1 Classification problems

1.1 Pima Indians Diabetes Dataset

This classification problem involves predicting whether or not a person has diabetes based on certain features.

The features are: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age.

The dataset has 768 entries.

This classification problem is interesting since it is not necessary clear whether someone has diabetes given those limited features, so there is not necessarily a clearly separable boundary. Running different algorithms in this problem should give some insight on how to find a proper boundary for such a dataset.

1.2 Predicting whether mushrooms are edible

The objective of this classification problem is to predict whether or not a mushroom is edible or poisonous, based on a set of given features.

The features are: cap-shape, cap-surface, cap-color, bruises, odor, gill-attachment, gill-spacing, gill-size, gill-color, stalk-shape, stalk-root, stalk-surface-above-ring, stalk-surface-below-ring, stalk-color-above-ring, stalk-color-below-ring, veil-type, veil-color, ring-number, ring-type, spore-print-color, population, habitat

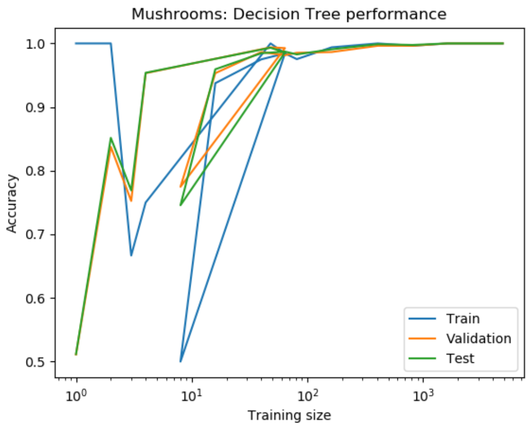
The dataset contains 8,124 data points in total.

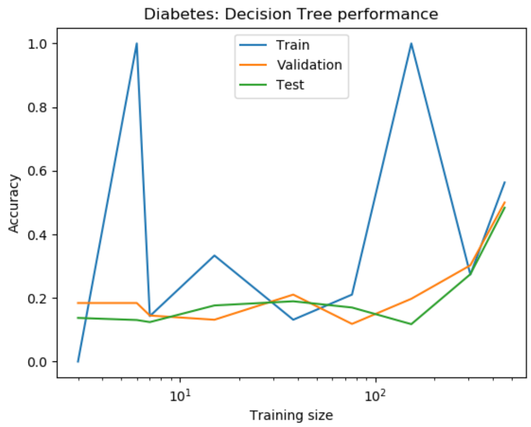
2 Results

2.1 Decision Trees

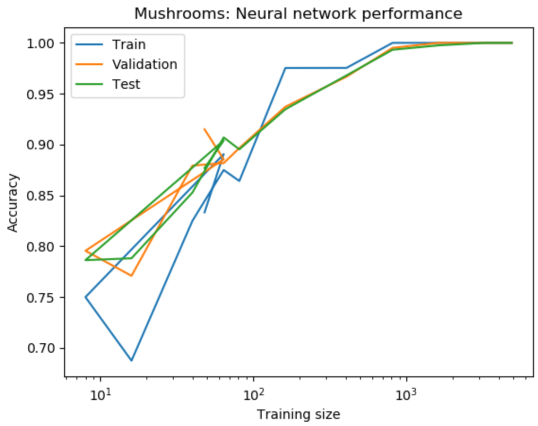
The pruning used for the decision trees was caused by setting a limit on the maximum depth.

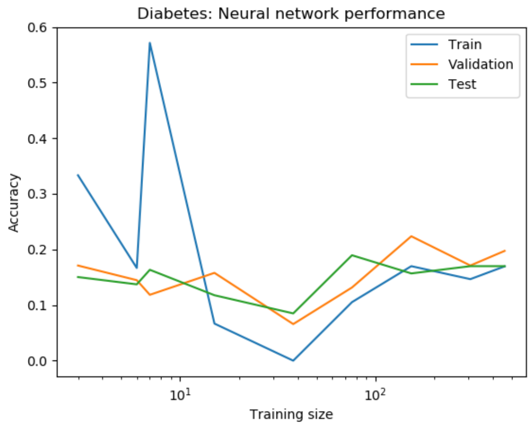
Here are the plots:





2.2 Neural Networks

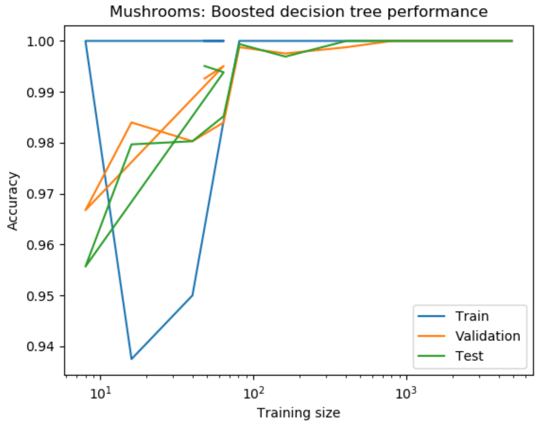


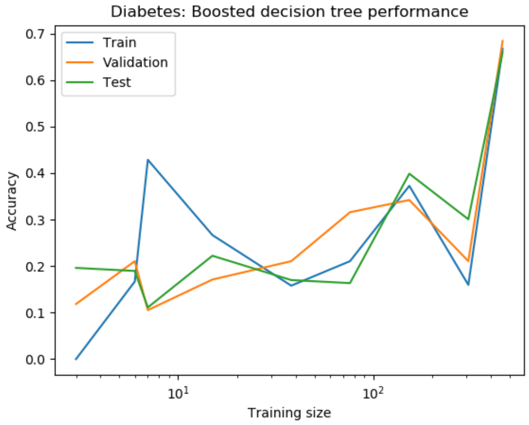


2.3 Boosting

The experiments in this project used an AdaBoost classifier with a base estimator of decision trees. The max depth of the decision trees (for sake of pruning) and the number of trees (base estimators) in the boosted classifier were varied.

Here are the plots taken for boosting:

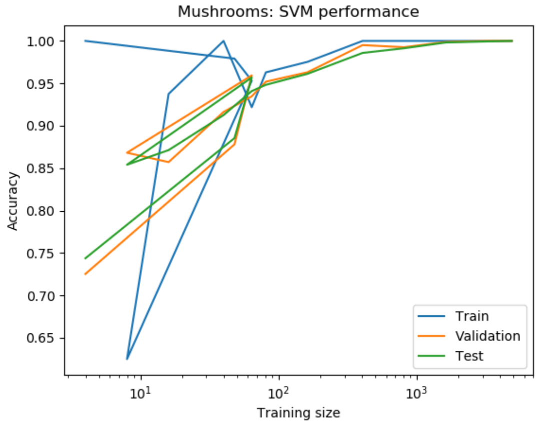


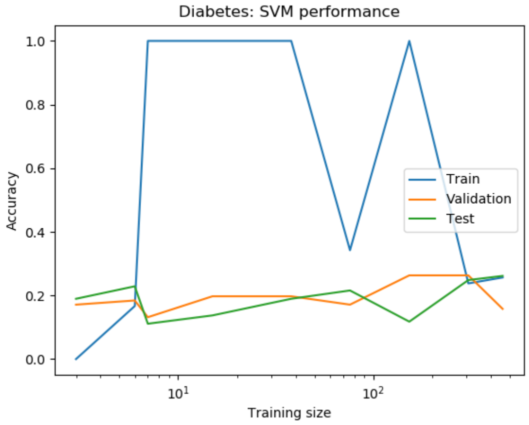


2.4 Support Vector Machines

The Support Vector Machines were tested with the linear kernel function and the RBF kernel function.

The SVM plots are shown below:

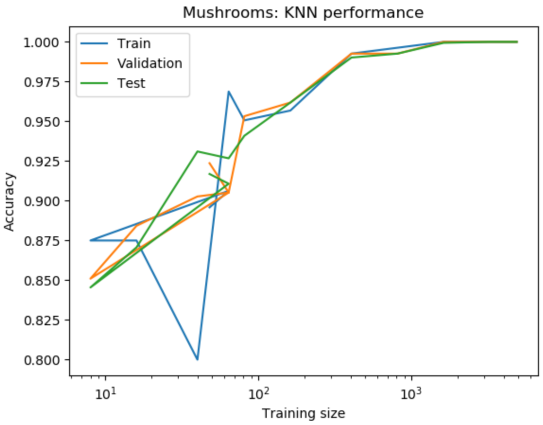


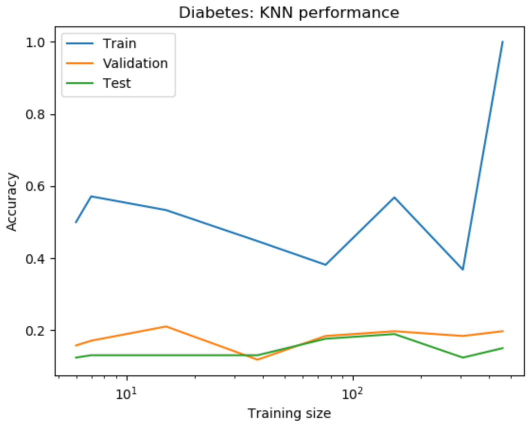


2.5 k-Nearest Neighbors

The KNN classifiers were tested with varying K values.

Here are the KNN plots:





3 Analysis

3.1 Decision Trees

Decision trees are very effective at learning the mushroom data even with a relatively small training size of 100 examples. This is because the values of the inputs are discretized, so there is a relatively small number of input combinations (a few orders of magnitude larger than 2^{25}), so several cases (at least 2^{20}, a weak lower bound for a decision tree with a max depth of 20) can be learned easily by the branching pattern of a decision trees.

3.2 Neural Networks

Neural networks were tested with multiple combinations of layers.

Below are the plots:

3.3 Boosting

3.4 Support Vector Machines

3.5 k-Nearest Neighbors