Wine Quality Analysis

Deepthi B
Computer Science Dept.
PES UNIVERSITY
Bengaluru, India
SRN:PESUG19CS107

Amiya Mishra
Computer Science Dept.
PES UNIVERSITY
Bengaluru, India
SRN: PESUG19CS034

Deboleena Mukherjee
Computer Science Dept.
PES UNIVERSITY
Bengaluru, India
SRN: PESUG19CS102

Deepali S Attavar
Computer Science Dept.
PES UNIVERSITY
Bengaluru, India
SRN: PESUG19CS106

Abstract — The main purpose of this study is to predict wine quality based on physicochemical data. Wine Dataset which contains 1599 rows with 11 attributes of physicochemical data such as alcohol, PH, and sulphates. This study explores performance of various regression and classification algorithms including binary, multiclass classification as well as continuous predicted output. Comparative studies show binary classification outperforms other techniques with Random Forest performing the best. Feature evaluation shows alcohol has greatest influence on quality.

I. INTRODUCTION

To determine the quality of wine, sensory tests are used which rely on human expert's knowledge, but physicochemical properties of wine can also be used. The relationship between physicochemical and sensory analysis are complex and not yet fully understood, but significant correlations can be found between quality of wine and physicochemical properties. Data mining techniques are powerful techniques to analyse relationships between different attributes of a dataset. They can be used for classification, clustering, forecasting, optimization, and summarization. In wine industry, DM is used to make recommendations on purchase of wine, based on wine ratings, consumer criticisms, and wine prices. There are a large number of websites and mobile applications that make recommendations for choosing wines based on that information (Ex: www.go-wine.com). In this project, we present a classification of wines based on their physicochemical properties that are easily measurable and accessible. This analysis can be valuable to wine producers (to improve the production process), to consumers (to select wine), and to experts to support their evaluation of wine and to potentially improve the speed and quality of their decisions.

II. PROPOSED METHODOLOGY

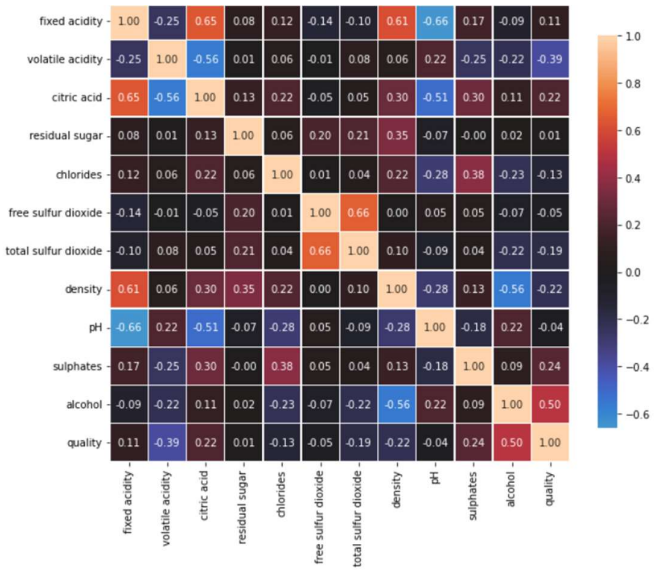
First data needs to be preprocessed for missing, inconsistent data. And then data needs to be checked for outliers and the outliers need to be removed. Since there are no categorical variables, there's no need for transformation. The data is then normalized. For model-building we need to divide the data into test and train data and the test size is taken as 0.2.

The x_train, y_train variables are then fit into various models. The y_test data need to be predicted. The accuracy of the train and test data need to computed. The performance needs to be optimized. This can be done by choosing different hyperparameters or hyperparameter tuning. It can also be done by reclassifying the output variable by making it a binary or n-ary valued variable. Various visualization can help choosing the right variables needed for this too.

III. PROPOSED SOLUTION

EDA and Visualization

- The dataset consists of following attributes volatile, acidity, citric acid, residual sugar, chlorides, free Sulphur dioxide, total Sulphur dioxide, density, pH, sulphates, alcohol, and quality.
- The dataset contains 1599 rows and 16 attributes. Most of the outliers have been removed as a part of exploratory analysis where initially outliers were present.
- There are no missing values on EDA and model building process.
- The heat map shows the correlation between attributes.



- The strongly correlated items are: 1. fixed acidity and citric acid. 2. Free Sulphur dioxide and total Sulphur dioxide. 3. fixed acidity and density. 4. Alcohol and quality. So, from above points there is a clear inference that alcohol is the

most important characteristic to determine the quality of wine. The weekly correlated items are 1.citric acid and volatile acidity. 2. fixed acidity and ph. 3. density and alcohol.

- These are some relations which do not depend on each other at all. This doesn't require dimensionality reduction as of now, as the attributes as different classifiers can take different no. of inputs. As there are only 11 attributes, which are all important in deciding the quality.
- However, as seen in the case studies, attributes can be removed while choosing and applying the model

The above it pair plot plots pair of attributes in the x and y axis with the target variable (quality), how the target variable clustered based on these two parameters.

1. Multiple linear regression

R^2 score is low and therefore this multiple regression model can't be used. Let us try classification methods

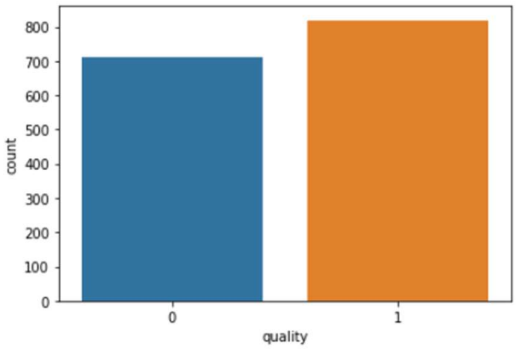
$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

2. Logistic Regression for Contiuous target variable

The accuracy is still not good. Let us try dividing it into 2 categories – good or bad.

3. 2-Class Classification

0 is considered bad and 1 is considered good. From this we can see that, quality is directly proportional to sulphates, alchohol, and inversely proportional to volatile acid, total sulpher dioxide, density, much more visibly than other attributes.



4. Logistic Regression
Training accuracy : 0.7440719542109566
Testing accuracy : 0.777777777777778

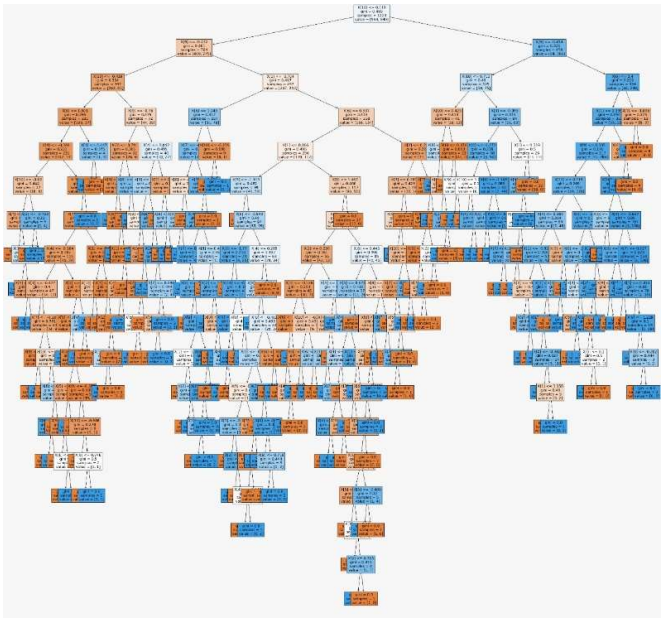
5. SVM

Hyper parameter tuning was done initially taking a large range of values of gamma from 0.01 to 1.2 skipping values and values of c. Once the best parameters were found, we reduced the range, so the search time reduces.
***C (Regularization)**: C is the penalty parameter, which represents misclassification or error term. The misclassification or error term tells the SVM optimization how much error is bearable. This is how you can control the trade-off between decision boundary and misclassification term. When **C** is high it will classify all the data points correctly, also there is a chance to over fit.*
***Gamma**: It defines how far influences the calculation of plausible line of separation. When gamma is higher, nearby points will have high influence; low gamma means far away points also be considered to get the decision boundary.*
Training accuracy : 0.8086672117743254
Testing accuracy : 0.7549019607843137

6. Decision tree classifier

Training accuracy: 1.0

Testing accuracy : 0.7124183006535948



7. Adaboost classifier
Training accuracy : 0.803761242845462
Testing accuracy : 0.7254901960784313

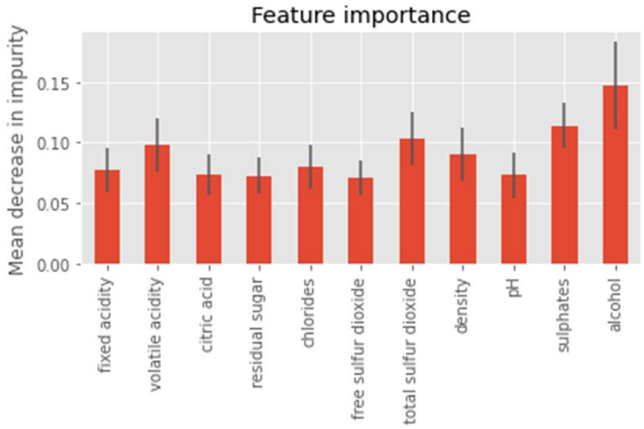
8. Naïve Bayes Bernoulli

It has the moderate accuracy.

Training set:0.716
Testing set: 0.728

9. Naïve Bayes Gaussian
Training set: 0.716
Testing set: 0.728

10. Random forest classifier:
Training accuracy : 1.0
Testing accuracy : 0.7908496732026143
Accuracy is best for Random forest.



11. Let us see how logistic regression classification into 3 output labels compares to classification into 2 output labels.

#3-Class-Logistic Regression Classification – second best
Training accuracy: 0.619233776387803

Testing accuracy: 0.571875

Accuracy is not as good as the 2-class logistic regression.

CONCLUSION:

Random forest is taken as the best classifier as it has highest test accuracy as the algorithm internally reduces the net variance.

REFERENCES:

1. How to classify wine using sklearn Naive Bayes mdeol in ML in python (projectpro.io)
2. TITLE: Classification based on Data-Mining Approach for Quality Control in Wine Prediction
   Authors: P.Appalasamy, A.Mustapha, N.D Rizal, F. Johari, A.F Mansor
   Publisher, year: Asian Network for Scientific Information 2012
   Dataset used: winequality-red.csv (winequality-white.csv, not taken in my
   analysis) Cortez et al, 2009
3. TITLE: Wine Quality prediction Model Using Machine Learning Techniques
   Authors: Rohan Dilip Kothawade
   Publisher, year: University of Skovde, 2021
   Dataset used: winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009
4. TITLE: Assessing wine quality using a decision tree.
   Authors: Seunghan Lee, Juyoung Park, Kyungtae Kang
   Publisher, year: 2015 IEEE International Symposium on Systems Engineering (ISSE)
   Dataset used: winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009
   Name of the model used: Decision tree: ID3
5. TITLE: Modelling Wine Quality from physicochemical properties.
   Authors: Dale Angus
   Publisher, year: Stanford University.
   Dataset used: winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009
   Name of the model used: Neural network model: binary classifier and multi-class classifier