# RED WINE QUALITY MODEL

Submitted by:

UJJWAL PRATIK

Contents

# ACKNOWLEDGMENT

For this study, I sincerely thank faculties of Data Trained for their detailed teaching method. Topic covered like Machine Learning and Neural language with python has helped a lot. There are also some references taken from YouTube videos and google search.

# INTRODUCTION

## About Wine

- Wine which was once viewed as a luxury product is increasingly enjoyed by a wider variety of customer today.
- Portugal is the 11th largest wine producer in the world and 9th largest wine exporter in the world.
- Quality of wine is graded based on the taste of wine and vintage, this process is time taking, costly and not efficient.
- A wine includes different parameters like fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphate, alcohol and quality.

## Problem Statement

- In industries, understandings the demands of wine safety testing can be a complex task for the laboratory with numerous analytes and residual to monitor.
- Our application's prediction, provide ideal solutions for the analysis of wine, which will make this whole process efficient and cheaper with less human interaction.

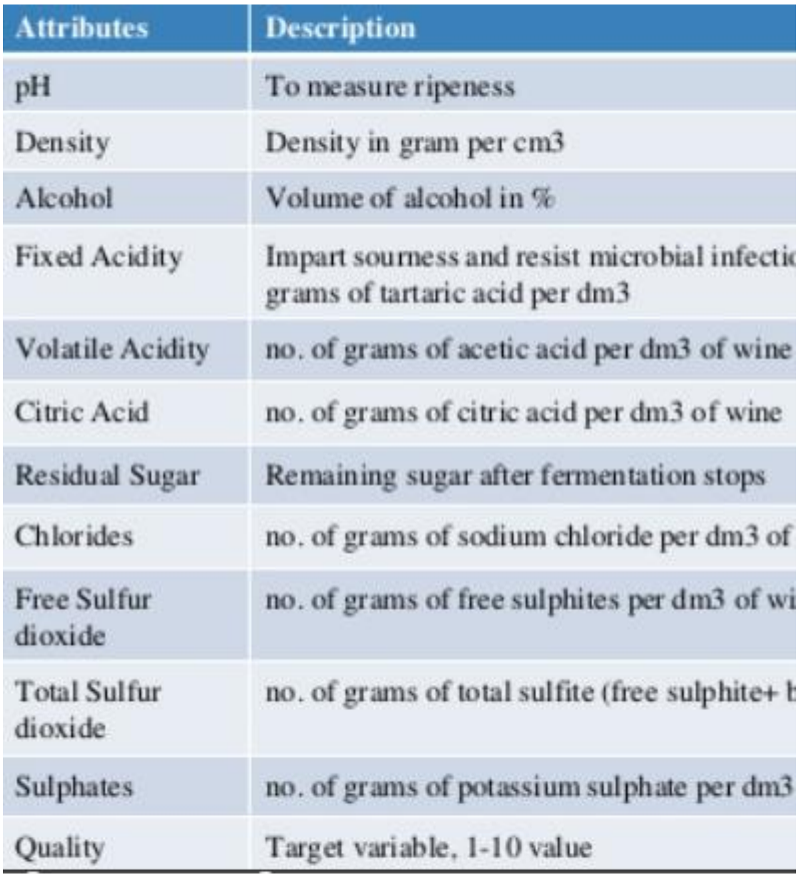
## Objective

- Our main objective is to predict the wine quality using machine learning algorithms.
- A large dataset is considered, and wine quality is modelled to analyse the quality of wine through different parameters like fixed acidity, volatile acidity etc.
- All these parameters will be analysed through Machine Learning algorithms like Decision Tree Classifier, Random Forest Classifier which will help to rate the wine on scale 1-10 or bad-good.
- It can support wine expert evaluations and ultimately improve the production.

## Data Description

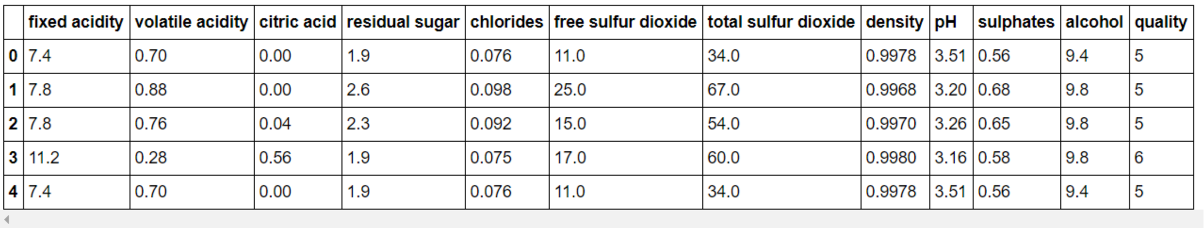
- The dataset contains chemical descriptions of Portuguese Red wine.
- The source of data is taken from GitHub.

# Data Formats



| Attributes | Description |
|---|---|
| pH | To measure ripeness |
| Density | Density in gram per cm3 |
| Alcohol | Volume of alcohol in % |
| Fixed Acidity | Impart sourness and resist microbial infecti grams of tartaric acid per dm3 |
| Volatile Acidity | no. of grams of acetic acid per dm3 of wine |
| Citric Acid | no. of grams of citric acid per dm3 of wine |
| Residual Sugar | Remaining sugar after fermentation stops |
| Chlorides | no. of grams of sodium chloride per dm3 of |
| Free Sulfur dioxide | no. of grams of free sulphites per dm3 of wi |
| Total Sulfur dioxide | no. of grams of total sulfite (free sulphite+ b |
| Sulphates | no. of grams of potassium sulphate per dm3 |
| Quality | Target variable, 1-10 value |

# Data sample looks using panda.



| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates | alcohol | quality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11.0 | 34.0 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 |
| 1 | 7.8 | 0.88 | 0.00 | 2.6 | 0.098 | 25.0 | 67.0 | 0.9968 | 3.20 | 0.68 | 9.8 | 5 |
| 2 | 7.8 | 0.76 | 0.04 | 2.3 | 0.092 | 15.0 | 54.0 | 0.9970 | 3.26 | 0.65 | 9.8 | 5 |
| 3 | 11.2 | 0.28 | 0.56 | 1.9 | 0.075 | 17.0 | 60.0 | 0.9980 | 3.16 | 0.58 | 9.8 | 6 |
| 4 | 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11.0 | 34.0 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 |

# Methodology

- It gives insights of the dependency of target variables on independent variables using machine learnings techniques to determine the quality of wine because it gives the best outcome for the assurance of quality of wine.
- The dependent variable is quality, whereas other variables i.e. alcohol, sulphur etc. are independent variables.
- While handling the effectiveness of the data model, various types of errors have occurred like over fitting, introduced from having too large of a training set and bias occur due to small of a test set.

# Model Building

a. Logistic Regression, where we got 45% accuracy.
**b.** Random Forest Classifier, where we got 81% accuracy.
**c.** Decision Tree Classifier, where we got 75% accuracy.
**d.** GaussianNB , where we got 48% accuracy.
**e.** Support Vector Classifier, where we got 38% accuracy.

# Hardware and Software Requirements and Tools Used

a. Hardware Requirement:
   i. Intel core i5
   ii. 8 GB Ram
b. Software Requirement:
   i. Python 3.x with packages:
      1. Pandas: Data analysis and manipulation tool
      2. NumPy: Provide support for mathematical functions, random number etc.
      3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
      4. Seaborn: is a library mostly used for statistical plotting in python.
      5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.
      6. Imb-learn for handling the imbalanced data.

# Identification of possible problem-solving approaches

Following models are used for solving the problem:

a. accuracy score: this function computes subset accuracy, the set of labels predicted for a sample must exactly match the corresponding set of labels in true.
b. Logistic Regression: Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results).
c. Random Forest Classifier: a collection of decision trees classifiers that each do their best to offer the best output.
d. Decision Tree Classifier: is a classification model that can be used for simple classification tasks where the data space is not huge and can be easily visualized.
e. GaussianNB: Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data.
f. Support Vector Classifier: is a widely used supervised learning method and it can be used for regression, classification, anomaly detection problems.
g. Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
h. Z-score: For checking and removal outliers in the dataset.
i. Datasets: overall data.
j. Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.

Following statistical and analytical approach followed:

a. regression coefficients are marginal results.
b. Started with univariate descriptive and graphs.
c. bivariate descriptive, again including graphs.
d. Model building and interpreting results.

# Testing of Identified Approaches (Algorithms)

a. Train Test Split
b. Logistic Regression
c. AdaBoost Classifier
d. Random Forest Classifier
e. Decision Tree Classifier
f. GaussianNB
g. Support Vector Classifier
h. Cross Validation
i. Hyper Parameter Tuning Using Grid Search Cv

## Key Metrics for success in solving problem under consideration.

1. Analysed data for any outliers and removed it by z-score method.
2. Analysed data for any skewness.
3. Handling class imbalance problem by oversampling the minority class.
4. Cross Validation for cross validates the accuracy-score from overfitting.
5. Hyper parameter tuning using Grid Search Cv for making the prediction better

## Conclusions

- Results will be used by wine manufacturers to improve the quality of the future wines.
- Certifications bodies can also use the result for quality control.
- Results can be used to make wine selection guides for wine magazines.
- Results can be used by consumers for wine selection.

## References

- Course Lectures and Notes
- Google Search
- YouTube
- GitHub
- http://archieve.ics.uci.edu/ml/datsetsWine+Quality

## Learning Outcomes of the Study in respect of Data Science

This study give me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc.

It also helped me to learn how to apply various model techniques on data and enable predications.