**Problem Addressed**

Predicting the quality of red wine using the features of the wine identified through tests.

**Description of the dataset**

Link to the dataset: <http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/>

The dataset was taken from UCI machine learning repository. It consists of 1599 red wine instances. There are 11 attributes and 1 output attribute.

The input variables are as follows:

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8. Density
9. pH
10. Sulphates
11. Alcohol

**Data Preprocessing**

Since the red wine quality dataset doesn’t contain missing attribute values any preprocessing, in order to remove them was not carried out.

The dataset was tested for null values and correct data types. Feature scaling was carried out in order to normalize the data. Here the features are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol.

**Classification using Decision Trees**

Decision tree is a predictive model used in machine learning, statistics and data mining. A Decision tree consists of a tree structure (Directed, acyclic graph) starting with some observations about an item (Data) to conclusions (target value).

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

The red wine quality dataset can be used for both classification and regression. I have used the dataset for classification to complete the assignment. With respect to the wine quality dataset the target variable is quality and the data features are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol. Therefore, decision tree approach has been used in order to predict the quality of red wine from the data features mentioned above.

I have used DecisionTreeClassifier in order to predict the quality of red wine using the provided data features. DecisionTreeClassifier is a class capable of performing multi-class classification on a dataset. Since there are 11 data features which affect the quality of the wine, those can be considered as the multiple classes that are being considered in the classification that I have carried out.

**Results**

The dataset was tested for null values and checked for correct datatypes as a part of preprocessing.

In order to visualize the distribution of quality ranks of red wine a bar plot is plotted.

According to the below figure it can be inferred that the quality rating of the red wine follows a fairly normal distribution.

The accuracy of the classification is identified using confusion matrix

In order to understand further connections of the dataset a heatmap consisting of the correlation coefficients between features are plotted.

When we look at the heatmap it can be identified that there are strong positive correlations between fixed acidity and density and total Sulphur dioxide and free Sulphur dioxide. And also, there are strong negative correlation between fixed density and pH, citric acid and pH.

sklearn-Efiicient tools for moodeling including classification,regression,ML,statistics & reduction

A confusion matrix is a summary of prediction results on a classification problem.

The confusion matrix shows the ways in which your classification model is confused when it makes predictions.  
It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.