CS-7641: Supervised Learning

# Intro:

The purpose of this project is to analyze five supervised learning algorithms. The learning algorithms that were analyzed are:

1. Decision tree with pruning
2. Neural Networks
3. Boosting
4. Support Vector Machines
5. k-nearest neighbors

The code for the project was written in Python on Jupyter Notebook

The libraries used for the project:

* Numpy – for numbers
* Pandas – data processing
* Scikit-learn – learners/data processing
* Matplotlib – plotting tool

# Dataset Description:

Both datasets were downloaded from UCI Machine Learning Repository. Datasets are classification problems with multivariate data types.

Dataset 0: Wine Quality dataset contains 4898 rows and 12 columns, the data contains physicochemical and sensory variables, no data such as wine brand or selling price is available. The job for the learners is to identify and predict good or bad wine given the variables. There are 3838 bad quality wine and 1055 good quality wine. What makes this dataset interesting is that there are wines that are neither good or bad but in the middle, also not all given variables are relevant to predicting the wine quality, which poses a challenge for the learning to identify which X values are relevant.

Dataset 1: Breast Cancer dataset contains 699 rows and 11 columns, the data contains different parameters of a tumor and in the “class” columns the indication whether or not the tumor is benign or malignant. Learners must predict if the tumor is benign or malignant given the parameters. There are 444 benign and 239 malignant tumors. What makes this dataset interesting is the fact that it is a clear yes or a no, it is a simple classification problem but not a lot of data is given to train on, also unlike the wine quality dataset, breast cancer input variables are all relevant with some indicating a definite malignant tumor.

# Learners Analysis

First the certain parameters plot between datasets will be compared before any tuning is done. Next, the tuned accuracy and confusion matrix will be analyzed. Finally, the learning curve for each dataset will be analyzed too look for thins such as bias or variance.

In the end all algorithms will be analyzed at once.