ML-MINOR-JULY

Machine Learning Project

Wine Quality Prediction

Project Statement:

The dataset is related to the red variant of the Portuguese "Vinho Verde" wine. For more details, consult the reference [Cortez et al., 2009]. 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.).These datasets can be viewed as regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones). Apply regression and find the quality of Wine.

Problem Definition & Target Variable:

This project aims to determine which chemical features are the best quality red wine indicators. To be more specific, we define below problems for this analysis:

- Show the contribution of each factor to the wine quality in our model
- Show which features are more important in determining the wine quality
- Show which features are less important in determining the wine quality

Our target variable will be wine quality, which is scored between 3 and 8.

Data Information:

Inputs or features:-

fixed acidity: most acids involved with wine or fixed or nonvolatile.

volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste.

citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines.

residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet.

chlorides: the amount of salt in the wine.

free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine.

total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine.

density: the density of wine is close to that of water depending on the percent alcohol and sugar content.

pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale.

sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant.

alcohol: the percent alcohol content of the wine.

Output or label:-

quality: output variable (based on sensory data, score between 3 and 8).

Tools used:

We used google colaboratory to work on this project.

Libraries used:

numpy
pandas
matplotlib.pyplot
seaborn
sqrt from math
confusion_matrix, accuracy_score, mean_squared_error, mean_absolute_error from sklearn.metrics
LinearRegression from sklearn.linear_model
train_test_split from sklearn.model_selection

Why linear regression?

We used linear regression because regression analysis allows us to understand the strength of relationships between variables**.** And regression analysis tells us what predictors in a model are statistically significant and which are not. In simpler terms, if we give a regression model 50 features, we can find out which features are good predictors for the target variable and which aren't. Regression analysis can give a confidence interval for each regression coefficient that it estimates. Not only can we estimate a single coefficient for each feature, but we can also get a range of coefficients with a level of confidence (e.g., 99% confidence) that the coefficient is in.

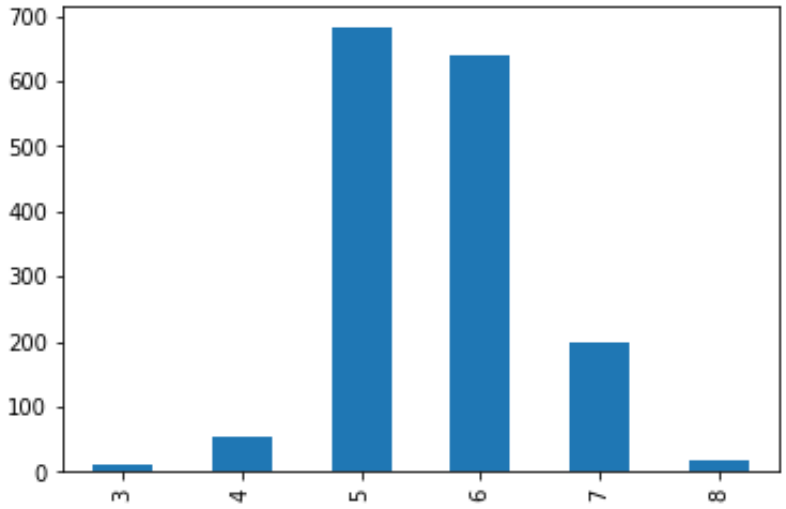
Steps involved:

1. Importing the libraries: In this step we import all the necessary libraries in order to work with the problem. Some of the libraries are Pandas, Numpy, etc.

2. Data collection: In this step we fetch the dataset into the notebook.

3. Data analysis and Visualization: In this step we use different pre-defined functions which help us to analyse and visualize the data in a more simple and

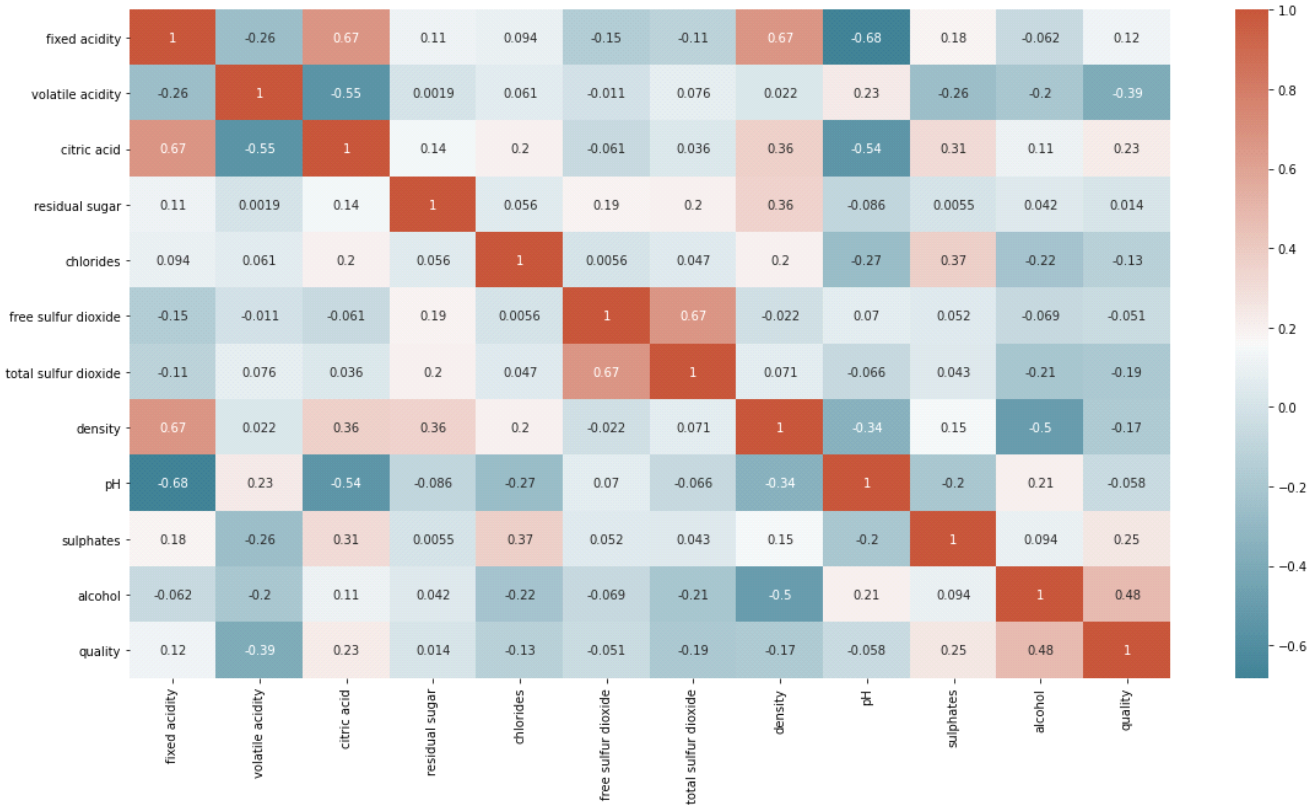
understanding manner. We used df.head(), df.info(), df.describe(). We see that there are no null values. Furthermore, the quality of wine is separated into three different groups, so that we can do things a little easier:-
LOW: contains the wines whose quality is 3 or 4.
MEDIUM: contains the wines whose quality is 5 or 6.
HIGH: contains the wines whose quality is 7 or 8.



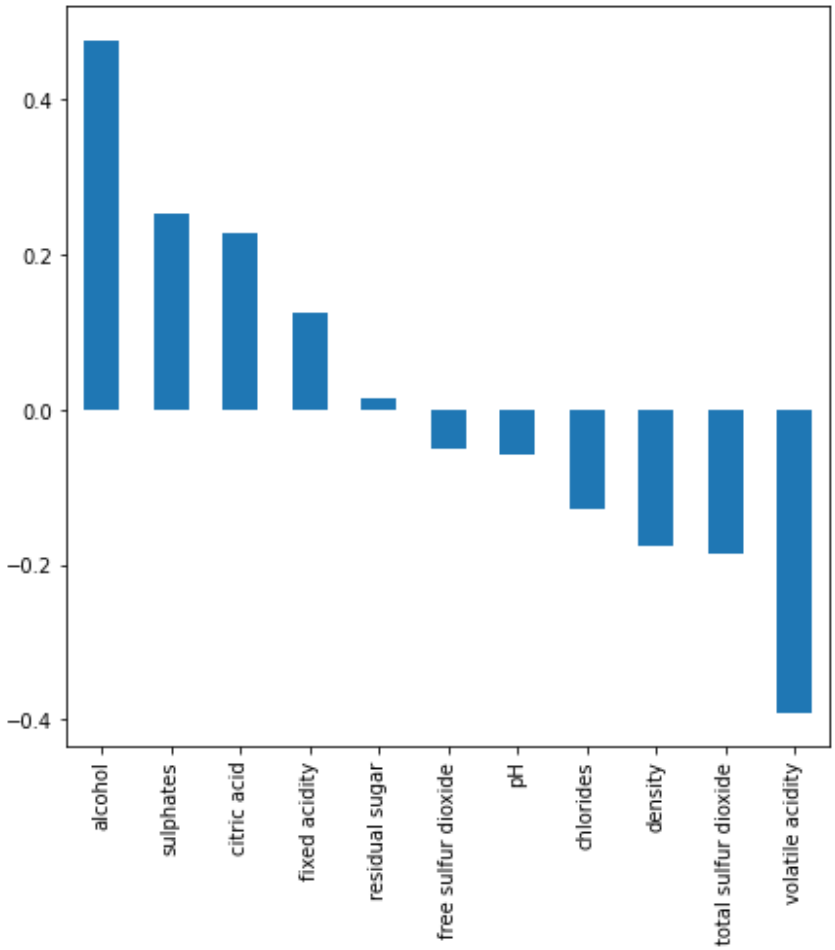
This graph shows that medium quality of wine is quantitatively more. While, low and high quality wine is scarce.

4. Features VS Quality : In this step we are plotting graphs to check how all the features are related with quality, as this gives us a better idea about the problem.

5. Correlation: Now that we have got information about the features in comparison to quality, in this step we find out the correlation between the features and quality and find out which variables play an important role in deciding the quality of the wine. Also we plot a heatmap to plot all the correlations between features and label.



For the better understanding of the features and to examine the correlation between the features we used a heatmap to display the correlation matrices.



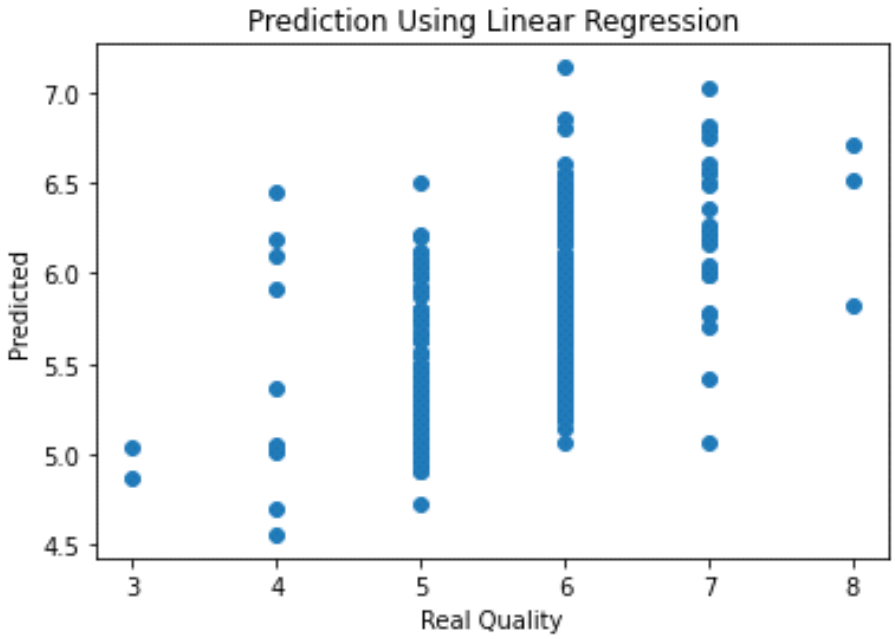
This graph shows that alcohol, sulphates, citric acid, fixed acidity and residual sugar is positively correlated with quality; which implies that they are directly priportional to quality. While, pH, chlorides, density, total sulfur dioxide and volatile acidity is negatively correlated with quality; which implies that they are inversely proportional to quality.

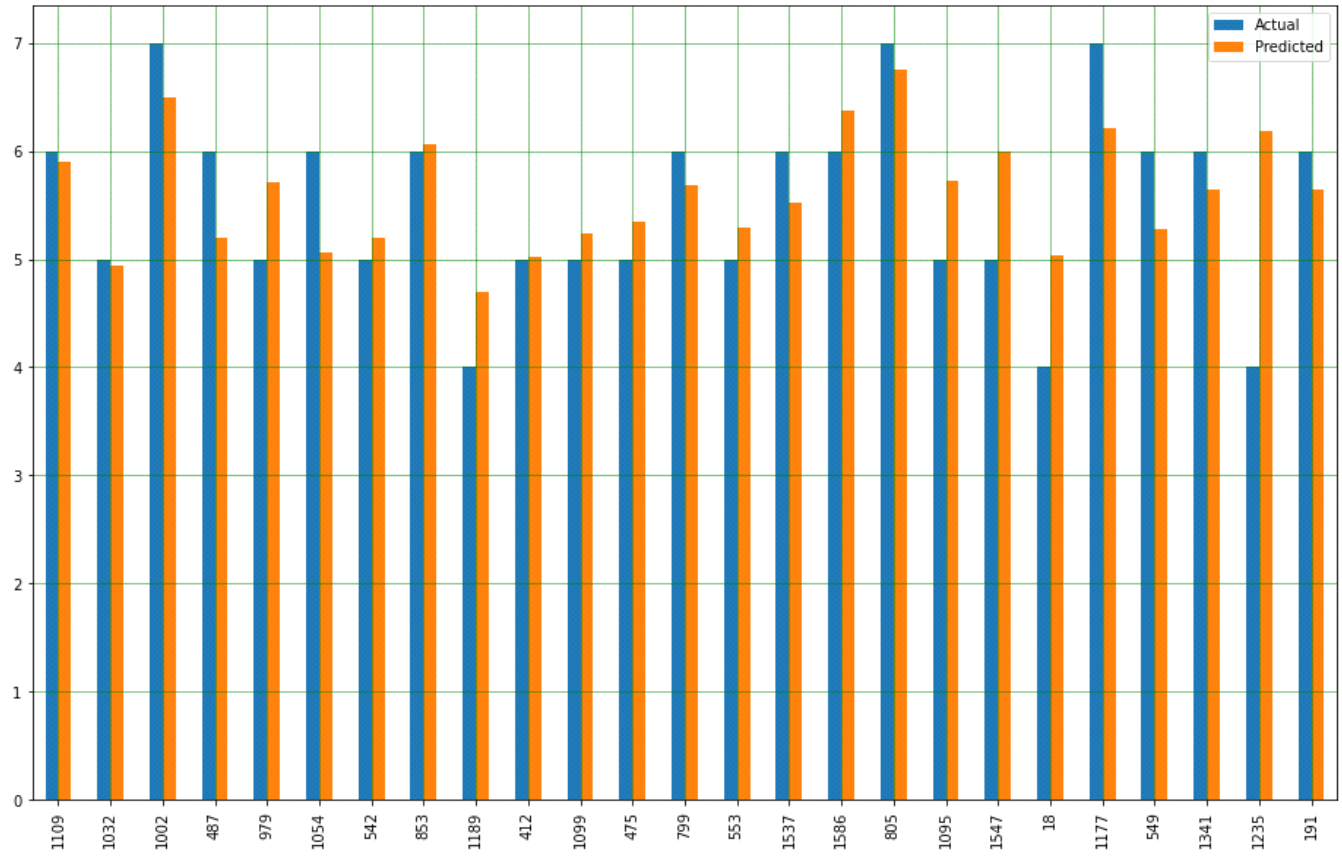
6. Data Preprocessing: In this step we select the dependent features. In the given eleven features, we have selected citric acid, volatile acidity, sulphates and alcohol. We have selected them after executing a user-defined function which displays the features that have absolute correlation with quality greater than 0.2.

7. Train and Test Split: In this step we use a procedure called Train-Test Split which is in Scikit-Learn library. So the function “train_test_split()” takes the

loaded dataset as input and returns the dataset split into two subsets which are for training data and testing data. With this function we don't have to manually divide the dataset.

8. Linear Regression: Linear regression is a basic and commonly used type of predictive analysis. It comes under supervised machine learning technique. So in this step we apply linear regression to our model to make an estimation of the quality of the wine based on the determinant features.





From the above figure we can see the actual and predicted values for quality are very close, but the predicted values are in a position where they can further be optimized.

9. Errors : We calculate the errors of the model.

Mean Absolute Error: 0.48176850659876125

Mean Squared Error: 0.40092630292327974

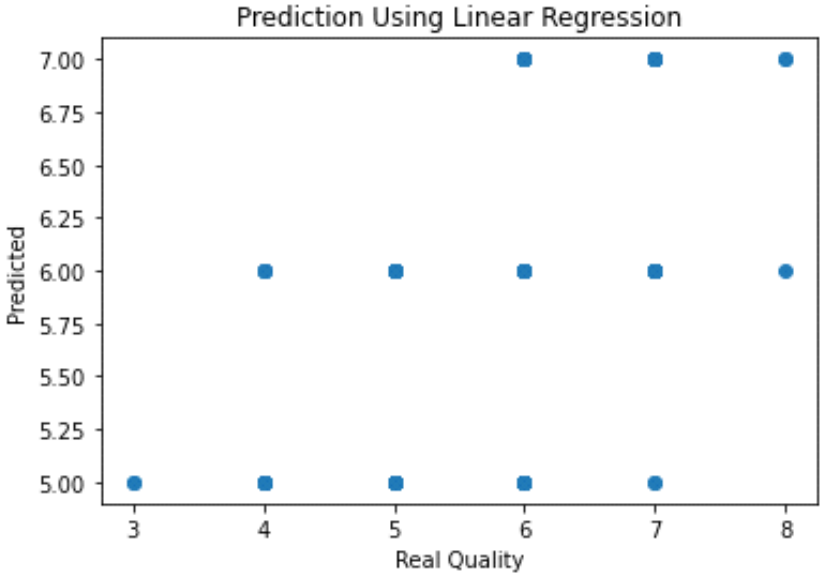
Root Mean Squared Error: 0.694095459284068

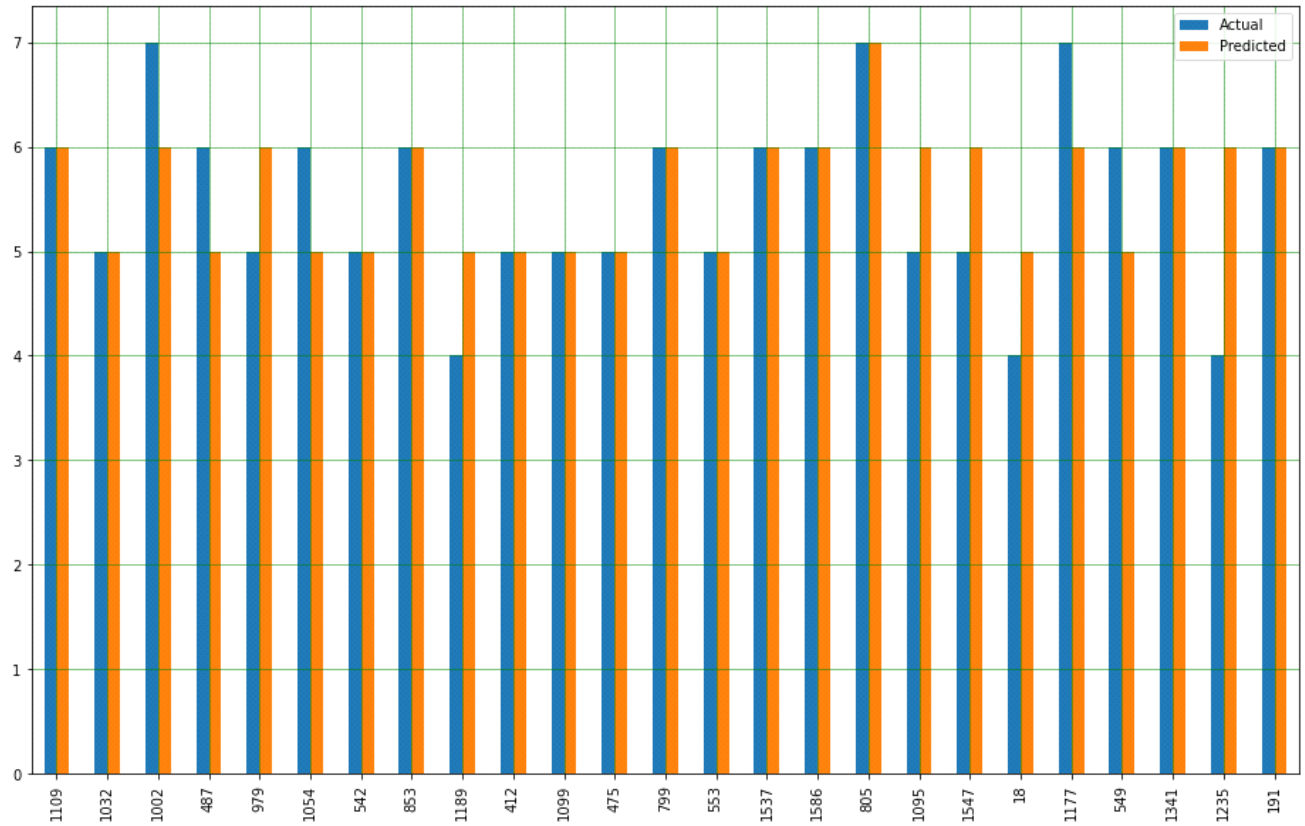
Mean Absolute Error(MAE): MAE takes the average of absolute errors for a group of predictions and observations as a measurement of the magnitude of errors for the entire group.

Mean Squared Error(MSE): To calculate the MSE, we take the difference between our model's predictions and the ground truth, square it, and average it out across the whole dataset.

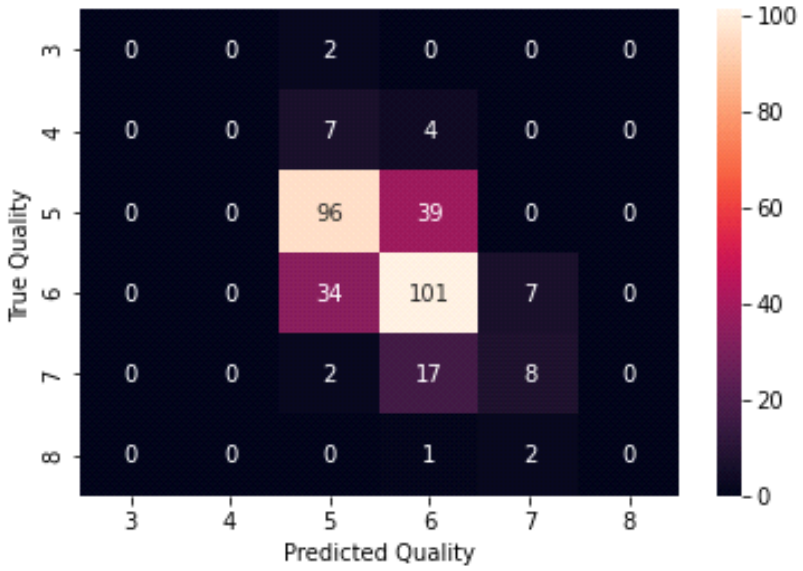
Root Mean Squared Error(RMSE): Root mean square error or root mean square deviation is used for evaluating the quality of predictions. It shows how far predictions fall from measured true values using Euclidean distance.

10. Optimizing the model: The model is optimized by rounding the variable (pred_y) to nearest integer. We reduced the error and increase the accuracy of the model.





From the above figure we can see that actual and predicted values are much closer than earlier, as we can see predicted values are much optimized than earlier.



The above confusion matrix shows the ways in which our regression model

is confused when it makes predictions.

11. New Errors: We see that the errors have decreased after optimising the model.

Mean Absolute Error: 0.3875

Mean Squared Error: 0.44375

Root Mean Squared Error: 0.6224949798994366

Accuracy score: 0.640625

Accuracy score: It is the ratio of correct prediction to the total number of predictions.

12. Output:

	Coeffecient
alcohol	0.314876
sulphates	0.671290

citric acid -0.076627

volatile acidity -1.334401

These numbers mean that holding all other features fixed, a 1 unit increase in suplhates will lead to an increase of 0.6 in Quality of wine, and similarly for alcohol.

These numbers mean that holding all other features fixed, a 1 unit increase in volatile acidity will lead to a decrease of 1.33 in Quality of wine, and similarly for citric acid.

Conclusion:

The interest has been increased in wine industry in recent years which demands growth in this industry. Therefore, companies are investing in new technologies to improve the wine(production and selling) quality. In this direction wine quality certification plays a very important role for both processes and it requires wine testing by human experts. The work done in this project will explore the usage of machine learning techniques. How linear regression determines the important features for prediction and how we used other methods and functions to predict the values. The benchmark dataset of Red wine is used, which contains 1599 rows x 12 columns of data samples. The dataset contains 12 physicochemical characteristics. This experiment shows that the dependent variable can be predicted more accurately if only important features are considered in prediction rather than considering all features. In future, Large datasets can be taken for experiments and other machine learning techniques may be explored for wine quality prediction.