**Red Wine Quality Prediction using physicochemical and sensory variables**

**Using python, Dataset from GitHub, pandas, NumPy, Importing various dependencies for Exploratory Data Analysis, Model Building**

The dataset is related to red and white variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g., there is no data about grape types, wine brand, wine selling price, etc.).  
  
This dataset can be viewed as classification task. The classes are ordered and not balanced (e.g., there are many more normal wines than excellent or poor ones). Also, we are not sure if all input variables are relevant. So, it could be interesting to test feature selection methods.

GitHub is a popular sharing website for sharing open-source projects and code repositories. For example, the TensorFlow repository contains the entire source code of the TensorFlow deep learning frame work.

Repositories in GitHub can be tagged using topics. For example, the tensor flow repositories have the topics python, machine learning, deep learning etc

Repositories can used to share the data set required for machine learning

The page <https://github.com/dsrscientist/DSData/blob/master/winequality-red.csv> provides a dataset on GitHub used for prediction of red wine quality.

In this article we will perform various steps for the prediction of red wine quality using physicochemical as Features and sensory as Label variable. We will use python libraries pandas, and Other various third-party dependencies for the prediction.

The followings are outlines of this project:

1. Download dataset from the GitHub
2. Data Analysis
3. Data Pre-processing
4. Train and Test split dataset
5. Model Building
6. Trained Model
7. Saving Best Model
8. Test Model & Quality Prediction

**Download dataset from the GitHub Account**

Starting: <https://github.com/dsrscientist/DSData/blob/master/winequality-red.csv> use the link to download the data set.

Middle: Click on the button of raw and go to the page containing data. Copy the whole data and paste to the note editor and then save it with extension of ‘.csv’ (Comma Separated Values).

End: Download the dataset from the account, convert it into csv file and then upload it into the Python Jupyter Notebook.

Now kick start the project.

**Data Analysis**

**Starting**: First step is to import all the dependencies libraries for the project.

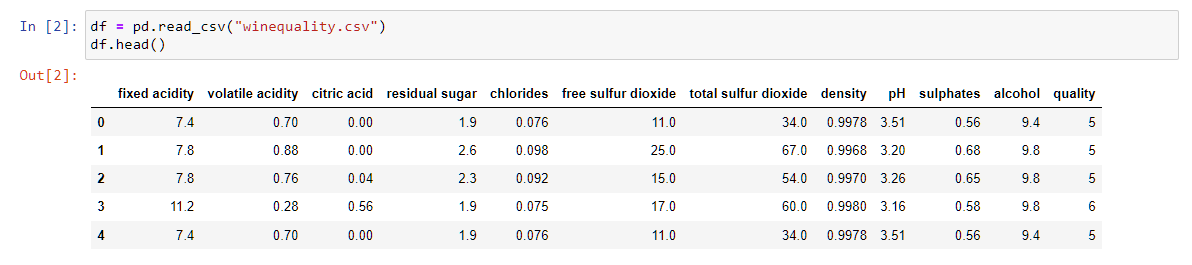
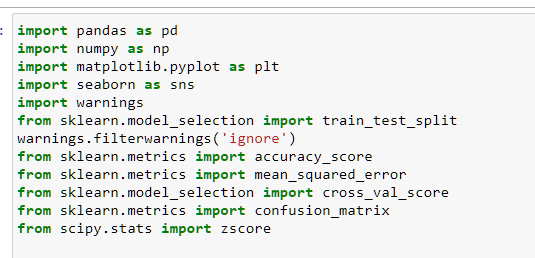
Second step is to load the dataset into data frame to convert it into tabular format.

Third step is to get information regarding data using various operation on it.

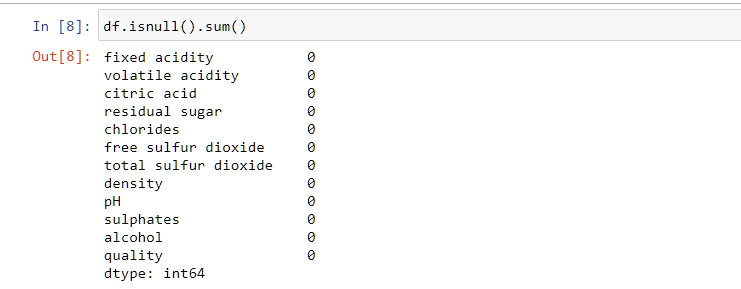
Fourth step is to get the insights of the data using the Statistics and Data Visualization.

Check the correlation between Features and label columns. If there is multi collinearity then accordingly we will have to use the Variance Inflation Factor to balance the collinearity dropping highest collinear column and check again

**Middle:** After importing the required dependencies in the jupyternote book, the actual task begins. The next step is to load the data set using panda's library which lead to dataset representation into tabular format, which is easy to analysis.



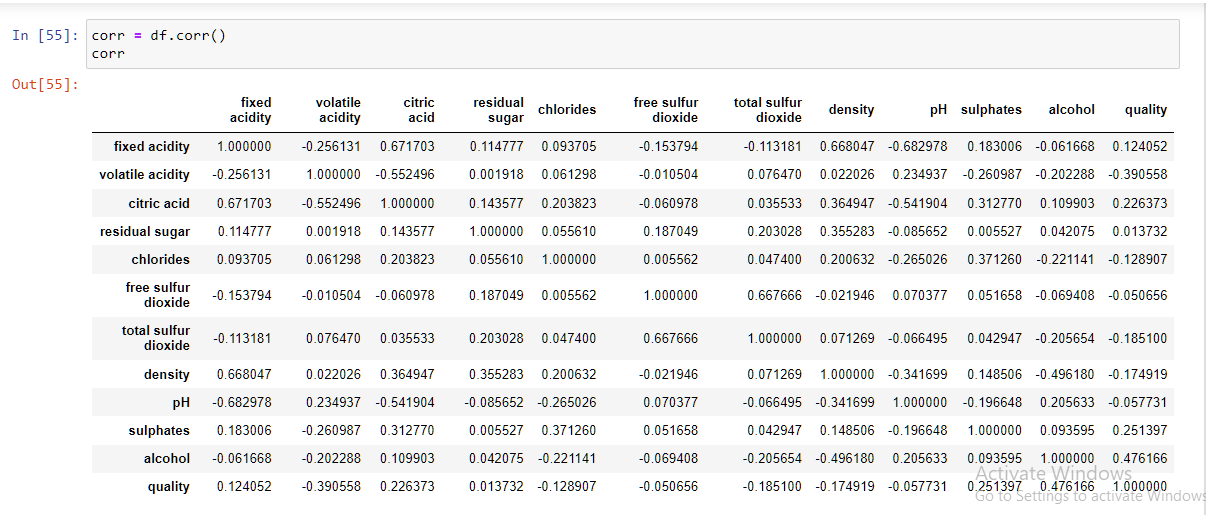
Next step finds out the statistic of the data and ask set of questions to get insight of the data. Find out if the data is having null values and then take effective measure to convert dataset into meaningful input for the machine learning.



Next step is Data Visualization of the data to get more insights through various types of plots and graphs using various dependencies libraries. Write the important observation to pay attention on.



Next step is to check correlation of feature columns with respect to label/target column using heatmap from seaborn library which gives the idea about the correlation. If there is multi- collinearity then by using VIF (Variance Inflation Factor) we can overcome the problem of multi-collinearity



**End:** Importing dependencies and loading data set using panda’s data frame and then check the information of the dataset. Start taking insight using statistics and visualization and find out the correlation between features and label column.

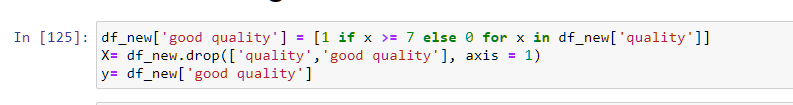
**Data Pre-processing**

**Start:** Data Pre-processing is divided into four sub parts they are has follow:

1. Identifying and handling missing values
2. Encoding the Categorical data
3. Splitting the dataset
4. Feature Scaling

**Middle:** To identify and handle the missing values present in the dataset. Mostly it consists of nan values. In some case it has unique characters instead of the nan values. In such case if we didn’t handle the condition, the model learning input will throw some errors.

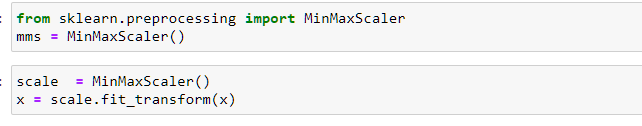
Encoding the Categorical data, Label Encoding library is used to convert the string datatype into numerical datatype so that it can be given as an input to the model.



Splitting the dataset is performed on dataset to divide data into two parts such as ‘x’ variable will be saved for feature columns and ‘y’ variable will be saved for label or the target column to perform training and testing of the model and to predict the test data after applying necessary model algorithm.



Feature Scaling is a method to standardize the independent variables of a dataset within a specific range. In other words, feature scaling limits the range of variables so that you can compare them on common grounds.



**End:** To identify and handle the missing values which is in the form of nan or to replace some unknown unique which are unable to identify we visualize it with fillna with mean method. If the column data is of numeric type and if it is categorical data, use mode to fill the nan values.

Encoding the categorical data is used to convert string type data into numeric type.

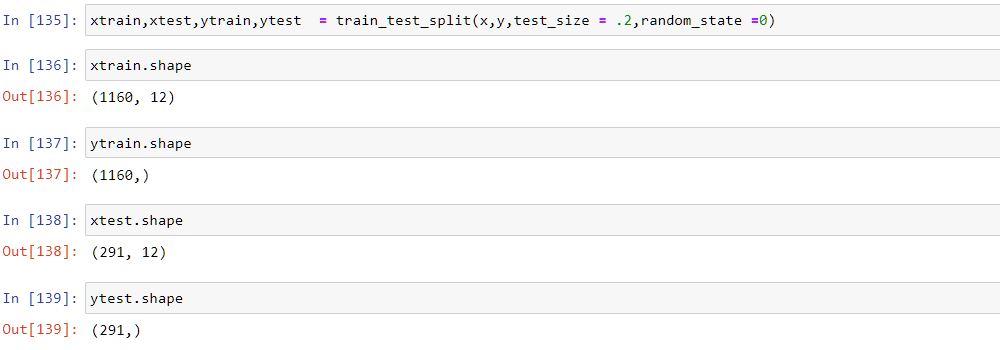
Splitting the data into two part one for training and other testing in machine learning

Feature scaling is used to standardize the independent variables of dataset within a range.

**Train and Test Split Dataset**

**Start:** The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data which are not used to train the model.

It is a fast and easy procedure to perform, the results of which allows you to compare the performance of machine learning algorithms for your predictive modelling problem.



**Middle:** The train-test split is a technique for evaluating the performance of a machine learning algorithm.

It can be used for classification or regression problems and can be used for any supervised learning algorithm.

Train Dataset: Used to fit the machine learning model.

Test Dataset: Used to evaluate the fit machine learning model.

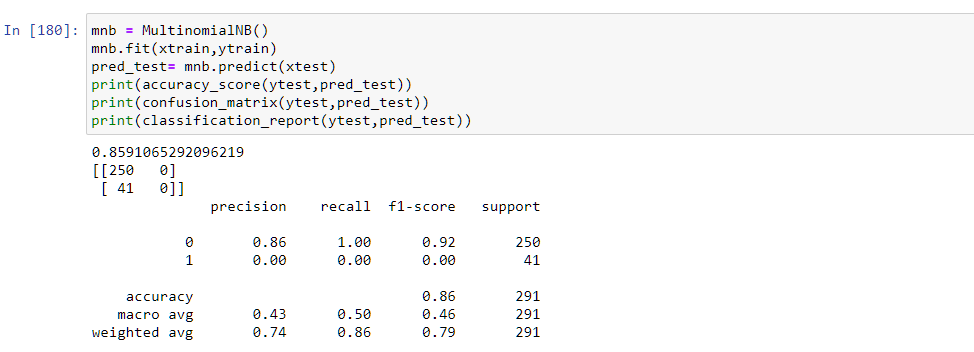
**End:** The train-test split used to estimate the performance of the machine learning algorithms. It is easy procedure to preform and allows us to compare the performance of machine learning algorithms for the predictive models. Train dataset is used to fit the machine learning model and Test data set is used to evaluate the fit machine learning model.

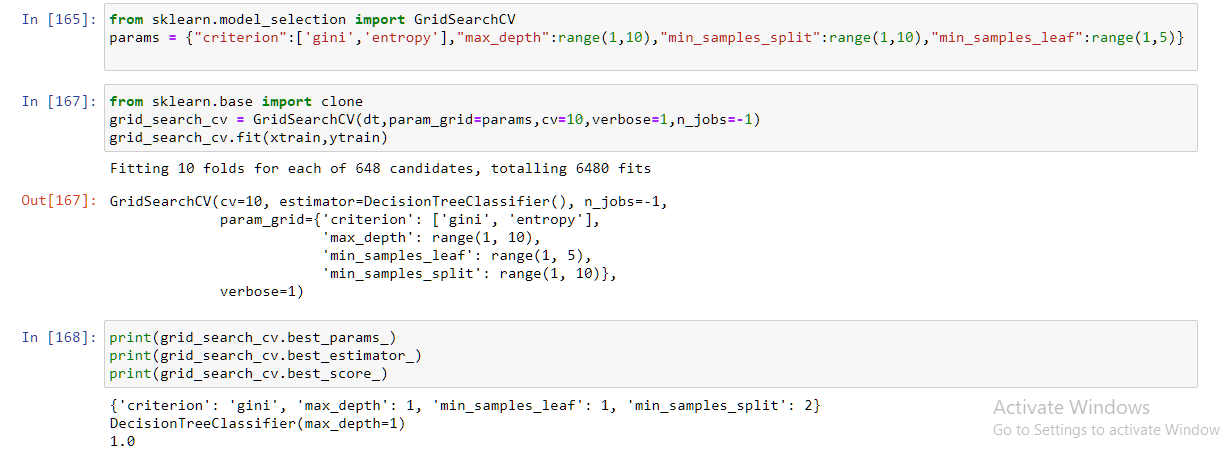
**Model Building**

**Start:** The evaluation of the machine models using a model metric approach, quality measurements, datasets, and matrix calculations. This phase is the quality assurance of a machine learning approach by using algorithms and Ensembling techniques and Hyper-Parameter-Tunning for getting the best score for the training models.



**Middle:** The evaluation of the machine models using various approach for the training of model. This project problem has multi label classification which I have converted it in to binary one by taking 7 as a good quality wine rest as not so good quality wine. The evaluation of the machine learning models using metric approach, Confusion matrix, mean squared error, accuracy score used for training the model. After training the model hyper-parameter- tunning is used to find the best training model using various algorithms and ensemble technique.





**End:** After splitting the data, z-score is used to normalize the data. Then use threshold value as 3 for normal standard deviation. Also, data is scaled using scaling method because in some columns data range is higher than the other.

Then using SMOTE, which is an over-sampling technique focused on generating synthetic tabular data. The general idea of SMOTE is the generation of synthetic data between each sample of the minority class.

Then various types of algorithms and ensemble techniques and hyper-parameter-tunning is applied for training of model, to find the best fitting model.

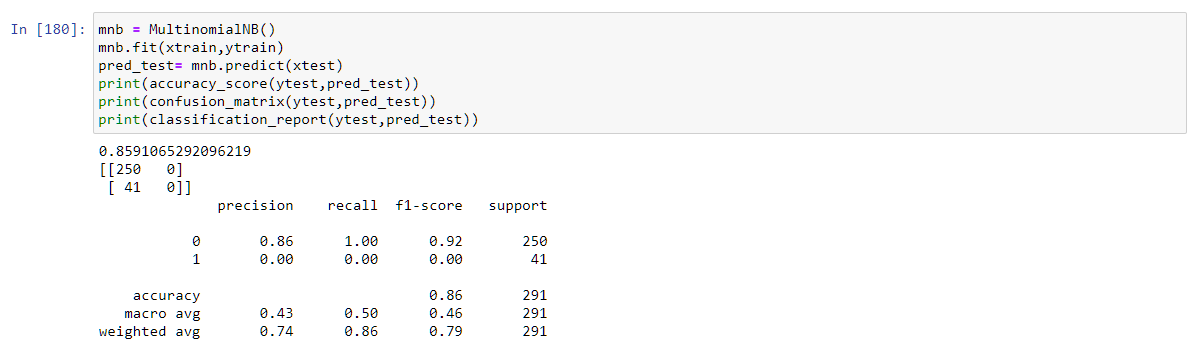
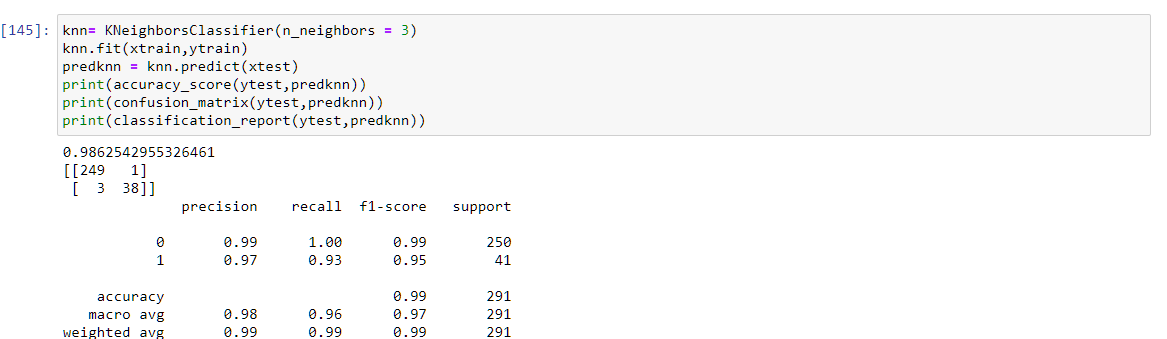
**Trained Model**

**Start:** In the machine learning model training refers to the process of allowing a machine learning algorithm to automatically learn patterns based on data.

In this step we will see the accuracy score, classification report and confusion matrix to see which of model is performing well and at which parametric tuning result is having much difference then previous training model. Then that model is saved.

**Middle:** In the machine learning model training we use various algorithms to learn the patterns based on the data set (which is pre-processed for training model) to learn good amount of data for the prediction on the test data.

In next step, algorithms, ensemble techniques are used for training the model and check which of the model is performing well. To enhance the performance more we use hyper-parameter-tuning and compare the result of each model with tunned model and save the best performing model.

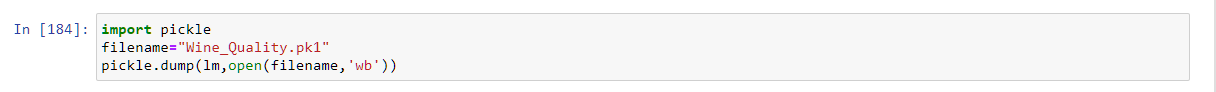


**End:** Model training is the process of allowing a machine learning algorithm to learn the pattern based on data. In this step we will see accuracy score, classification report and confusion matrix to finalize which model is preforming well and then apply hyper parameter tunning to well performed model to see if it performs better after tunning and save the best model.

**Saving Best Model**

**Start:** The best model saving is decided on the base of which model is learning very well amount of dataset pattern. Then it is hyper parameter tunned to check if that specific model training has some percentage difference. After applying the parameter tunned, it is saved. If the difference is not much then very well-trained model is saved.

**Middle:** Saving the model plays an important role in machine learning. After saving trained model, it is used for prediction of the test data-set (which are the new inputs) given to the model that predicted the pattern and see how many percentages model is learning.



**End:** The algorithm, ensemble technique and hyper parameter tunning are used for best model training. They learn the data patterns, predict the model and then new inputs are given to the model to check the prediction. The actual target columns are compared to the predicted test data set to check the how many percent model training is accurate.

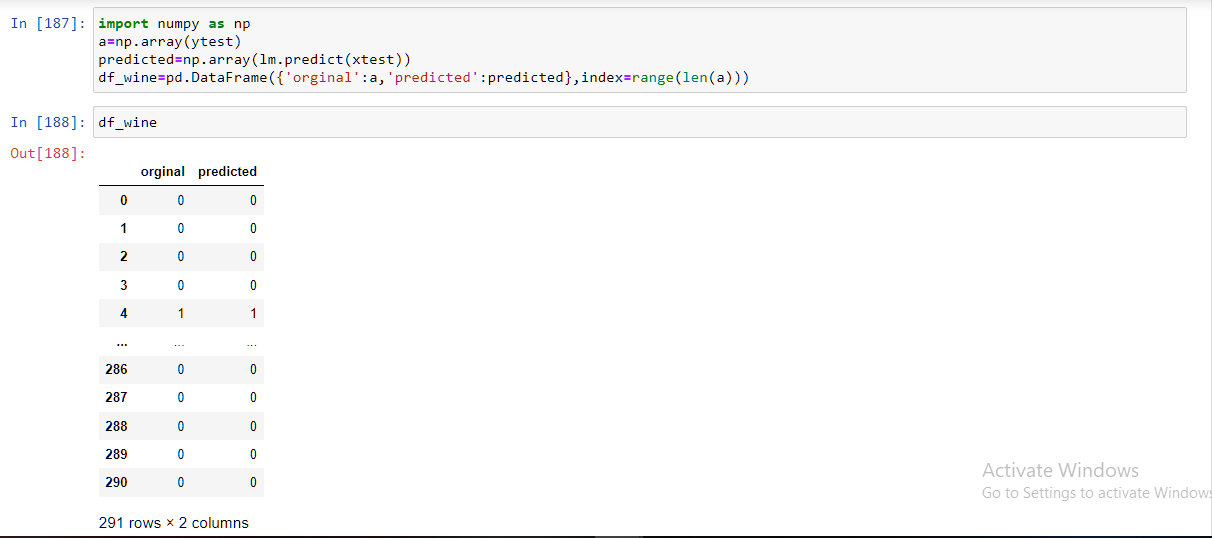
**Test Model & Conclusion**

**Start:** In machine learning, model testing is the process where the performance of a fully trained model is evaluated on a testing set. The testing set, consisting of a set of testing samples should be separated from both training and validation sets, but it should follow the same probability distribution as the training set.

The conclusion is based on actual target compared to predicted target column and to see the difference between training and testing model. Finally, it is concluded which model is performing great and then training model is dumped with help of pickle library and it is saved for future, so that there is no need to do all the task for the new test data prediction.

**Middle:** The main objective of test model is to evaluate the performance of trained model on a testing set. The testing dataset must have testing samples and should be separated from training data. Also, it should follow the same probability distribution as the training data.

The conclusion to the project is done on the basis of trained model and test model prediction. At the end to conclude the project in which algorithm, ensemble method model, with or without hyper parameter tuning preforms best are saved. In the next step both the actual and predicted result of test data is compared to see the difference between the actual and predicted values of the target column.



**End:** The main motive of the test data is to cross verify the training data learned pattern to predict the result for the project, by taking the new input for the Logistic Regression model. Observe the difference between the actual target columns and the predicted target column. To conclude that this Logistic Regression model is learning the maximum data pattern while training, visualize the difference in the testing model. According to my analysis the is performing great with 100 percent of accuracy in my case.