Student Name: Tarek Ahmed

Student ID: 300315679

Topic: Report on predicting wine quality as final project for course CSIS-3275

* **Introduction and discovery:**

The dataset has been collected from UCI Machine Learning repository. Main problem I tried to solve by using this database is to predict wine quality. Organizations that make wines or the restaurant industries that consumes wines can be benefitted from this report and it may help to take good decisions regarding choosing best wine qualities.

There are thousands of qualities of wines available to the consumers. If consumers as well as wine creators can predict about wines quality by their raw materials it would be easier for them to calculate wine prices, estimate market value, estimate market demand thus they can improve their business and reduce work in producing wines that will not have that quality.

The dataset has 10 features that can participate in building the quality of the wine. Initially, I am assuming all 10 features will contribute to predict wine quality.

* **Data Preparation:**

The dataset has been downloaded from UCI Machine Learning repository. There are two dataset containing different wines like red wine and white wine. I have downloaded the dataset directly from their site.

The dataset has total 11 columns in total. These are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality.

Input variables (based on physicochemical tests):

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

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

So, from the two datasets, I have used red wine dataset because it has more rows 4898 which is very less than white wine dataset which has 1599 wines. All the features are having float data type and there is no none values.

To prepare the dataset, I had to rename the columns because it will be easier for me use the column name if it is separate. Another thing was important, which is transforming the quality of wines. Wine quality values ranges from 1 to 10 where 5,6,7 was the most frequent. So, I made the median column as threshold for the wine quality column. Values below 6 will get 0 indicating as not good quality and values more than 6 indicates values were good.

* **Model Planning and Implementation:**

To select the model properly, I have used several techniques taught in the course. Firstly, I used pipeline technology to see the accuracy of each individual classifiers. These are:

Naive Bayes, KNN, Decision Tree ,Random Forest ,Ada BOOST ,SVC.

And I also used feature transformation. Below is the data for different model with and without scaled data.

|  |  |  |
| --- | --- | --- |
| Algorithm | Accuracy with Scaled data | Accuracy without scaled data |
| Naïve Bayes | 0.675 | 0.711 |
| KNN | 0.745 | 0.724 |
| Decision Tree | 0.697 | 0.745 |
| Random Forest | 0.788 | 0.843 |
| Ada Boost | 0.706 | 0.752 |
| SVC | 0.726 | 0.768 |

From the chart above it is evident that RandomForestClassifier accuracy is the most. Single model achieved very good accuracy score among the others. Let us try to get the best hyperparameters for this model. For further investigation I used Random Forest Classifier because this model had more accuracy, so tweaking its parameters can help increasing the model accuracy.

I tried to get the best parameters by using GridSearchCV technique. Using

the best parameters I got from GridSearchCV. Best result has not improved much. It was still the same.­­­­

* Results Interpretation and Implications:

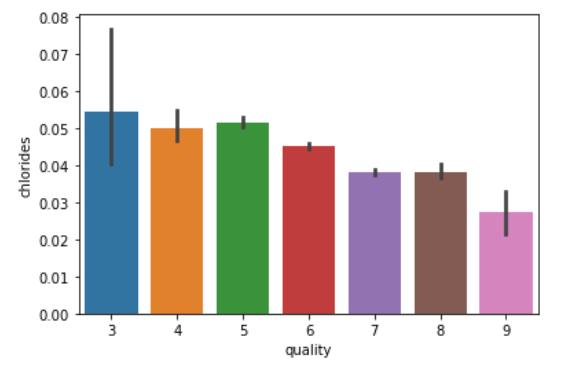
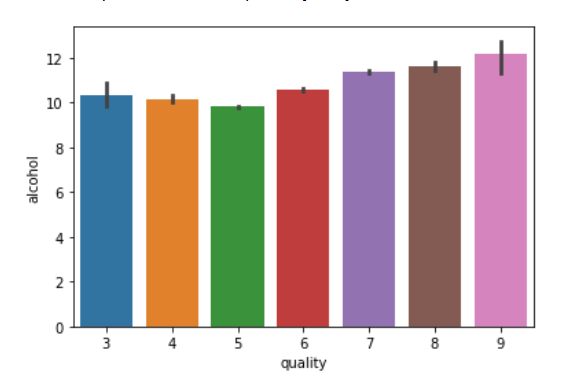
Let us look the accuracy table that I generated from the test data.

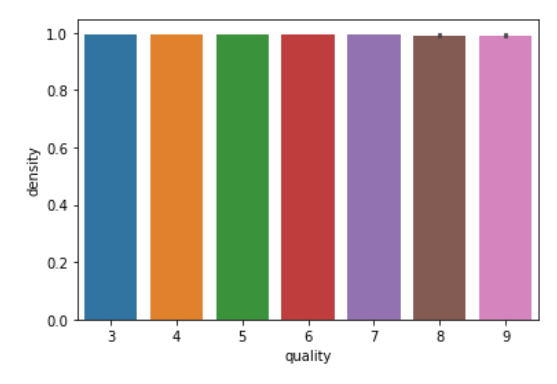
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| 0 | 0.78 | 0.72 | 0.75 | 321 |
| 1 | 0.87 | 0.90 | 0.89 | 659 |

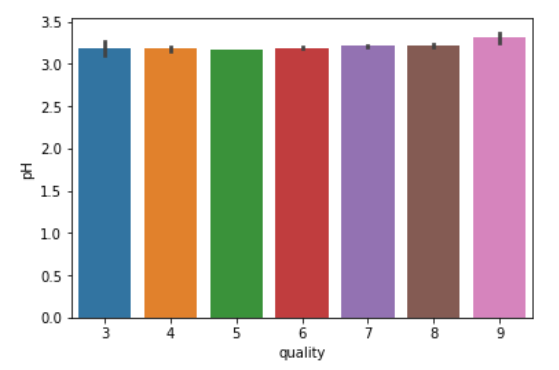
The model successfully predicted 0.834% data correctly.

According to my observation, more data will be required to predict the model perfectly. The more we have sample data, the more it will help to get perfect model and predict better.

Let us also some data analysis we can conclude from my experiment.







From the above graph we see

* When alcohol rate is increased it will be increased wine quality.
* When chlorides rate is increased quality has also increased.
* Quality and density seem to have no connection as it is parallel all the time.
* When pH rate is increased quality has also been increased.
* Concluding Remarks

After analyzing the project and considering the findings the manager or business owners can take some managerial decisions.

* They should improve the wine quality emphasizing on alcohol rate, chlorides, and pH rates.
* As the accuracy rate of the model is not above 95% they should try to improve the wine such that it goes above median.