CS7641 Machine Learning Assignment 1

Supervised Learning

Student Name: Qianwen Shi Date: Feb 9th, 2020
GTid: qshi39

1. Introduction

In this assignment, five supervised learning algorithms are applied to solve classification problems using two datasets. The analysis implemented to data are decision trees, neural networks, boosting, support vector machines and k-nearest neighbors.

2. Datasets Description

The two data sets selected are (a)wine quality data and (b)air pollution data, both from UCI machine learning repository. The reason of choosing these two datasets is air quality (especially ozone O_3) can affect the health of plants in turn impacts the wine quality. Reducing the air pollution to achieve a higher crop yield and healthier crop is one of the issues that farmers are concerned about. Details of each data set are described in the following part.

2.1Wine Quality Dataset

The wine quality data include white and red 'Vinho Verde' wine samples from a study performed by University of Minho in Portugal. In this assignment, only the dataset of red wine is used. This dataset contains 12 attributes and 1599 instances. The first 11 attributes are chemical compositions measured from the sample and the 12th attribute is the classifying attribute 'quality' scored between 0 and 10.

2.2 Air Quality Dataset

The air quality data is hourly air pollutants data from a monitoring site in Beijing based on observation in 2013 with missing data denoted as NA. This dataset contains 10 attributes and 7344 instances. The first 9 attributes are major air pollutants concentration and meteorological parameters, and the last attribute is the classifying attribute PM2.5.

3. Datasets Preprocessing

3.1 Wine Quality Data

This is a quite clean data and have no missing points therefore doesn't need much preprocessing. However, for this classification task, I have separated the 'quality' attribute into two categories when loading the data in python. With quality rating lower than 6, the data converted to 0 indicating low quality wine, while quality rating equals to or above 6 the quality data is converted to 1 indicating high quality wine.

3.2 Air Quality Data

This dataset contains 1211 NAN therefore are removed using pd.dropna() function when loading the data. The attribute PM2.5 is concentration between 2 to 443 hence we need to convert it to categories for classification problem. Based on PM2.5 data distribution, data points with value <= 22 are converted to 0 indicating relatively lower health risk and data larger than 22 are converted to 1 suggesting high health risk.

4. Experiments

The five supervised learning algorithms are all from scikit-learn in python. The training data size percentage is set to 30% for all experiments. The k-fold cross validation is used to test the performance and the k value is set to 10 for all runs. The best performance parameters for each model were found using hyperparameter tuning. The GridSearchCV tool from scikit-learn is used to perform automatic tuning. Results after tuning are presented in this report, while results before tuning are in support documents. (Note to ensure results reproduction, random_state is set to 0 for all experiments)

Table 1 shows the accuracy scores of five models before and after using GridSerachCV for tuning. There is obvious increase of accuracy for neural network, support vector machines and k-nearest neighbors after tuning when applied to wine quality data. The accuracy scores are higher for air pollution data, but tuning does not improve accuracy much except for decision tree and support vector machine model.

Table 1. Summary of Model Accuracy Scores Before and After Tuning

	Accuracy	**Decision Tree**	**Neural Network**	**Boosting**	**Support Vector Machine**	**K-Nearest Neighbors**
Wine Quality Data	Before Tuning	71.67%	**63.33%**	72.50%	**61.25%**	**66.46%**
	After Tuning	72.08%	**73.33%**	73.75%	**70.42%**	**71.25%**
Air Pollution Data	Before Tuning	**85.97%**	90.38%	91.55%	**89.00%**	92.08%
	After Tuning	**91.50%**	90.65%	91.45%	**92.24%**	92.03%

4.1 Decision Tree

To avoid overfitting issue, the decision tree was pre-pruned by limiting its maximum tree depth and minimum leaf size. If the maximum tree depth is too low, underfitting might occur and cause low scores for both training data and cross-validation tests. However, overfitting can happen as maximum tree depth increased to a certain level. After tuning, the best parameters for wine quality are max tree depth = 7 and minimum leaf size =1, and for air pollution are max tree depth = 7 and minimum leaf size = 4.

Confusion matrix can be used to evaluate the quality of classifier output. The diagonal elements are the number of points that prediction equals to true label, while off-diagonal are number of mislabeled data points by the classifier. Therefore, the higher the diagonal values the better the confusion matrix, which also means more correct predictions. Figure 1 presents the confusion matrix of decision tree results for both datasets. Decision tree algorithm for wine quality gives a better confusion matrix than that for air pollution data.

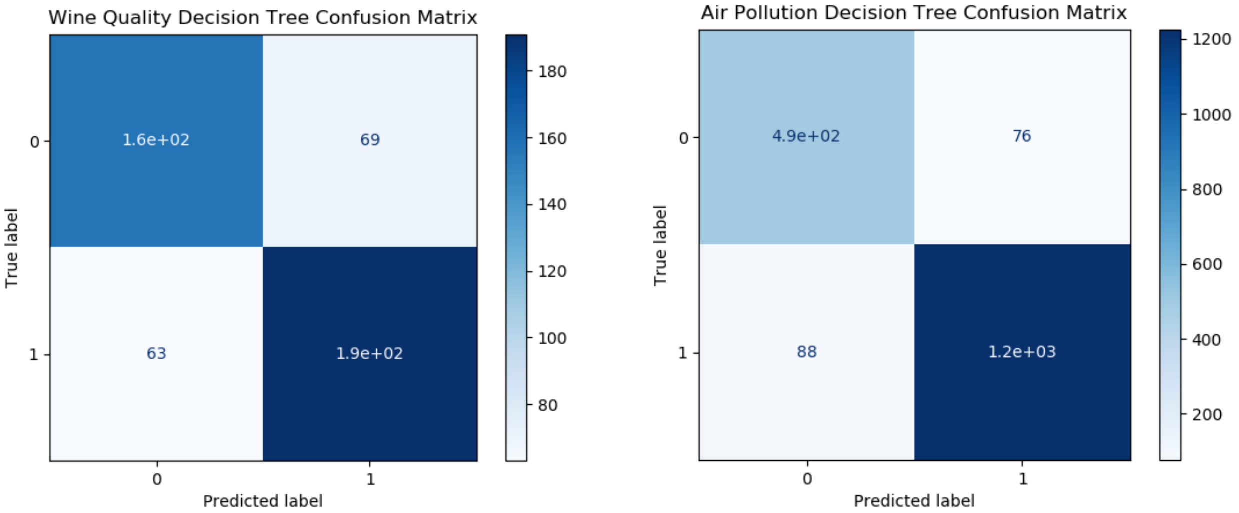

Figure 1. Confusion matrix of decision tree model for two datasets

Another way to check the classifier performance is by plotting the learning curve, which shows the how validation and training score of an estimator change as training sample size increases. This is a useful tool to see if the estimator can benefit from more training data. The learning curve and model scalability for wine quality and air pollution datasets are presented in Figure 2 and Figure 3 respectively. For water quality learning curve, the difference between training and cross-validation scores is very small and quickly converges at a small sample size (we won't benefit much by more training data). This means the classifier has a low variance. The training score and cross-validation score is around 70%, not a very high fitting. The scalability of the model shows the time needed to train the model with various training data size. For the results of air pollution data (Figure 3), it shows a higher cross validation score (86%) but the curve is flat suggesting larger training size won't increase the performance much. However, it requires more time for training compared to that of wine quality.

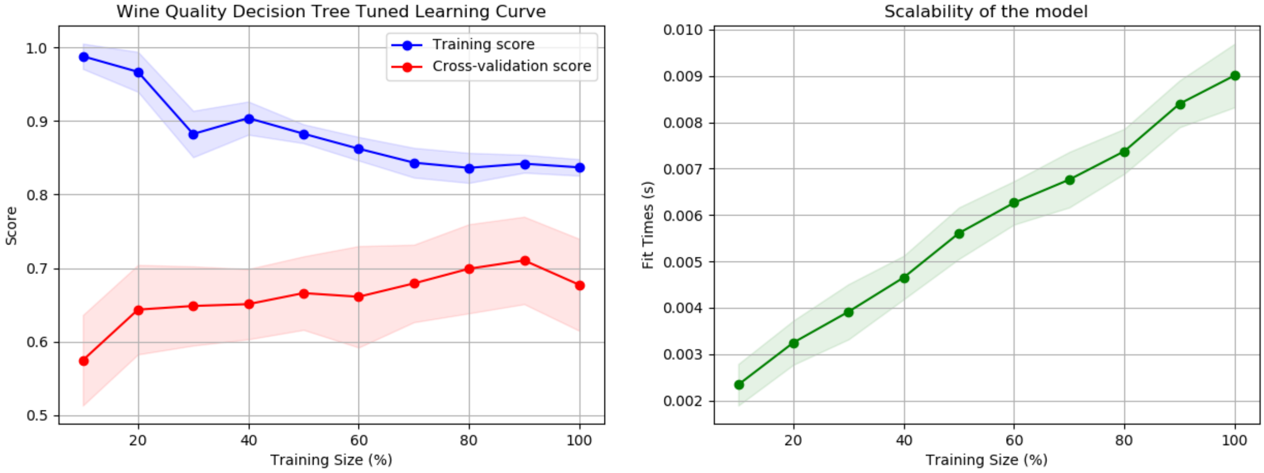

Figure 2. Learning curve and model scalability of wine quality dataset using decision tree algorithm. Dotted lines are means and colored areas represent the corresponding standard deviation.

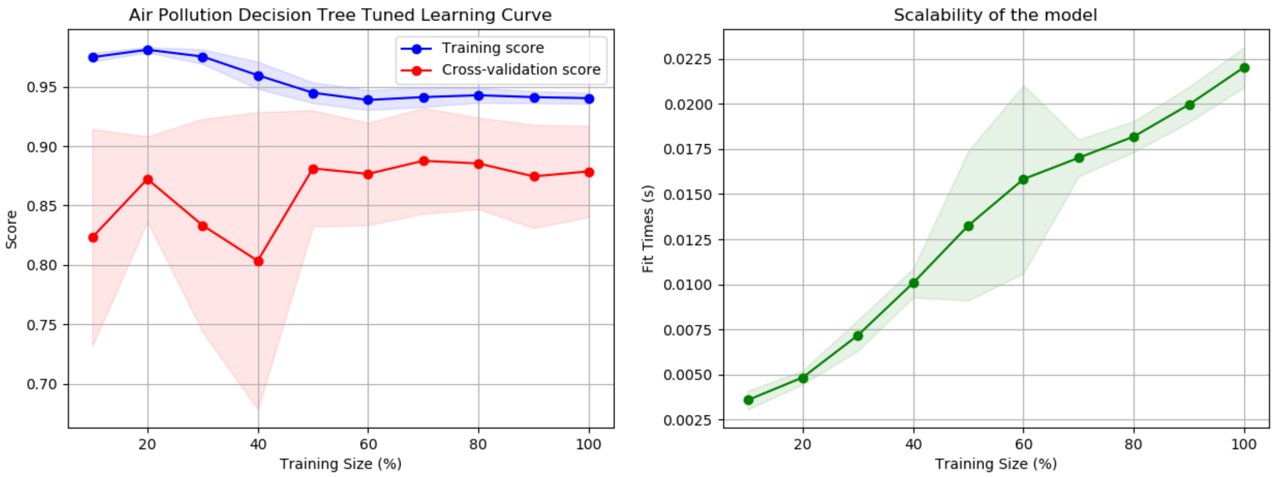

Figure 3. Learning curve and model scalability of air pollution dataset using decision tree algorithm. Dotted lines are means and colored areas represent the corresponding standard deviation.

4.2 Neural Network

Multi-layer Perception (MLP) in neural network models is used in this experiment. After trying with certain parameters, the models are tuned to one hidden layer with 50 units for wine quality data and 100 units for air pollution data. Figure 4 shows the confusion matrix is quite good for air pollution data but not so good for wine quality data. The learning curves of wine quality data (Figure 5) show a quick convergence at 20% training data size while the scalability curve increases sharply when training data size is larger than 70%. Therefore, using smaller training size for wine quality is more efficient. Figure 6 show that MLP mean cross validation score for air pollution is higher and two curves converge around 40% training data size. The scalability shows it is faster to train air pollution data than wine quality. MLP performs better for air pollution data.

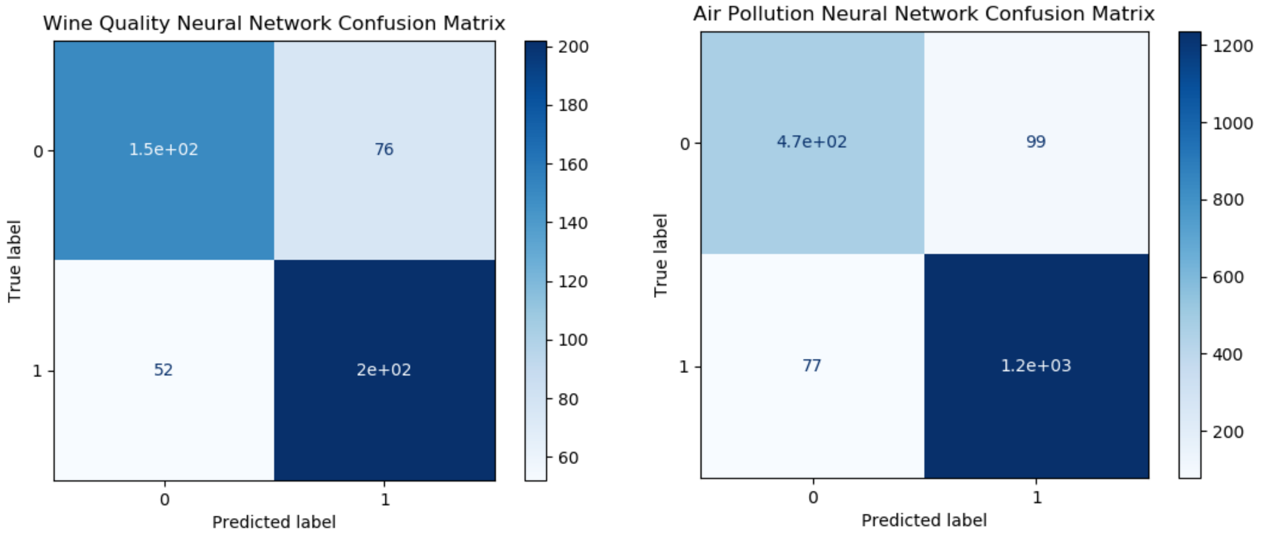

Figure 4. Confusion matrix of neural network for two datasets

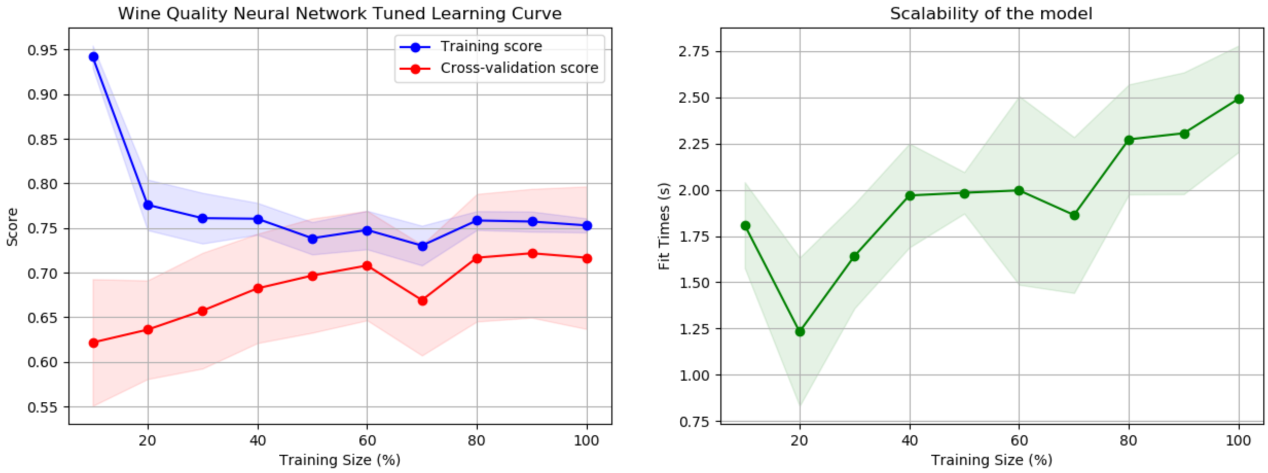

Figure 5. Learning curve and model scalability of wine quality dataset using neural network algorithm. Dotted lines are means and colored areas represent the corresponding standard deviation.

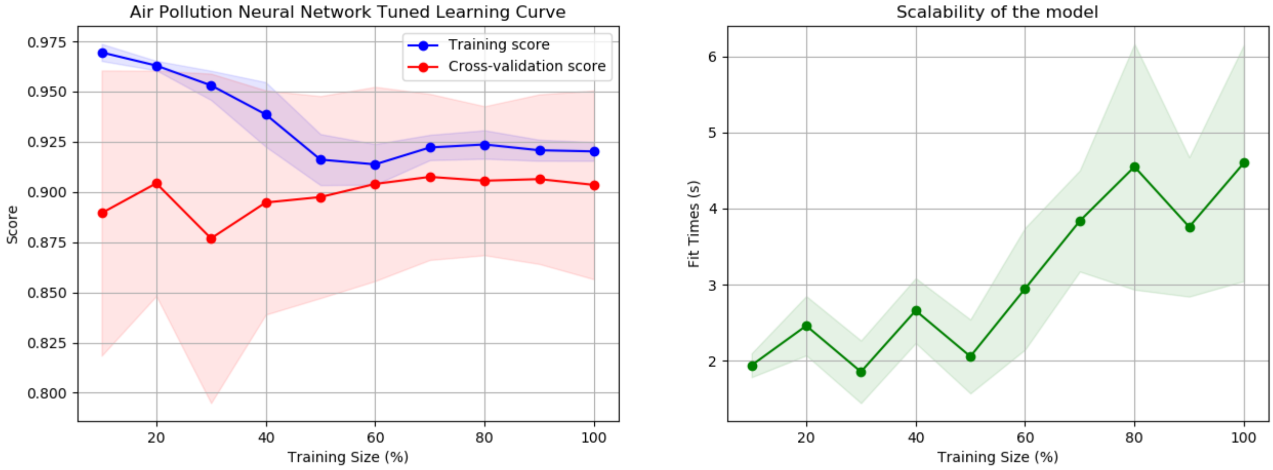

Figure 6. Learning curve and model scalability of air pollution dataset using neural network algorithm. Dotted lines are means and colored areas represent the corresponding standard deviation.

4.3 Boosting

Adaptive boosting (AdaBoostClassifier) is implemented as the boosting algorithm for these two datasets. The base estimator that boosted ensemble is built on is decision tree classifier. Figures 7 to 9 show the results of AdaBoost classifier after tuning. The parameter tuned is n_estimators, which is the maximum number of estimators at which boosting is terminated. After hyperparameter tuning, the best performance is found with n_estimators set to 9 and 73 for wine quality and air pollution data respectively. Figure 7 shows that confusion matrix looks good for both datasets as diagonal values are large. Figure 8 and 9, for two datasets, the learning curves converge at low sampling size, therefore adding more training data won’t provide much benefit. Figure 8 has lower cross-validation scores than Figure 9, suggest AdaBoost performs better for air quality data. The scalability plot suggest that it takes longer to train air quality data, and the time needed increases quickly as training data size increases.

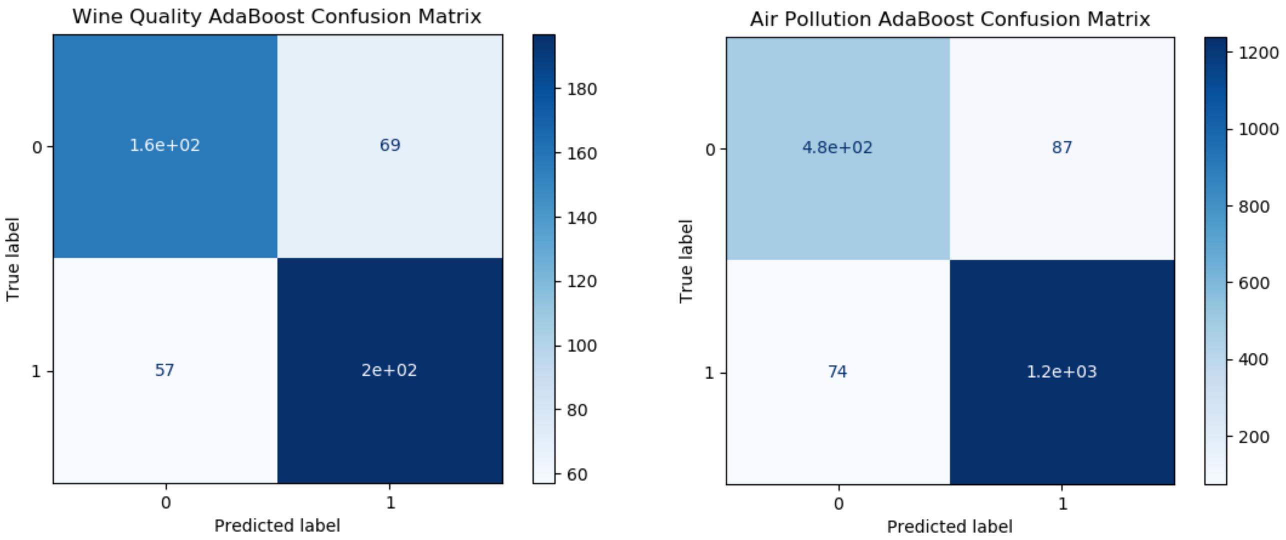

Figure 7. Confusion matrix of AdaBoost for two datasets

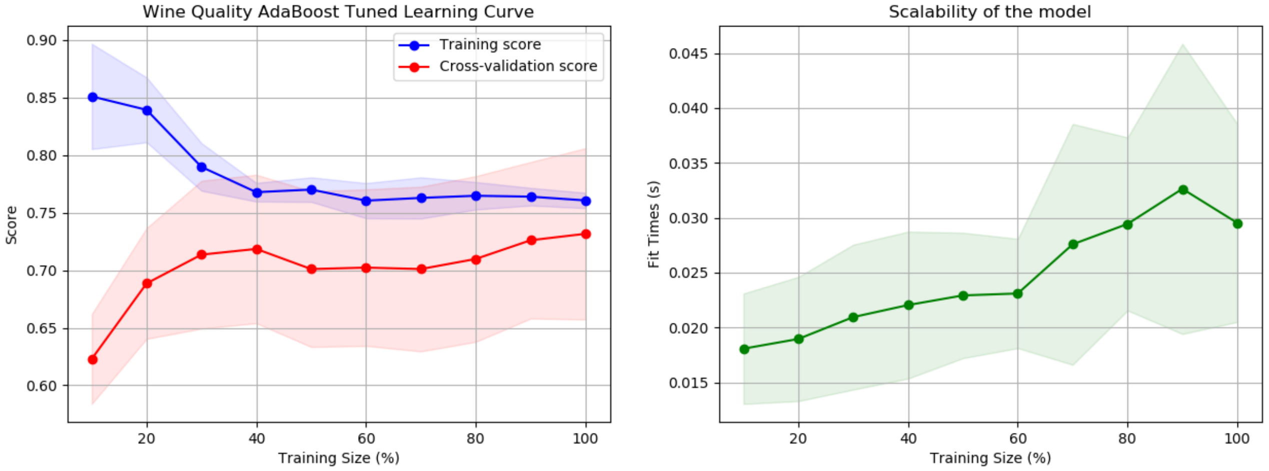

Figure 8. Learning curve and model scalability of wine quality dataset using AdaBoost algorithm. Dotted lines are means and colored areas represent the corresponding standard deviation.

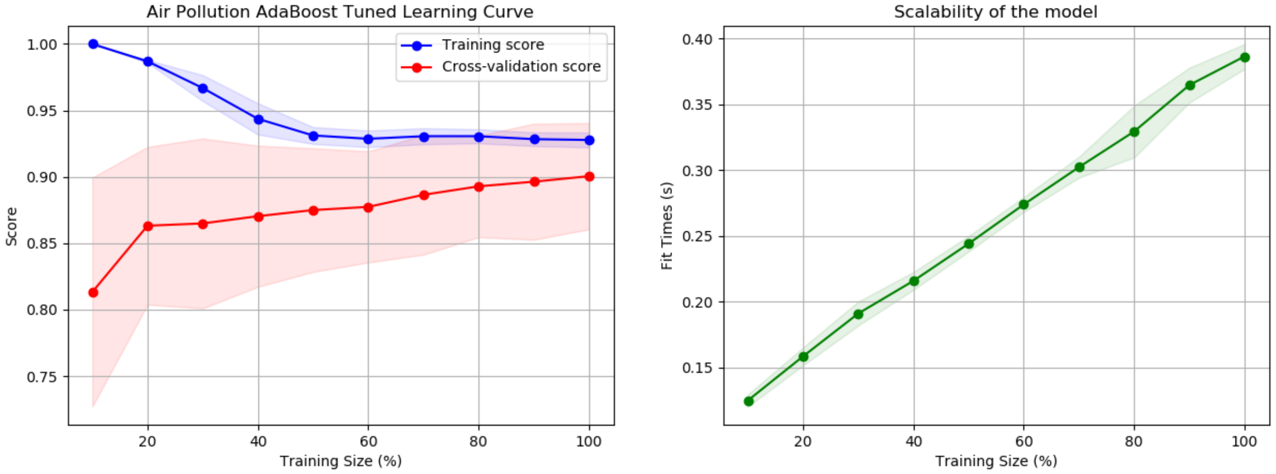

Figure 9. Learning curve and model scalability of air pollution dataset using AdaBoost algorithm. Dotted lines are means and colored areas represent the corresponding standard deviation.

4.4 Support Vector Machines

The analysis is performed using C-support Vector Classification (SVC) in support vector machines with kernel set to radial basis function (rbf). This kernel is used to choose a non-linear decision boundary for our classification problem. The tuning parameters are penalty parameter C of the error term and gamma (for rbf). The best parameters for wine quality and air quality datasets are: C=10, gamma=0.01; C=1, gamma=0.001 respectively. Results are summarized in Figures 10 - 12. Learning curves converge quickly at smaller training data size for both datasets, air pollution has a higher cross-validation score. The scalability curve looks similar for two datasets, but it will take longer to train air quality dataset.

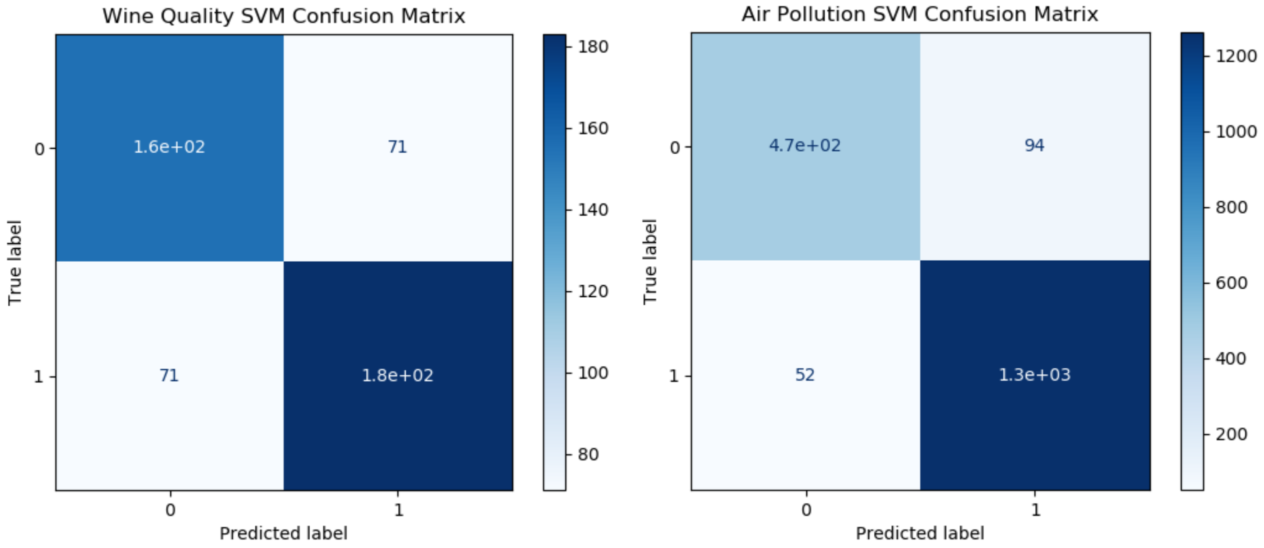

Figure 10. Confusion matrix of SVM for two datasets

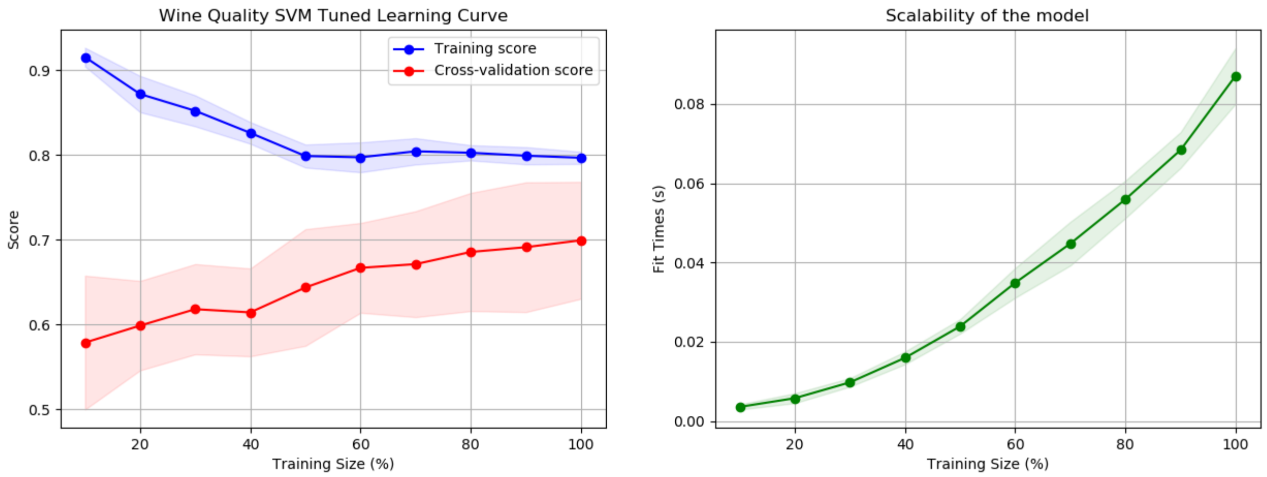

Figure 11. Learning curve and model scalability of wine quality dataset using SVM algorithm. Dotted lines are means and colored areas represent the corresponding standard deviation.

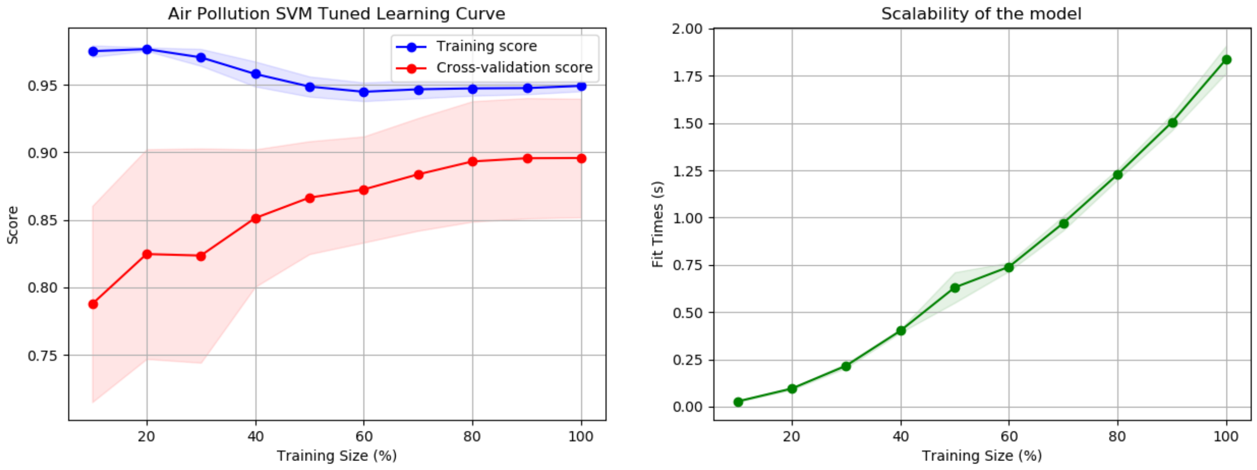

Figure 12. Learning curve and model scalability of air pollution dataset using SVM algorithm. Dotted lines are means and colored areas represent the corresponding standard deviation.

4.5 K-Nearest Neighbors (KNN)

K Nearest Neighbors (KNN) is a simple classifier based on the assumption that samples in the same class have similar features. The key parameter to tune is the value of k, namely how many neighbors are to be chosen for the class of a new sample. K value too small can cause overfitting, while the model complexity decreases as K value increase. By testing k value between 1 to 50, the k value after tuning for wine quality and air pollution data is 36 and 10 respectively.

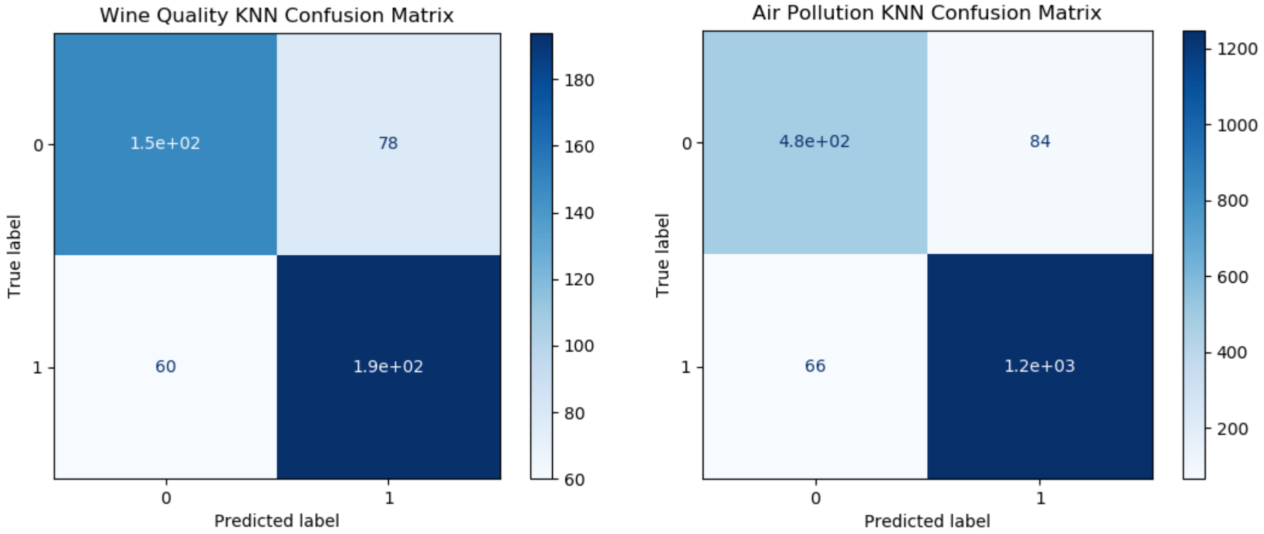

Figure 13. Confusion matrix of KNN for two datasets

The confusion matrices for two datasets are similar with high diagonal values suggesting good prediction. However, from the learning curve (Figure 14 & 16), the difference between training score and cross-validation is large and two curves do not converge. This indicates that KNN for both datasets suffer from high variance. In future, higher K values can be tested to see if performance can be improved.

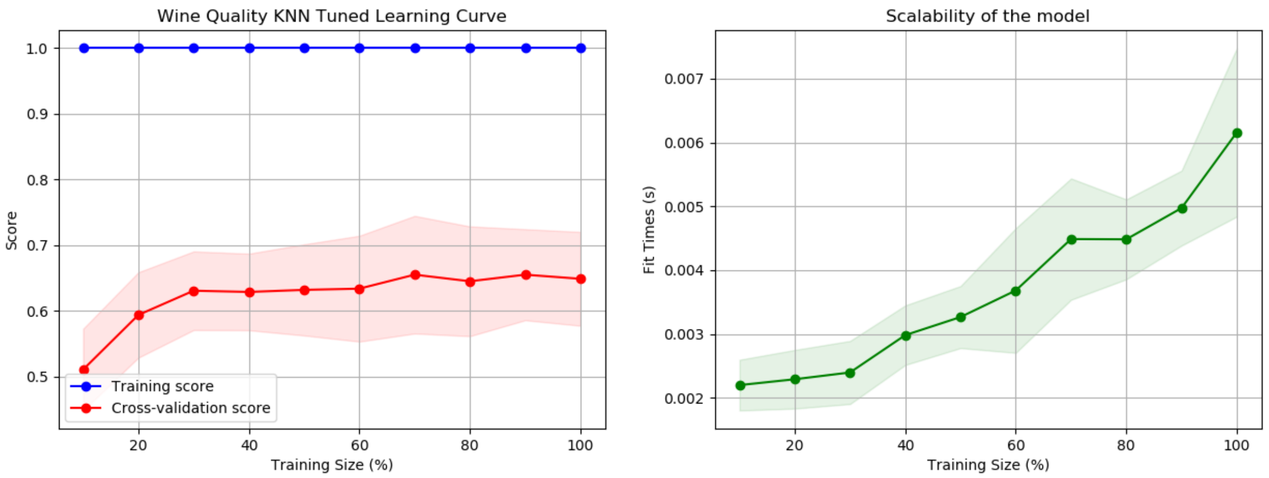

Figure 14. Learning curve and model scalability of wine quality dataset using using KNN algorithm. Dotted lines are means and colored areas represent the corresponding standard deviation.

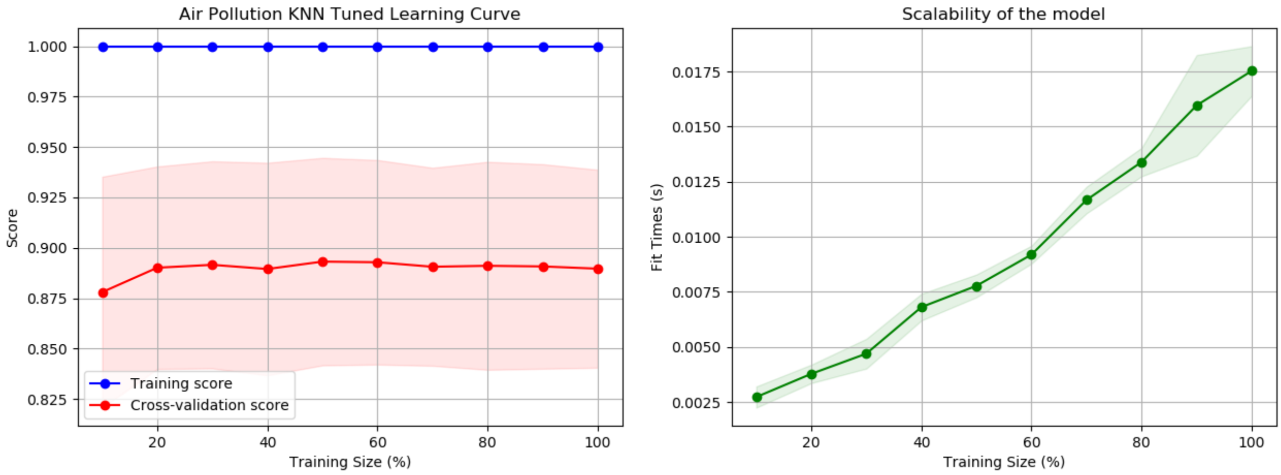

Figure 15. Learning curve and model scalability of air pollution dataset using KNN algorithm. Dotted lines are means and colored areas represent the corresponding standard deviation.

5. Algorithm Comparisons

The model prediction accuracy and runtime for above five algorithms are summarized in Figure 16 to 18. The accuracy score between true test y data and predicted y data is used to present the performance of each algorithm. According to Figure 16, all five classifiers perform similarly on wine quality data with Decision Tree and AdaBoost models slightly higher accuracy score. As for air pollution data, five classifiers give similar accuracy and the accuracy score (around 0.85) is generally higher than that of wine quality data (around 0.65).

Regarding the time to train the data set and time to predict data, the neural network model takes longest than other models when training two datasets, which is expected as neural network is much more complex. As for the inference time on test data, AdaBoost and SVM are much longer when predicting wine quality data, while SVM takes more time to predict air pollution data. Decision tree is the fastest model and a good accuracy score for both datasets, hence probably the most efficient methods in this case.

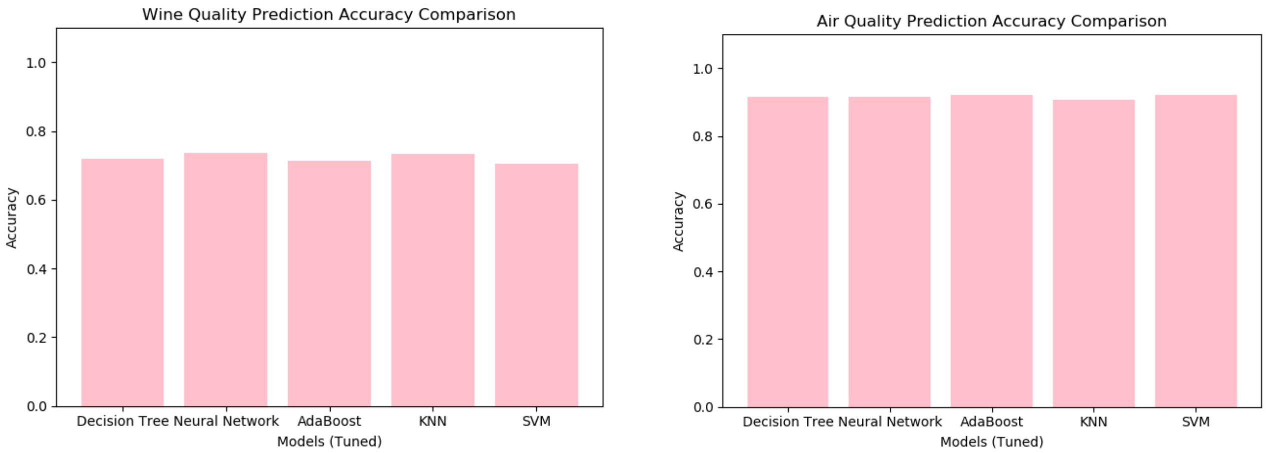

Figure 16. Accuracy score comparison among five models (tuned) for wine quality and air pollution data.

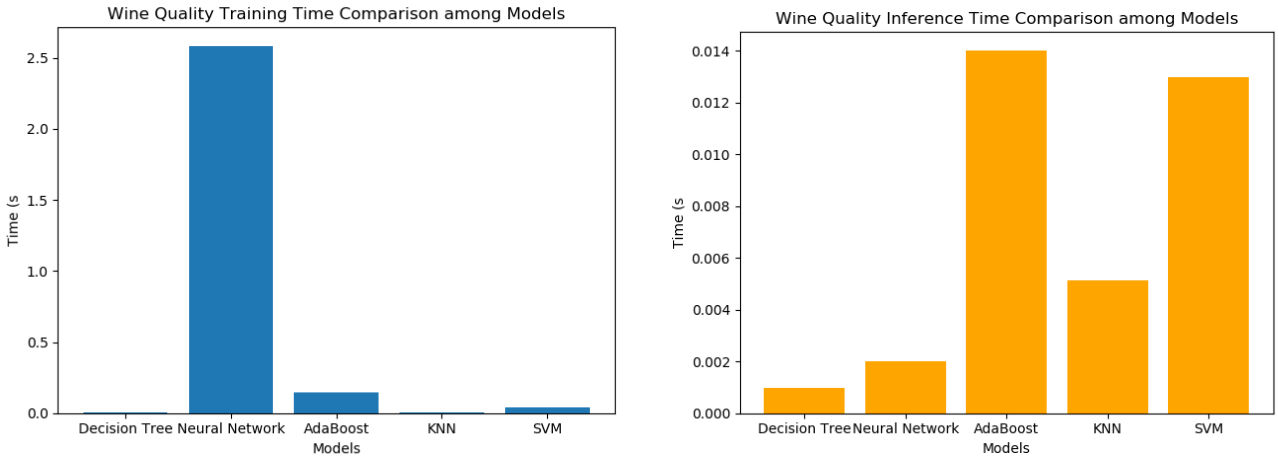

Figure 17. Training and inference time comparison among five models for wine quality data.

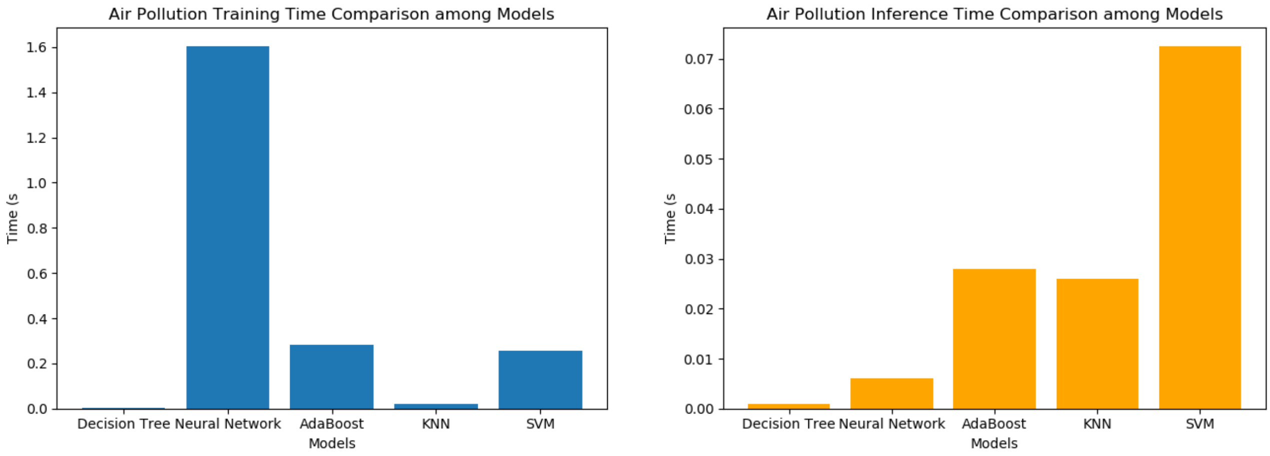

Figure 18. Training and inference time comparison among five models for air pollution data.

References:

Wine Quality Data: https://archive.ics.uci.edu/ml/datasets/Wine+Quality
Air Quality Data: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data
scikit-learn (0.22.1) User Guide: https://scikit-learn.org/stable/supervised_learning.html