MACHINE LEARNING(ML) MODEL TO PREDICT THE WINE QUALITY

This ML model is built to predict the quality of wine based on given features. The data set given contains eleven features that are key in wine quality. As the first step of Exploratory Data Analysis (EDA) the basic statistics of the data set was captured and visualized in boxplots, scatter plots and density plots. The EDA and research about the features that determine the quality of wine aided in spotting missing values and outliers which could have caused biased to our ML model. The outliers and missing values were dropped from data set.

Though there are 11 features in the data set not all of them affect the quality of the wine. To check this correlation matrix and bar graph was plotted. In figure 1 we can see that residual sugar is not affecting the quality of fine as the other features are, therefore, the residual sugar is removed from the data set.

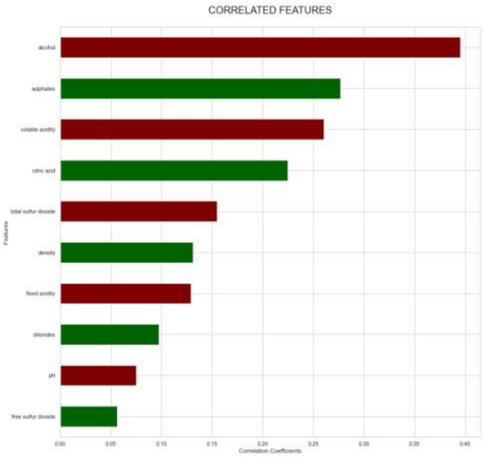


Figure 1 Correlation of features

To train the model, the data is split into two: train and test data in ratio 4:1. Since the features were all in different magnitudes and range, hence it was scaled to be uniform across all features.

I have built the following 3 models: a) k- nearest neighbor (KNN)b) Gaussian Naïve Bayes and c) Bernoulli Naïve Bayes. The accuracy, precision and MSPE measures of the three models are in table 1.

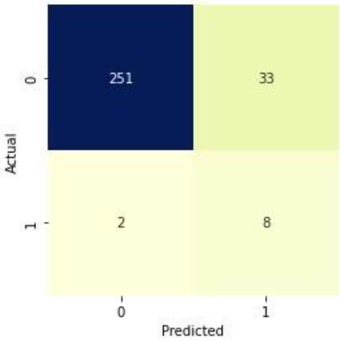
	KNN NB	Gaussian NB	Bernoulli NB
Precision	87.2	86.8	74.1
Accuracy	88.1	84.7	86.1
MSPE	0.119	0.153	0.139

From the above table it can be inferred that the mean squared prediction error by K- nearest neighbors is the smallest, there for the prediction of quality is closer to the actual value. On running the Knn classifier with k values 2, 3 , 4 and 5 , k value 4 yielded more accurate result.

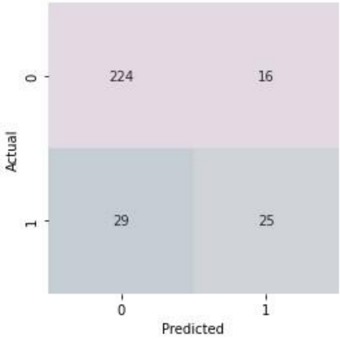
Therefore, to predict the quality of wine , knn classifier can be used to get the most accurate prediction which 88.1% with a probability of 33 False Negative and 2 false positive, which means that the probability of model to predict a bad wine as good is very very low.

CONFUSION MATRIX

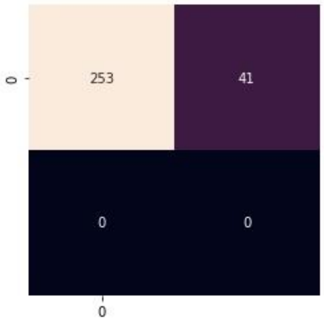
knn classifier



Gaussian



Bernoulli



Compared to the confusion matrix of Gaussian and Knn classifier the Bernoulli classifier does not predict any where close to the actual. This is because the Bernoulli classifier works better when the features are less and the target is binary.