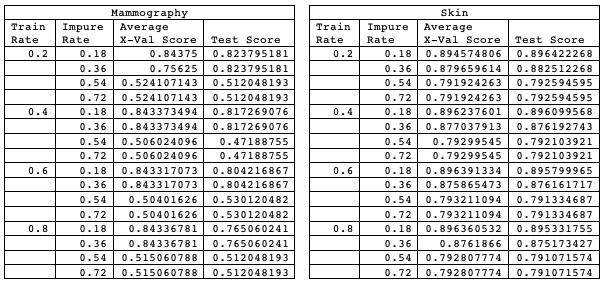
Experiment

The data was processed using the Scikit-learn library for Python. In each run, the training set made use of a 4-folded cross-validation procedure and the average was compared with the results of the testing set.

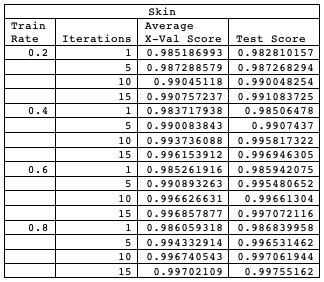
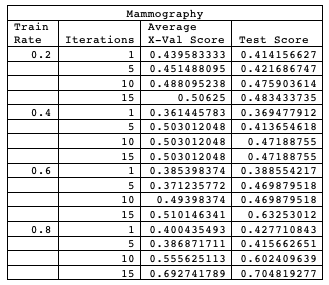
The decision tree algorithm used information gain to split the data as this is a highly recommended [CITATION NEEDED]. Pruning was applied by preventing further splits if the level of impurity that may result is less than some threshold.



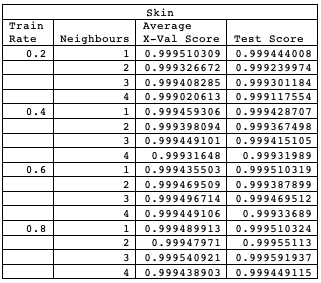
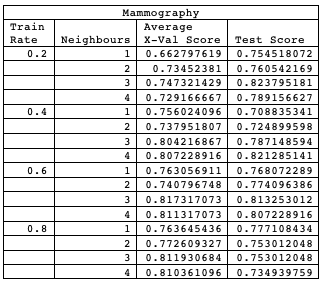
The boosted tree version of this made use of the AdaBoost algorithm to focus in on attributes that have more predictive power.



Neural network.



KNN.



SVM.

