[](http://jntuhceh.ac.in/) 

project report

Predicting the Quality of Wine

**TEAM MEMBERS:**

1. **M.Ramya (JNTUHCEH)**
2. **A. Sai Joshitha (JNTUHCEH)**
3. **T. Indu (JNTUHCEH)**
4. **Md.Hafiza Rizwana (Sri Vishnu Engineering College for Women)**

**Certificate**

This is to certify that \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, enrolled in the B. tech degree programme (Electronics and Communication Engineering) of the Jawaharlal

Nehru Technological University College of Engineering Hyderabad has successfully completed the four-week internship cum hands-on training program conducted by Smartbridge at Jawaharlal Nehru Technological University in ‘Introduction to Machine Learning using Python 3’ during the time period from 13th May, 2019 to 1st June 2019 under the guidance of Mr. K. Venkat Rao, Senior Data Consultant. During this period of internship with us he was found punctual, hardworking and inquisitive.

Acknowledgement

We hereby thank Dr.K.Venkat Rao who has given us the great opportunity to work on this project. He has been a great source of inspiration and his timely support and guidance has helped in the successful completion, whose creative suggestions and timely interventions helped us a lot during the process.

Then we would like to thank Mr. Sai Radha Krishna for his enthusiastic approach and dedication which has indeed been a great source of inspiration and support to us. We thank the department of Electrical and Communication, JNTU Hyderabad for providing us with necessary infrastructure to do our work.

**TABLE OF CONTENTS**

1. Abstract……………………………..………..04

2. Introduction……………………………..……05

3.Objectives of Research………………………..06

4.Industrial Profile………………………………07

5. Methodology…………….…………...............09

5.1 Exploratory Data Analysis………………10

5.2 Outliers…………………………………..11

5.3 Correlation……………………………….12

5.4 Multi Class Classifiers………………..13,14

5.5 Binary Classifiers………………………..15

6. Results and discussions ……………………...16

7. Accuracy Chart……………………………….17

8. Conclusions…………………………………..18

9. References……………………………………18

**ABSTRACT**

Wine is an alcoholic drink made from fermented grapes. Yeast consumes the sugar in the grapes and converts it to ethanol, carbon dioxide, and heat. Different varieties of grapes and strains of yeasts produce different styles of wine. These variations result from the complex interactions between the biochemical development of the grape, the reactions involved in fermentation and the production process. Many countries enact legal appellations intended to define styles and qualities of wine.

Wine quality strongly depends on the grape quality. To obtain high-quality wines, it is necessary to process healthy grapes at the correct ripeness stage and for this reason the farmer has to be especially careful in the prevention of parasite attacks on the grapevine. It is the result of a complex set of interactions, which include geological and soil variables, climate, and many variables, climate, and many viticulture decisions. Taken as a whole, they may be described as the terroir effect.

FACTORS THAT EFFECT THE QUALITY OF WINE:

While grape quality and climate plays a significant role, post-harvest winemaking techniques such as maceration, fermentation, extraction, and aging also influence wine flavors immensely.

**INTRODUCTION**

**Red wine** is a type of wine made from dark-colored (black) grape varieties. The actual color of wine can range from intense violet, typical of young wines, through to brick red for mature wines and brown for older red wines. The juice from most purple grapes is greenish-white, the red color coming from anthocyanin pigments present in the skin of the grape; exceptions are the relatively uncommon teinturier varieties, which produce a red-colored juice. Much of the red-wine production process therefore involves extraction of color and flavor components from the grape skin.

**What makes the quality of wine?**

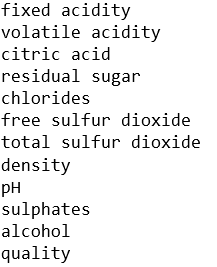
Many factors determine the quality of wine.

The types of grapes used, the condition of the environment, as well as the growing methods can all affect wine quality.

-There is an official formula that wine tasters use to determine the quality of taste.

*Sweet Taste (sugars + alcohols)🡸🡺acid taste(acids)+bitter taste(phenols)*

-Ultimately, there are many factors (both official and unofficial) that contribute to wine’s quality, and many organizations use different methods to measure their wine’s quality.

****

**OBJECTIVE OF RESEARCH:**

The scope of this analysis is to understand relationship of various parameters which impact the quality ratings for wine.

Description of attributes:

1. **Fixed acidity**: most acids involved with wine or fixed or non volatile (do not evaporate readily)
2. **Volatile acidity**: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
3. **Citric acid**: found in small quantities, citric acid can add ‘freshness’ and flavour to wines
4. **Residual sugar**: the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/litre and wines with greater than 45 grams/litre are considered sweet
5. **Chlorides**: the amount of salt in the wine
6. **Free sulphur dioxide**: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulphite ion; it prevents microbial growth and the oxidation of wine
7. **Total sulphur 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
8. **Density**: the density of water is close to that of water depending on the percent alcohol and sugar content
9. **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

10**. Sulphates**: a wine additive which can contribute to sulphur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant

11. **Alcohol**: the percent alcohol content of the wine

12. **Quality**: output variable (between 3 and 8)

**INDUSTRIAL PROFILE**

1. **Business understanding:**

This phase involves clearly defining the project objectives and goals, and translating these goals into a problem statement.

2. **Data understanding:**

This phase involves collection of data and performing a preliminary analysis on the data to evaluate the data quality. Data understanding phase may also contain making subsets of data that may have any actionable patterns.

3. **Data preparation:**

This phase is the most time taking one in the data mining process. It involves cleaning the data, performing certain transformations on the data to get the final dataset

4. **Modeling:**

This phase involves selecting the appropriate modeling technique

5. **Evaluation:**

The models generated during the modeling phase are evaluated for quality and also it’s determined whether the business objective is which means whether the problem statement is solved or not

6. **Deployment:**

In this phase the effective models are finally put to use. It may be making a simple report or using the insights in the daily functioning of a company

**DATA COLLECTION**

1. Name: Wine Quality Data Set

2. Input variables:

* fixed acidity
* volatile acidity
* citric acid
* residual sugar
* chlorides
* free sulfur dioxide
* total sulfur dioxide
* density
* pH
* alcohol

3. Output variable: Quality (score between 3 and 8)

4. Number of Observations: 1599

5. Number of Attributes/Variables: 12

6. Missing Values: N/A

**METHODOLOGY**

Steps we followed to build our models for predicting the quality of wine:

1. **Getting a feeling of the dataset**

First step is to see what kind of data you have at hand: How many attributes are available per instance, if the attributes are in numerical format, if preprocessing is required and more importantly if there are any missing attributes. The red wine dataset has only numerical data and no missing attributes which means that no special preprocessing is required.

2. **Exploring our attributes**

By exploring the attributes we observed that few inputs are positively correlated and few are negatively correlated. From heat map these are made clear and observations are done. Using the distribution plots, we observed which attributes show the normal curve.

3. **Outlier Detection**

Even after our preprocessing there are some values which are highly unexpected in comparison to the rest of the values of the same attribute in the dataset. We call these outliers. We observe such extreme values mainly in residual sugar and chlorides and only very few on pH and sulphates. The instances that have such extreme values in their attributes are <5% of the full dataset. We decided to remove completely these instances as they are expected to make learning more difficult and the scoring less accurate.

4. **Exploring our target, the quality of wine**

Our dataset is highly imbalanced as we have lots of wines being classified with qualities of 5 or 6 while only very few wines are classified are very bad (3) or as extremely good (8) which is expected. We used Classification task as good quality wine from 2-6.5 and bad quality wine from 6.5-8

5. **Splitting for Testing**

We are splitting our dataset in a way such that all of the wine qualities are represented proportionally equally in both training and testing dataset. Other than that the selection is being done randomly with uniform distribution.

5.1 Exploratory Data Analysis:

Statistical analysis of wine quality

|  | **volatile acidity** | **citric acid** | **chlorides** | **free sulfur dioxide** | **total sulfur dioxide** | **sulphates** | **alcohol** | **density** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 | 1599.000000 |
| **mean** | 0.527821 | 0.270976 | 0.087467 | 15.874922 | 46.467792 | 0.658149 | 10.422983 | 0.996747 |
| **std** | 0.179060 | 0.194801 | 0.047065 | 10.460157 | 32.895324 | 0.169507 | 1.065668 | 0.001887 |
| **min** | 0.120000 | 0.000000 | 0.012000 | 1.000000 | 6.000000 | 0.330000 | 8.400000 | 0.990070 |
| **25%** | 0.390000 | 0.090000 | 0.070000 | 7.000000 | 22.000000 | 0.550000 | 9.500000 | 0.995600 |
| **50%** | 0.520000 | 0.260000 | 0.079000 | 14.000000 | 38.000000 | 0.620000 | 10.200000 | 0.996750 |
| **75%** | 0.640000 | 0.420000 | 0.090000 | 21.000000 | 62.000000 | 0.730000 | 11.100000 | 0.997835 |
| **max** | 1.580000 | 1.000000 | 0.611000 | 72.000000 | 289.000000 | 2.000000 | 14.900000 | 1.003690 |

**wine['quality'].value\_counts()**

5 681

6 638

7 199

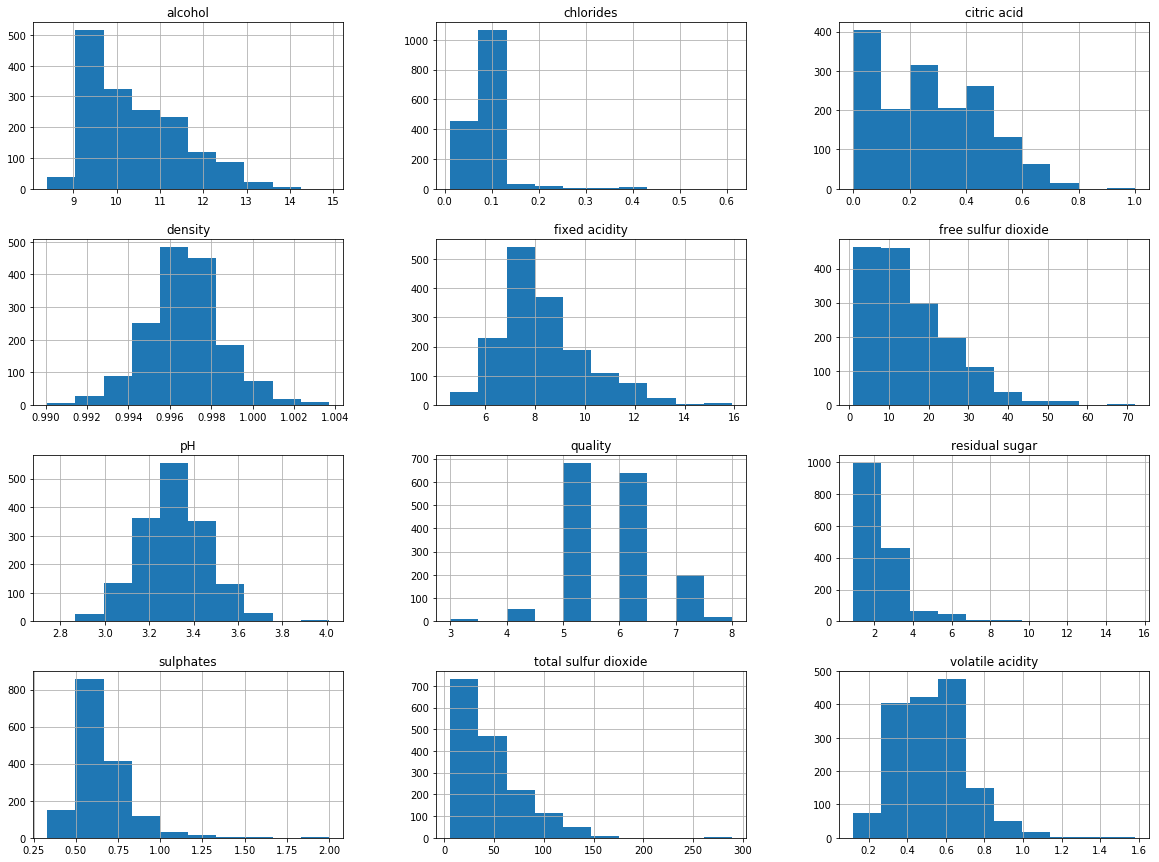
4 53

8 18

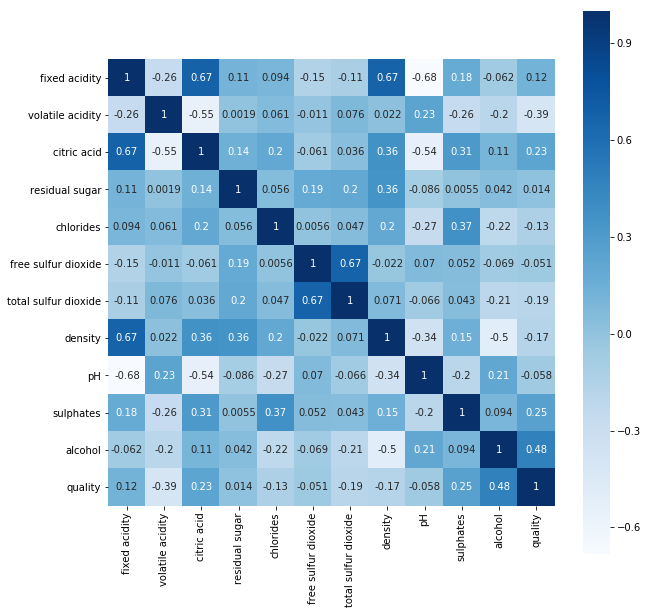
3 10

Name: quality, dtype: int64

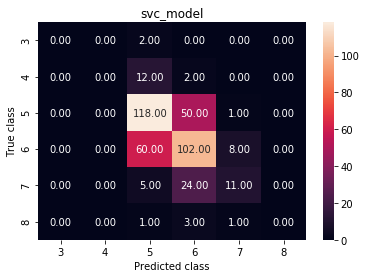
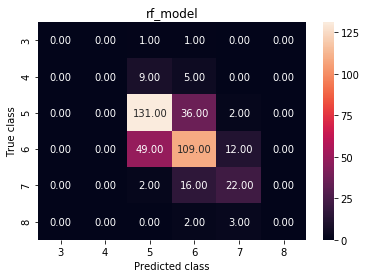
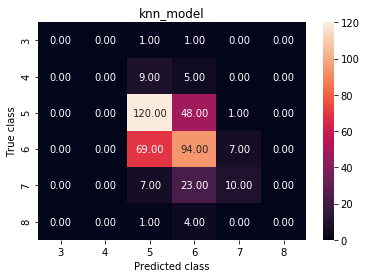
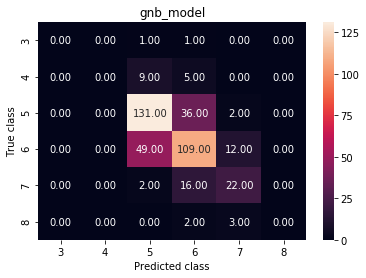
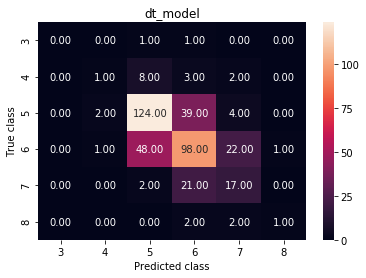
5.2 Outliers



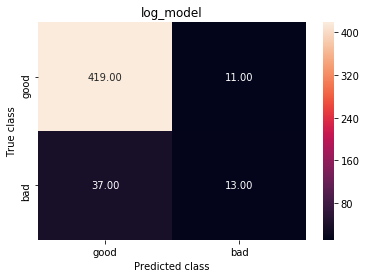
5.3 Correlation:



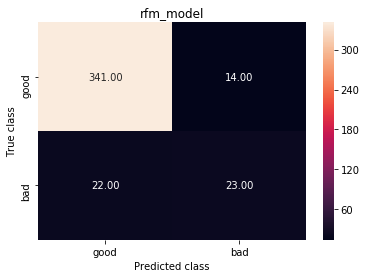
5.4 Multiclass Classifiers



5.5 Binaryclass Classifier



|  |  |  |
| --- | --- | --- |
|  | **Condition Positive (CP)** | **Condition Negative (CN)** |
| **Test Outcome Positive (OP)** | True Positive | False Positive |
| **Test Outcome Negative (ON)** | False Negative | True Negative |

****

**RESULTS AND DISCUSSIONS**

From data exploration, we can observe that few attributes are postively correlated wheras few are negatively correlated.This shows the relation between outputs and the inputs. Data visualisation gave a clear view of the given data. Histograms , Bar plots gave the outliers. The binary and multiclass classification analysis is measured according to the accuracy , For a good model the accuracy must be high.

By considering the attributes:

volatile acidity, citric acid, density,alcohol, total sulfur dioxide, free sulfur dioxide, chlorides,sulphates….

We used both multiclass and binary classification models,

In multi class classification we have to classify our output to 6 discrete classes(3,4,5,6,7,8) .

As a result accuracy decreases. So, we move on to Binary classification.

In this , 2-6.5 quality rate is classified as “Bad” wine and 6.5-8 as “Good” wine,which improved our accuracy.

From the accuracy chart for different models , the best model can be selected whose accuracy rate is high.

**ACCURACY CHART**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl.no** | **Classification model** | **Test size** | **Accuracy(%)** | **Cross\_val**  **Accuracy(%)** |
|  | Binary classifiers |  |  |  |
| 1. | Knn model | 0.1 | 60% | 48.5% |
|  |  | 0.25 | 56% | 48.5% |
|  |  | 0.4 | 52.1% | 48.5% |
| 2. | Decision tree | 0.1 | 60% | 53.7% |
|  |  | 0.25 | 59.5% | 53.9% |
|  |  | 0.4 | 62.3% | 53.85% |
| 3. | Svm | 0.1 | 62.5% | 48.9% |
|  |  | 0.25 | 57.75% | 48.9% |
|  |  | 0.4 | 53.2% | 48.9% |
| 4. | Random Forest | 0.1 | 69.3% | 57.36% |
|  |  | 0.25 | 65.25% | 56.2% |
|  |  | 0.4 | 63.5% | 55.49% |
| 5. | Gaussian Naive Bayes | 0.1 | 60.6% | 55.67% |
|  |  | 0.25 | 56.5% | 55.6% |
|  |  | 0.4 | 56.0% | 55.67% |
|  | Binary Classification |  |  |  |
| 1. | Logistic Regression | 0.1 | 90.6% | 86.49% |
|  |  | 0.25 | 90.5% | 86.49% |
|  |  | 0.4 | 88.7% | 86.49% |
| 2. | **Random Forest** | **0.1** | **91.2%** | **87.8%** |
|  |  | 0.25 | 91% | 87.5% |
|  |  | 0.4 | 90.5% | 87.2% |
|  |  |  |  |  |

**CONCLUSIONS**

From the present study of assesment of quality of red wine, it is found that some of the parameters such as alcohol, sulphates, volatile acidity, citric acid have much affect on wine quality. The higher value of volatile acidity tends to poor quality of wine. Similarly, lower amount of alcohol and citric acid tends to poor quality. Hence, we conclude that for maintaining good quality of wine:

1.Alcohol should be high in amount.

2.Volatile acidity should be low.

3.citric acid should be moderate.

4.sulphates amount must be moderate.

We suggest to maintain the same levels of other attributes as there may be no much effect on quality but it partially effects.

REFERENCES

-wikepedia.com

-github.com

-kaggle.com

-stackoverflow.com