Name: Sameer Patil

Student ID#: 862060515

UCR Net ID: spati007

Week 4 Assignment Solution:

1. Understanding and explaining the data set.

The dataset consists of 12 columns in total, which helps in prediction on wine quality. Below is the names and description of each variable:

1. **fixed acidity:** it is the feature whose mean is 8.31 and standard deviation is 1.74
2. **volatile acidity:** it is the feature whose mean is 0.52 and standard deviation is 0.17
3. **citric acid:** it is the feature whose mean is 0.27 and standard deviation is 0.19
4. **residual sugar:** it is the feature whose mean is 2.53 and standard deviation is 1.40
5. **chlorides:** it is the feature whose mean is 0.08 and standard deviation is 0.04
6. **free sulfur dioxide:** it is the feature whose mean is 15.87 and standard deviation is 10.46
7. **total sulfur dioxide:** it is the feature whose mean is 46.46 and standard deviation is 32.89
8. **density:** it is the feature whose mean is 0.99 and standard deviation is 0.001
9. **pH:** it is the feature whose mean is 3.31 and standard deviation is 0.15
10. **sulphates:** it is the feature whose mean is 0.65 and standard deviation is 0.16
11. **alcohol:** it is the feature whose mean is 10.42 and standard deviation is 1.06

Other information related to each column is given in the notebook.

Output variable (based on sensory data):

12. **quality** (score between 0 and 10)

2. Processing data, cleaning up.

For cleaning of data, I have used standardscaler which is pre-built in scikitlearn library, as this is the standardized formula used by most of the feature normalization methods.

This problem can be treated as classification as well as regression problem, but we are treating it as a classification problem as quality ranges between 0-10 and in the dataset which I have at hand has qualities (3,4,5,6,7,8).

As few numbers of examples, I have converted this problem into binary classification problem (good, bad). I set limit of 6.5, means 2 to 6.5 is transformed into bad quality and 6.5 to 8 is transformed into good quality. For more details please find the notebook.

3. Dividing your data into a training and test set.

In order to perform validation on the data – I have divided data sets between training data and testing data. The split is 80% and 20%

4. Choosing the relevant algorithm.

I have tried different machine learning models (RandomForest, SVM, SGD), for more details regarding accuracy and other measures please find the attached notebook.

5. Evaluating your learning performance.

Radom Forest giving better accuracy post accuracy improvement measures.

For increasing accuracy, I used GridSearchCV (grid search cross validation). The best accuracy is 91%, please find the attached notebook.