**Wine Quality Prediction.**

**Name: Vinod Kumar Nethi Chinna**

**Student\_ID: 300304630**

**Introduction**

This project is about predicting the quality of wine. In this project we have 11 feature columns or attributes and 5 different categories for target column (i.e., Quality). The target class have values ranging from 4 to 8. 4 being the least quality of wine and 8 indicates good quality of wine. Quality is very important in service oriented businesses like restaurants, lounges and key for customer satisfaction.

**Uses of Project.**

* This model can determine the quality of wine based on various attributes.
* This will help restaurants and lounges to decide best quality wine among various options available in market.
* Maintaining high quality wine collection will sure attract customers and business can be developed for restaurants and hotels.
* This will help individual who drinks wine a lot and looking for better wines to choose high quality wine.
* Wine Manufacturers can decide what effects their wine quality and necessary steps can be taken to ensure the quality and achieve customer satisfaction.
* Price can be determined based on quality by wine manufacturer and sellers.

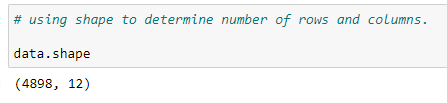
Null Hypothesis – This states that all the features will not have any effect on quality of wine.

Alternate Hypothesis – when null hypothesis is rejected then we can say alternate hypothesis is supported and attributes have some effect on determining quality of wine.

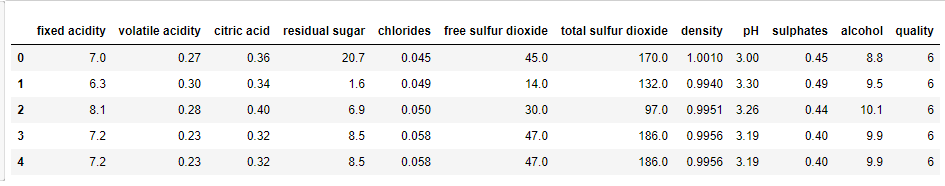
**Data Preparation.**

Dataset is downloaded from the link <https://www.kaggle.com/maitree/wine-quality-selection>. This data set contains 11 features of white wine namely (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality.

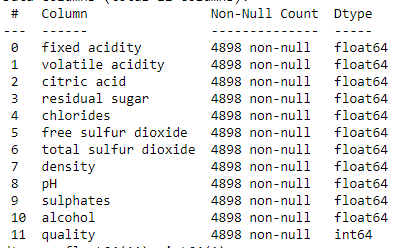
Quality is the target column and rest all attributes can be treated as features. Based on the values of feature columns we will build Machine Learning classification model to predict the quality of wine. Data contains 4898 records and 12 columns.



Below are the first 5 records of data.

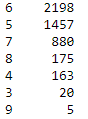


Below screenshot shows that dataset do not have any missing values.



The columns names are renamed to maintain standard and Data is not evenly distributed among target classes, this may affect model performance. I have merged class labels with low records with subsequent class.

Initial count. Modified Count.

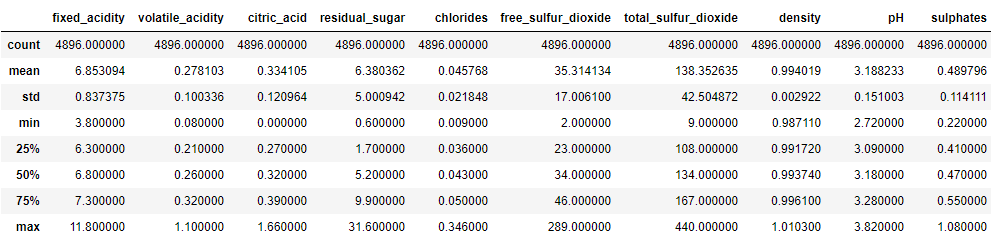
Below is the code for data manipulation. Labels with 3 are modified as 4 and same with 9 class label is replaced with 8.



Outliers are removed by observing statistics of data. We only have one record for outlier conditions, this may affect our model, so we are dropping the outliers.



Here are the descriptive statistics of our data after dropping outliers.



**Model Planning & Implementation.**

For this project I am using XGB Classifier as my final model and used to predict the test data and out of sample data. XGBoost is an implementation of Gradient Boosted decision trees used for high speed and better performance of model. In this model we will make sure of resources very well. Weak learners are trained on Strong learners’ errors and this make sure we give weight to data which are not correctly classified by strong learners. Models are created in sequential manner and feedback is taken from previous version of model to decrease the amount of data which are misclassified.

This is purely classification problem, so I did not get chance to work with multiple models in prediction. Tried various classification techniques and model with good accuracy is selected. Random Forest with more estimators is giving good accuracy but it is consuming most time and resources. When increasing estimators, I also faced problem of overfitting the data.

Now it time to tune our XGB model with hyperparameters. To achieve this RandomSearchCV is used to find the best parameters for our model. Once this is done, we now have the best parameters to use in our classifier and this can boost the accuracy.

In creating this model, I used pipeline to chain the actions of scaling and model fitting. Before hand I have performed feature selection using variance threshold feature selection method.

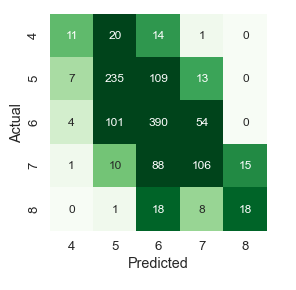
However, scaling the features do not give us best results, this completely depends on type of problem. Scaling will significantly help us in regression problems but in classification decision-based models we can observe scaling will not have significant change. Tried both the ways and opted the method which has more accuracy.

Initially I did not remove outliers in data and continued to model and later point I removed outliers. Surprisingly, accuracy with outliers in data is higher. Accuracy is dependent on various aspects of data, type of data, noise in data, hyperparameters of model, selecting important features, weight to features.

In this project I tried my best to increase accuracy by following the standards of machine learning. Data with outliers gave me higher accuracy but I removed to ensure data is clean and free from outliers. Due to lack of sufficient data model is not performing well, but if we can add more data then our model can improve in accuracy aspects.

**Results Interpretation and Implications.**

Below is the confusion matrix.



Above figure can be understand by below calculations.

We have total 11+20+14+1 = 46 records with label 4 and our model predicted 20 records correctly into class 4.

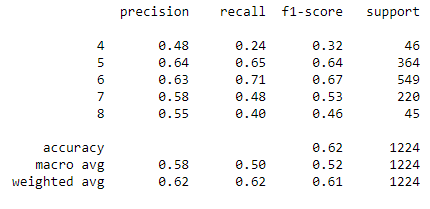
We have total 7+235+109+13 = 364 records with label 5 and our model predicted 235 records correctly into class 5.

We have total 4+101+390+54 = 549 records with label 6 and our model predicted 390 records correctly into class 6.

We have total 1+10+88+106+15= 220 records with label 7 and our model predicted 106 records correctly into class 7.

We have total 1+18+8+18= 45 records with label 8 and our model predicted 18 records correctly into class 8.

Here is the classification report.



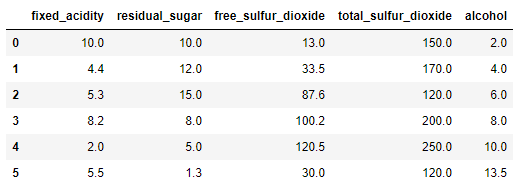
Accuracy of model is 62 %.

This model can be improved if more training data is provided. More training data increases the model performance.

Random forest also gives satisfying results but using higher number of estimators consumes resources and computing power. Powerful machines are needed to handle.

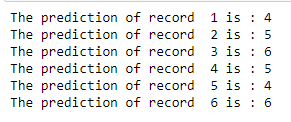
**Out of Sample Predictions.**

Here is the snapshot of out of sample data.



This is created by mixing and matching of unique values of each attributes from data.

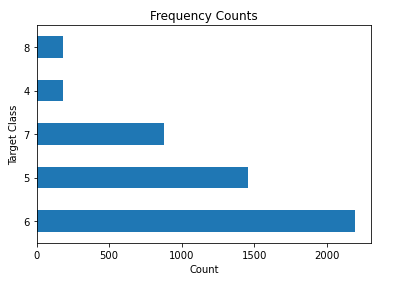
Below are the predictions of data made by XGB Classifier model.



Our model classified quality of wine with record1 as attributed into class 4 and results of other records are displayed in above figure.

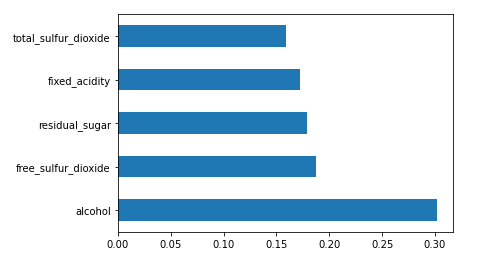
**Conclusion.**

This graph shows how the data is distributed among different classes.



Majority of data falls under label 6 and very less data for classes of 4 & 8. This shows data is biased and not evenly distributed. Model cannot train on such less data and predict well for every scenario. More data which is evenly distributed among class labels will improve model performance.

Below chart shows the feature importance.



Based on important features graph it is clear this attributes effects quality of wine. We can reject the null hypothesis we made in starting of this project. We have support to alternative hypothesis.

Among 5 features which are selected using variance threshold selection method, alcohol has more weight than other features. I observe alcohol % of value in wine will plays major role in predicting the quality of wine. Free sulfur dioxide takes next place followed by residual sugar then fixed acidity and total sulfur dioxide being least weighted attribute, among others.

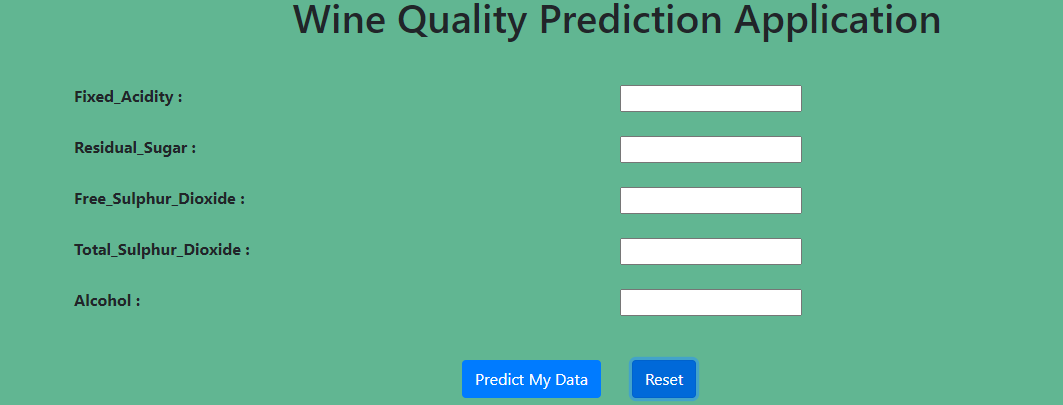
Percentage of alcohol in wine will majorly affect its quality, based on correlation report also we can see it has more weight and we can consider this as important aspect. This will help manufacturers to produce good quality wine.

We observed total sulfur dioxide, fixed acidity, residual sugar inversely affects quality of wine. Less the concentration of above 3 will give us good quality of wine. Manufacturers need to make necessary steps to reduce this proportions while producing wine.

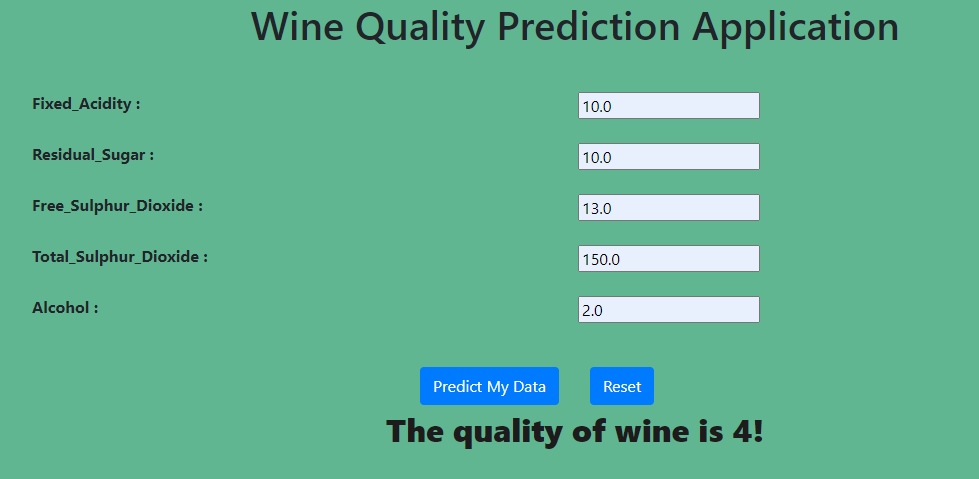
This will give user a clear idea to decide weather wine is good quality or not by looking at this features or attributes.

**Flask Application.**

I have developed a flask application which provides GUI interface for user to predict the quality of wine using above 5 features.



This takes 5 inputs from user and predict the quality of wine. Below is the test run of our flask model with 1st record of out of sample prediction.



We got the same result in both predictions using flask and Jupyter notebook.

**Instructions to run.**

1. Go to Flask Application folder.
2. Open command prompt
3. Install the required libraries by running the below command.

**pip install -r requirements.txt**

1. Now run the app.py

**Python app.py**

This will give us a link and if we open we can access the above webpage and predict the quality of wine.

\*\* End of Document \*\*