1. Description of classification problems

Red Wine Quality[i]: The first problem uses red wine quality data from the Portugal. This data has 11 features and a 0-10 quality measurement. Some features are more interesting than others. Specifically, the measure of *alcohol* is often important to wine drinkers for obvious reasons. Also, you probably never heard someone say, "This vinegar tasting wine is excellent!" One can measure whether wine has turned to vinegar by looking at the amount *volatile acidity* it has.[ii] Finally, the level of *fixed acidity* correlates to how sour a wine tastes.[iii]

A great variety of wines exist, and it's often difficult to choose. So for this problem, I seek a model that determine if the wine should be tried or not. To accomplish this, I divided the wines into two categories: wines with quality ratings of 6 or higher are worth trying and everything else is a pass. This roughly split the wine data in half.

Titanic Survivor Data[iv]: This data contains a passenger list from the infamous Titanic shipwreck. Also famous is the phrase "woman and children first." This problem explores the accuracy of this phrase by analyzing *passenger age* and *gender*. Additionally, there are many sayings about the privilege of wealth and perhaps one of them is has to do with being at the head of the life boat line-first come, first served. So additionally, *passenger cabin class*-1st being the best-is included.

2. Training and testing error

The following tables summarize training and testing error for the two classification problems. Additionally I include an absolute time measurement to visualize relative performance.

	Tree	SVM	KNN	Boosting	Neural Net
Training Accuracy	72.10%	76.50%	81.10%	87%	79.85%
Testing Accuracy	71.25%	72.91%	72.25%	72.70%	80.59%
Time	6	15	15	32	160
Variance	0.85%	3.59%	8.85%	14.30%	-0.74%

Figure 1 Red Wine Quality Performance Data

	Tree	SVM	KNN	Boosting	Neural Net
Training Accuracy	80.50%	92.20%	84.90%	80.90%	81.80%
Testing Accuracy	79.80%	79.80%	77.90%	78.70%	79.50%
Time	6	6	6	27	128
Variance	0.70%	12.40%	7.00%	2.20%	2.30%

Figure 2 Titanic Survivor Model Performance Data

Below are all the learning and validation curves for each algorithm:

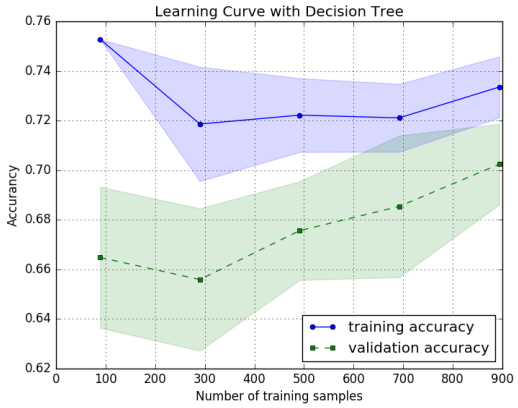

Figure 3 Red Wine Tree Learning Curve

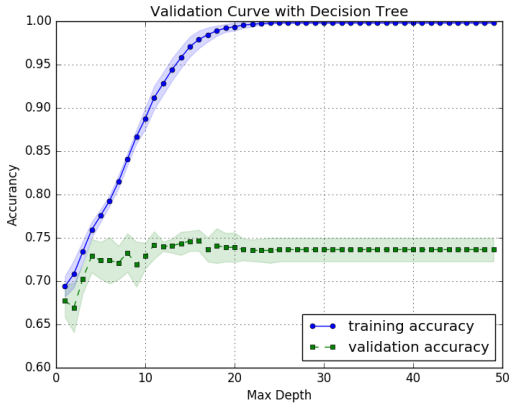

Figure 4 Red Tree Wine Validation Curve

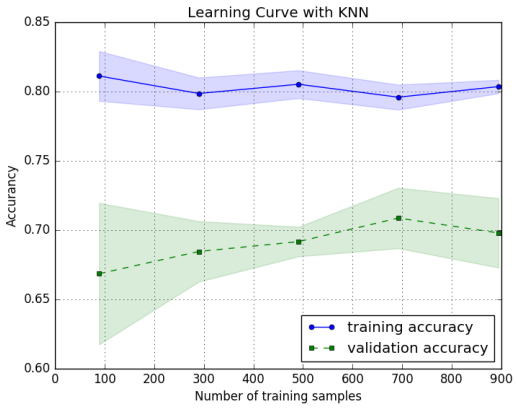

Figure 5 Red Wine KNN Validation Curve

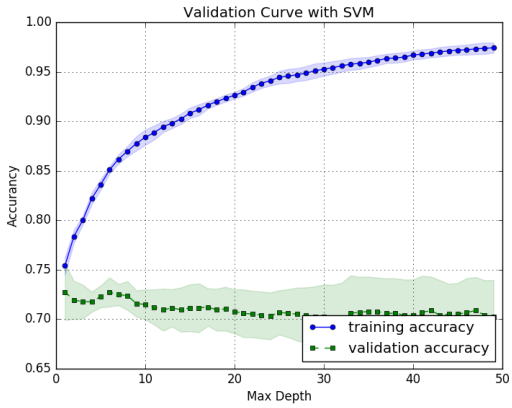

Figure 8 Red Wine SVM Validation Curve

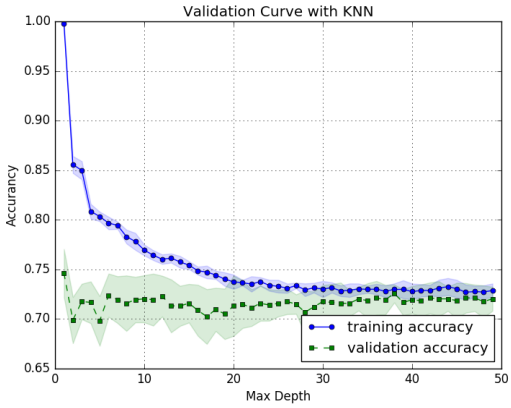

Figure 6 Red Wine KNN Validation Curve

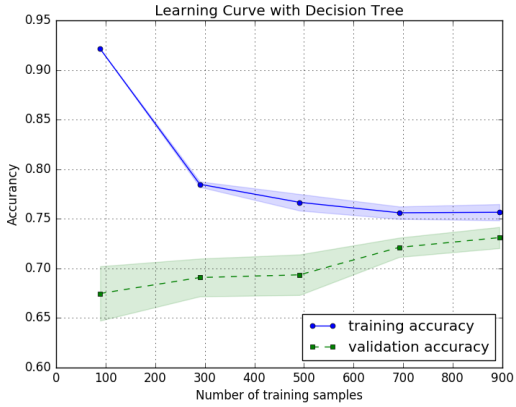

Figure 9 Red Wine Boosting Learning Curve

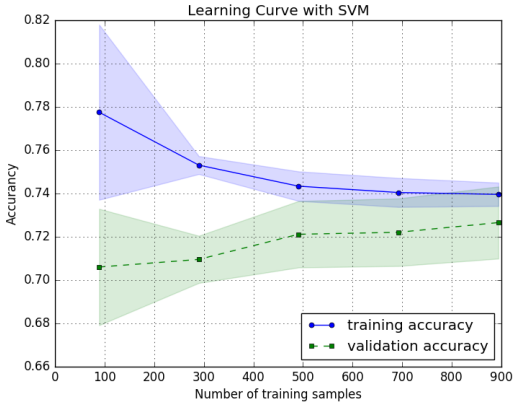

Figure 7 Red Wine SVM Learning Curve

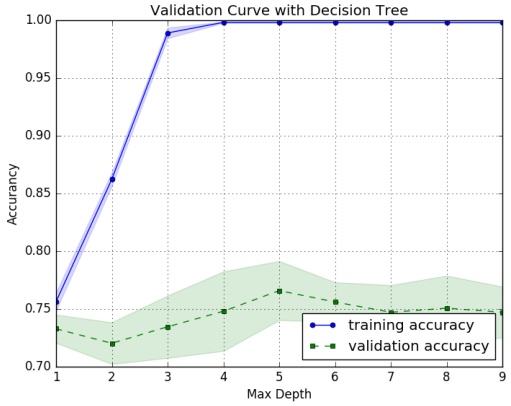

Figure 10 Red Wine Boosting Validation Curve

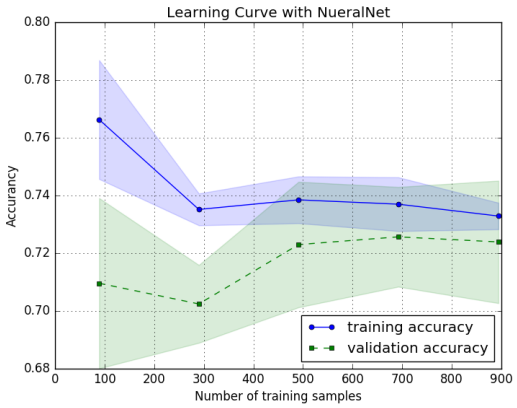

Figure 11 Red Wine Neural Net Learning Curve

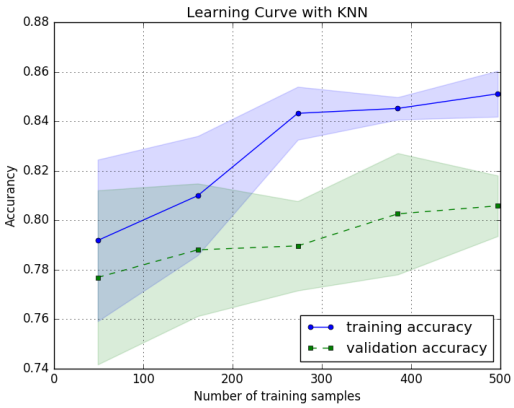

Figure 14 Titanic KNN Learning Curve

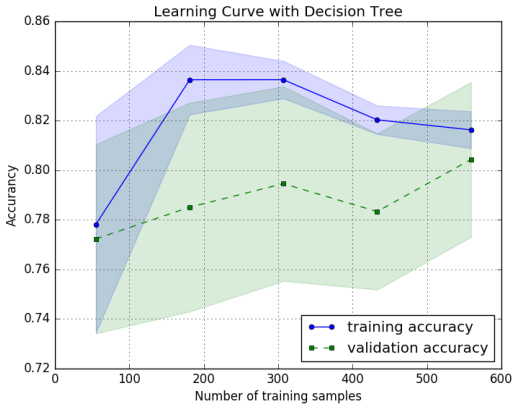

Figure 12 Titanic Tree Learning Curve

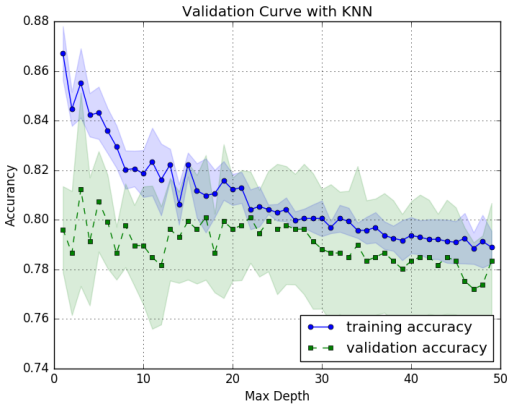

Figure 15 Titanic KNN Validation Curve

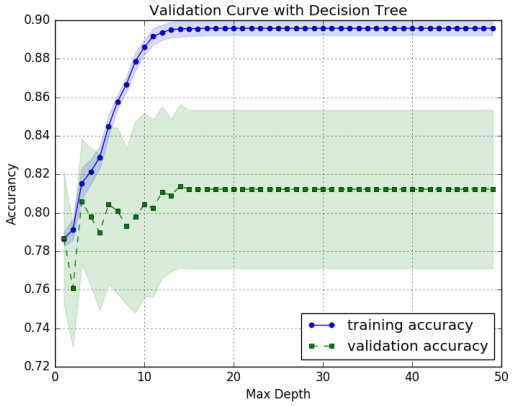

Figure 13 Titanic Tree Validation Curve

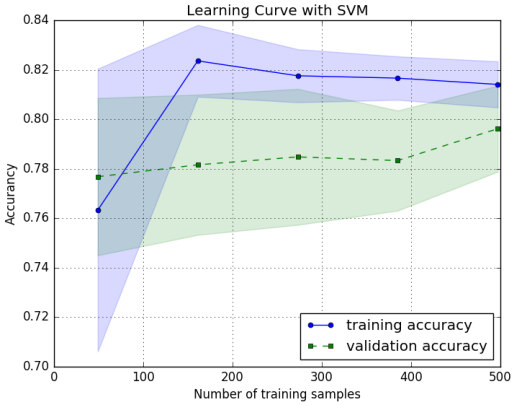

Figure 16 Titanic SVM Learning Curve

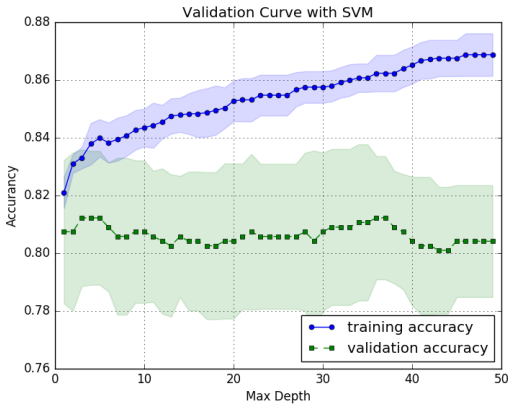

Figure 17 Titanic SVM Validation Curve

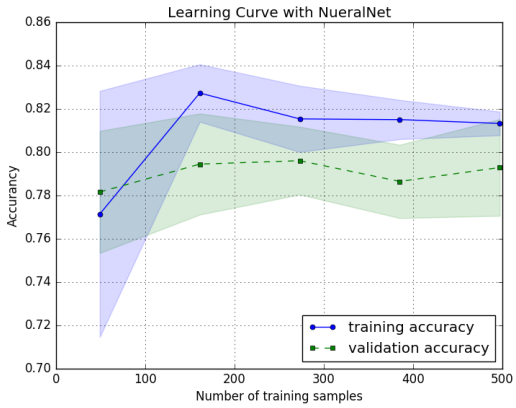

Figure 20 Titanic Neural Net Learning Curve

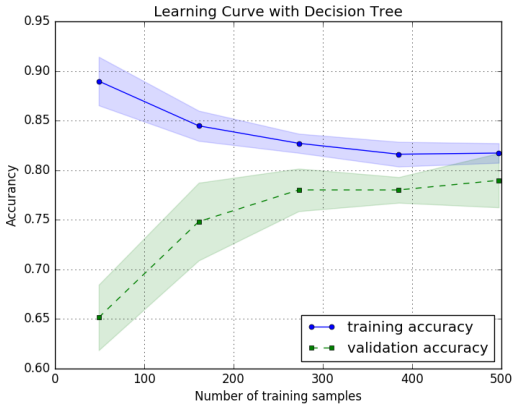

Figure 18 Titanic Boosting Learning Curve

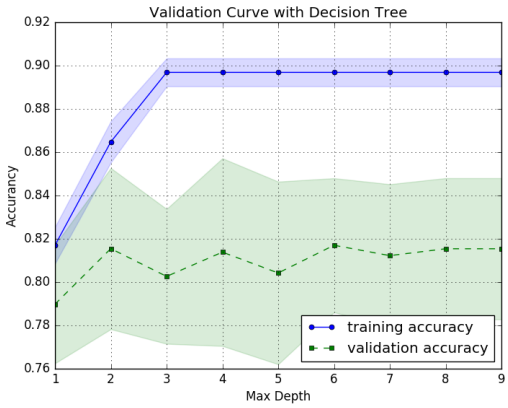

Figure 19 Titanic Boosting Validation Curve

A number of the learning curves for the wine quality classification problems demonstrate hints of bias, specifically the decision tree where the overall model performance is relatively poor at 70%. The wine decision tree learning curve also shows quite a bit of variance. The wine KNN shows a lot of variance as well. The boosting curve initially has a lot of variance, but as the max depth for the underlying tree model increases, the variance drops substantially. Since this is a week learner, there should be little risk of over-fitting despite the large tree depth.

With respects to the Titanic data, all the models performed about the same on the test data, but the SVM model shows high variance.

3. Results Analysis

Overall, the titanic model performed well with all algorithms over 80%. The simple tree model had the lowest variance. There was greater disparity on the performance of the algorithms on the wine data. I went through much iteration with the wine data set with results mostly in the 50% area. Implementing aggregation of the

quality measure such that that it became a yes or no question was helpful.

There were implementation differences as well. The tree provided a nice visual of the reasoning behind the conclusion. Additionally it was not necessary to standardize the feature set when using the tree. I did standardize for all the other algorithms.

Once, I got over the learning curve so to speak, I found that python's sklearn, and specifically the pipelining, made cross validation implementation easy to evaluate. All algorithms were implemented with both stratified kfold CV and hold out testing. I split the data into 70:30 training:testing, performing Kfold in the 70%.

Python ML is completely new to me, so I relied heavily on Sebatian Raschka's book "Python for Machine Learning". My code is largely sourced from his book with the exception of Neural Nets which uses MLPClassifer in the sklearn development version 18. Please see my code for verbose notes.

---

[i] https://archive.ics.uci.edu/ml/datasets/Wine+Quality

[ii] http://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity

[iii] https://en.wikipedia.org/wiki/Acids_in_wine

[iv] https://www.kaggle.com/c/titanic