Bank Note:

Decision Tree, split on first attribute, Non-linear. Support-Vector, Non-linear kernel is good for the overlapping. KNN and Naïve Bayes not as good because of the overlapping data.

Results:

Logistic Regress actually worked really well (99.01 %). We think it is because it weighted the one attribute that had a clear split much heavier that the other attributes.

Naïve Bayes didn’t work as well (84.01 %) as we expected.

KNN didn’t have any variance in results between 5 and 7 folds (99.85%). We think this is the best because it weighted the distances between nodes in favor of the first attribute that was clustered.

Decision Tree also performed well at around (85 %)

Support Vector Machines with Linear Kernel worked as expected (97.97 %).

Fertility:

We don’t think KNN will work very well because there are so few of one class (12) and so there will be a lot more of one class surrounding the smaller class. Naïve Bayes not good because of the small data set and very small of the one class. Support Vector Machine will probably do well because it will be able to handle the overlap of this dataset.

Results:

Support Vector Machine performed the best as expected (88 %).

Naïve Bayes was third (86.1 %)

We think the decision tree didn’t perform well because there wasn’t enough data for one class in order to create good branches.

Logistic regression was the worst (84 %) because it didn’t have much features to weight heavier than any others.

KNN performed second (not as expected – (87 %). We think it was because it just did a good job on the one class with more significantly more points than the other class.

Heart:

We think decision trees will do will because it can create a strong root node from one of the attributes. KNN will do well because it accounts for certain interactions between features.

We don’t think Naïve Bayes will do very well because there aren’t any significantly differing features. Vector machine, logistic regression we expect to do well.

Results:

Naïve Bayes performed the best (84.48%). We note that Naïve Bayes has performed about the same each time.

Decision Trees performed worse (76.15 % crossing our threshold of 0.01 statistical significance). We think this is because certain attributes may have led it down a wrong path (higher false positive and false negative rates). It actually performed slightly better (>1%) unpruned.

SMO performed about (83.59 %)

Logistic Regression didn’t do as well (82.78 %). We think this is because there weren’t many distinguished features.

KNN did about 82.04% (5 NN) and 82.59% (7 NN). This is the first time that we saw a difference between our two folds.

Mushroom:

Logistic Regression we expect to do well because there are some clustered features. SMO won’t do as well because it doesn’t scale well to larger data sets (it tends towards memorization). Decision tree won’t do as well because there are a high number of attributes. KNN will do pretty well because the data will be clustered around certain attributes. We expect Naïve Bayes to perform in the middle (no significant statistical difference between any of them).

Results:

Decision Tree, SMO, and KNN, Logistic Regression all got 100% and Naïve Bayes got 95.76 %. We think this is because the data set had enough clustered features so that it could be easily classified.

Sensor Readings (4 classes):

Decision trees will not do well because of overlapped data, significantly skewed data (larger class counts vs smaller). Logistic regression will do well because it can weight vs the different classes. SMO may do well because of the non-linear data. Naïve Bayes will Perform about in the middle again.

Results:

Decision Tree 99.61%

KNN (5-fold) 87.95 %

KNN (7-fold) 87.77%

SMO 71.51%

Logistic Regression 70.42%

Naïve Bayes 52.50%