Wine Quality Prediction using Machine Learning Algorithms

S. Shiva Keerthi[1*], S.Ramya Teja[2], Ramsha Mehreen[3]

[1]Department of, Computer Science and Engineering SR University, Warangal, Telangana, India.
[2]Department of, Electronics and Communication Engineering SR University, Warangal, Telangana, India.
[3]Department of Computer Science and Engineering, SR University, Warangal, Telangana, India.

*Email: **sirigirishivakeerthi@gmail.com**

Abstract:
Wine classification is a difficult task since taste is the least understood of human senses. A good quality wine prediction can be very useful in the certification phase since currentlythe sensory analysis is performed by human tasters, being clearly a subjective approach. An automatic predictive system can be integrated into a decision support system helping the speed and quality of the performance. Furthermore, a feature selection processcan help to analyse the impact of the analytical tests. If it is concluded that several input variables are highly relevant to predict the wine quality, since in the production process some variables can be controlled, this information can be used to improve the wine quality. Classification models used here are Logistic Regression, Decision tree, and Random Forest.

Keywords: Wine classification, wine prediction, automatic predictive system, wine quality, Classification models, Logistic Regression Algorithm, Decision Tree Algorithm, Random Forest algorithm.

1. INTRODUCTION

The quality of the wine is a very important part for the consumers as well as the manufacturing industries. Industries are increasing their sales using product quality certification. Nowadays, all over the world wine is a regularly used beverage and the industries are using the certification of product quality to increases their value in the market. Previously, testing of product quality will be done at the end of the production, this is time taking process and it requires a lot of resources such as the need for various human experts for the assessment of product quality which makes this process very expensive. Every human has their own opinion

about the test, so identifying the quality of the wine based on humans experts it is a challenging task. There are several features to predict the wine quality but the entire features will not be relevant for better prediction.

The aim of this project is to predict the quality of wine whether it is good or bad with given a set of features as inputs. The dataset used is Wine Quality Classification Data set from Kaggle Website. Input variables are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, sulphates, alcohol. And the output variable is quality (0-Bad and 1-Good).

2. PROBLEM DEFINATION

This is a study of the wine quality prediction using various models of machine learning. This predicts whether or not the wine's quality is good. We use numerous Machine learning frameworks including pandas, matplotlib, sci-kit-learn, Keras etc. to analyze such a model.

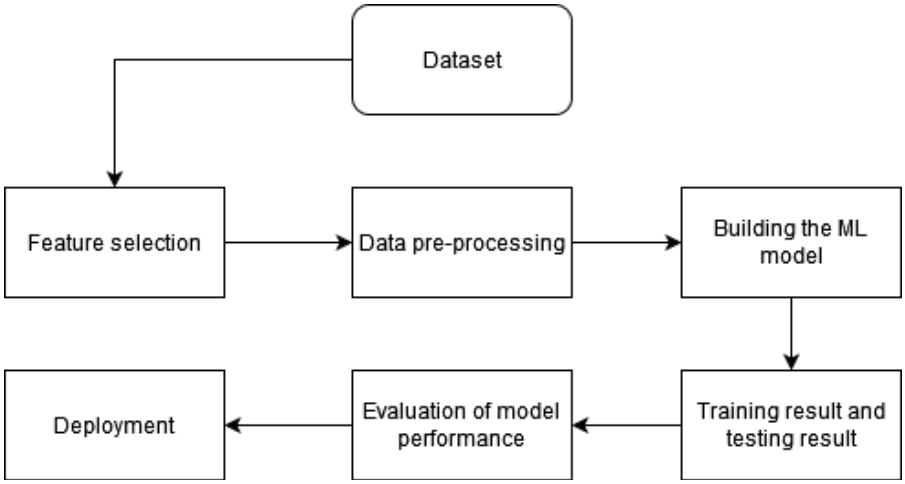


Figure 1. Processing Steps for Machine Learning for Spectacles Prediction

3. DATASET AND ATTRIBUTES

We collected the data from the Kaggle website. We use only 10 attributes to predict whether wine is good or bad.

Input features:

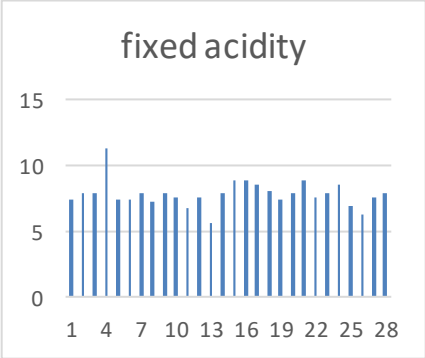
- Fixed acidity : are non-volatile acids that do not evaporate readily
  - Value : Numeric input
- Volatile acidity : are high acetic acid in wine which leads to an unpleasant vinegar taste
  - Value : Numeric input
- Citric acid : acts as a preservative to increase acidity (small quantities add freshness and flavor to wines)
  - Value : Numeric input

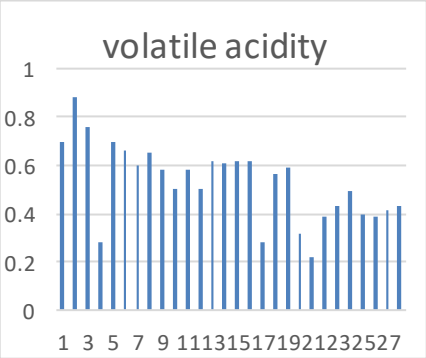
- Residual sugar : is the amount of sugar remaining after fermentation stops. The key is to have a perfect balance between — sweetness and sourness (wines > 45g/ltrs are sweet).
    - Value : Numeric input
- Chlorides : the amount of salt in the wine
    - Value : Numeric input
- Free sulfur dioxide : it prevents microbial growth and the oxidation of wine
    - Value : Numeric input
- Total sulfur dioxide : is the amount of free + bound forms of SO2
    - Value : Numeric input
- Density : sweeter wines have a higher density
    - Value : Numeric input
- Sulfates : a wine additive that contributes to SO2 levels and acts as an antimicrobial and antioxidant
    - Value : Numeric input
- Alcohol : the amount of alcohol in wine
    - Value : Numeric input

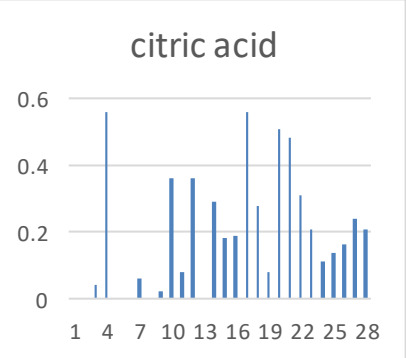
Output feature:

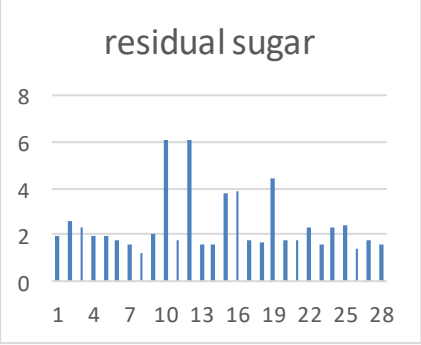
- Quality
    - Value : Good/Bad (Binary output)

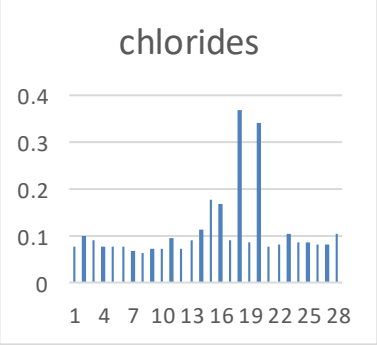
In total, we have 10 input features and 1 output feature i.e., 11 features for Wine Quality Prediction.











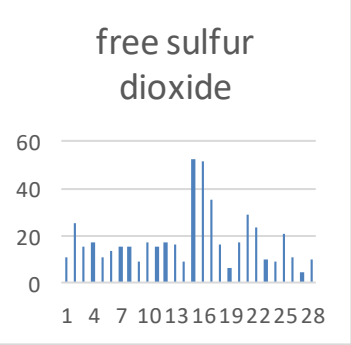


Figure 2: Visualizing attributes of the dataset

4. DATA PRE-PROCESSING

Real-world data collection has its own set of problems. It is often very messy which includes missing data, presence of outliers, unstructured manner, etc. Before looking for any insights from the data, we have to first perform preprocessing tasks which then only allow us to use that data for further observation and train our machine learning model. We use missing values treatment, outliers detection, normalization and data split to process our data before feeding it to the machine learning model.

Data info:

The wine prediction dataset is taken from Kaggle website it contains a large collection of datasets that have been used for the machine learning community. The wine dataset contains 1599 instances. And dataset have 10 input variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, sulfates, alcohol, and 1 output variable: quality (0-bad and 1-good). Below Table 1 description of the attributes.

Fixed acidity	Numeric [range= 0 to 15.9 to 1.58;mean= 8.319]
Volatile acidity	Numeric [range= 0.12 to 1.58;mean=0.527]
Citric acid	Numeric [range= 0 to 1;mean=0.270]
Residual sugar	Numeric [range= 0.9 to 15.5;mean=2.538]
Chlorides	Numeric [range= 0.012 to 0.611;mean=0.087]
Free sulfur	Numeric [range= 1 to 72;mean=15.874]

dioxide	
Total sulfur dioxide	Numeric [range= 6 to 289;mean=46.467]
Density	Numeric [range= 0.99 to 1.003;mean=0.996]
Sulfates	Numeric [range= 0.33 to 2;mean=0.658]
Alcohol	Numeric [range= 8.4 to 14.9;mean=10.422]
Quality	Numeric [range= 0 or 1;mean=0.534]

Table 1: Kaggle Dataset range and datatype.

Missing values treatment:

The real world's dataset often has many missing values which can be treated by using certain methods. But in our dataset, there are no missing values, because we collected the data from Kaggle website directly. To treat the missing values, we generally use the following strategies:

- Remove the entire row (If missing values are less in number)
- Replace the missing value with either mean or median
- Replace the missing value with most frequent value in the column (This is generally used only for large dataset)

```
print(data.isnull().sum())

fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
sulphates               0
alcohol                 0
quality                 0
dtype: int64
```

Figure 3: Missing values

Outliers detection and treatment:

Outliers are data points that don't fit the pattern of rest of the numbers. They are the extremely high or extremely low values in the data set. A simple way to find an outlier is to examine the numbers in the data set. We can also detect outliers by Z-score method, its formula is given by**:**

- **$Z = (x- \mu)/ \sigma$**
  - Where, x is each value in dataset,
  - μ is mean
  - σ is standard deviation

Normalization:

Normalization is a technique for organizing data in a database. Data normalization is the process of rescaling one or more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0. It is important that a database is normalized to ensure only related data is stored in each table and to avoid biasing towards huge values. When we normalize the data while

feeding it to the model, we also have to de-normalize it. This process can be done using the formulas below:

- $x_{nor} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$
- $y_i = y_{nor}(y_{max} - y_{min}) + y_{min}$

Data split:

To train any machine learning model irrespective what type of dataset is being used, we have to split the dataset into training and testing data. The reason to split the data is to give the machine learning model an effective mapping of input to outputs and to evaluate the model performance. We pass the training data to train our machine learning model and then test the model on testing data. In our model, we used 70:30 data split. That is the data is split 70% for training and 30% for testing. We can do the data split using train_test_split module in python.

Correlation Matrix

A correlation matrix is simply a table that displays the correlation. The measure is best used in variables that demonstrate a linear relationship between each other. The fit of the data can be visually represented in a scatterplot. coefficients for different variables.

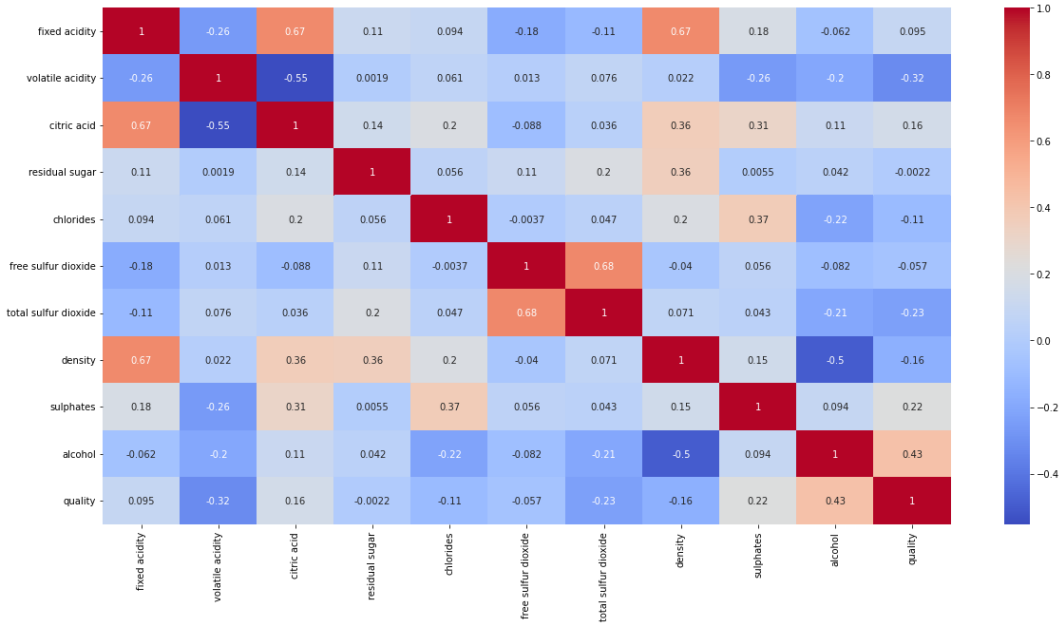


Figure 4: Correlation matrix

5. ALGORITHMS

We use 3 different machine learning models to solve our classification problem:

1. Logistic Regression
2. Decision tree algorithm
3. Random Forest Classifier

So, let us make our data ready for training and testing our machine learning model.

Logistic Regression

Logistic regression is a supervised algorithm for study classification. The likelihood of a destination variable was predicted. The nature of the target or dependent variable is dichotomous, meaning that only two possible classes are available [7]. Here, we use binary logistic regression which uses the sigmoid function to convert real numbers into binary and to calculate accuracies we use binary cross entropy loss function.

Decision tree

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute. This process is then repeated for the sub tree rooted at the new node [8]. We can perform decision tree using entropy and gini index.

Random Forest

Random forest is used for both regression and classification-based applications. This algorithm is flexible and easy to use. Most of the times this algorithm gives accurate results even without hyper tuning the parameters. It builds many decision trees which on merging forms as a forest. While building the decision trees, adds more randomness to the model. This algorithm searches for the best feature in the random subset of features, which results in the formation of a better model.

6. RESULTS

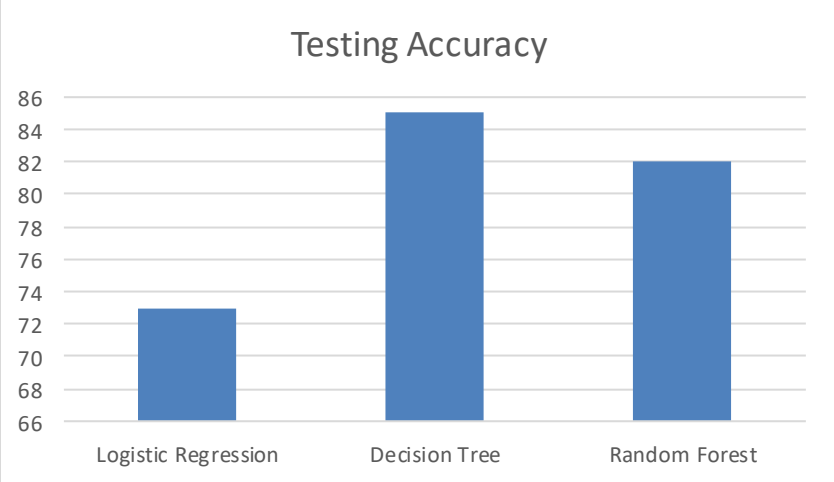


Figure 5: Graph depicting the accuracies of all models

Algorithm	Training error	Testing error
Logistic Regression	74%	73%
Decision tree	99%	85%
Random Forest	94%	82%

Table 2: Accuracy of all models

Cost	1 - 4	5	6	7	None
Entropy	0.699	0.789	0.815	0.832	1.0
Gini	0.699	0.78	0.729	0.727	1.0

Table 3: Decision tree training accuracy table

Cost	1 - 4	5	6	7	None
Entropy	0.699	0.718	0.78	0.852	0.76
Gini	0.699	0.725	0.725	0.747	0.76

Table 4: Decision tree testing accuracy table

Cost	1	2	3	4	None
Entropy	0.897	0.892	0.957	0.955	1.0
Gini	0.879	0.895	0.952	0.950	1.0

Table 5: Random forest training accuracy table

Cost	1	2	3	4	None
Entropy	0.731	0.691	0.781	0.735	0.82
Gini	0.718	0.714	0.768	0.766	0.81

Table 6: Random forest testing accuracy table

From the above tables, we can see that the highest accuracy is for Decision tree with 85%. Hence, we use it to deploy our model. The other models are also showing good accuracies of 82% and 73% for Random forest and Logistic regression models respectively.

7. CONCLUSION

Based on the above accuracy tables, we can conclude that for this project, decision tree has given us the highest accuracy of 85%. We have developed the web application using this decision tree model. There is a further scope of improvement with accuracy of the model, we can try several other models to improve the accuracy.

8. REFERENCES

[1] Yunhui Zeng1 , Yingxia Liu1 , Lubin Wu1 , Hanjiang Dong1. "Evaluation and Analysis Model of Wine Quality Based on Mathematical Model ISSN 2330-2038 E-ISSN 2330-2046,Jinan University, Zhuhai,China.

[2] Paulo Cortez1, Juliana Teixeira1, Ant´onio Cerdeira2."Using Data Mining for Wine Quality Assessment".

[3] Yesim Er*1 , Ayten Atasoy1. "The Classification of White Wine and Red Wine According to Their Physicochemical 2147-67992147-6799,3rd September 2016

[4] Ebeler S. (1999) "Flavor Chemistry — Thirty Years of Progress: chapter Linking flavour chemistry to sensory analysis of wine". Kluwer Academic Publishers, 409–422.

[5] Legin, Rudnitskaya, Luvova, Vlasov, Natale and D'Amico. (2003) "Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception". Analytica Chimica Acta 484 (1): 33–34.

[6] Sun, Danzer and Thiel. (1997) "Classification of wine samples by means of artificial neural networks and discrimination analytical methods". Fresenius Journal of Analytical Chemistry 359 (2) 143–149.

[7] Vlassides, Ferrier and Block. (2001) "Using historical data for bioprocess optimization: modeling wine characteristics using artificial neural networks and archived process information". Biotechnology and Bioengineering 73 (1) 55-68.

[8] Moreno, Gonzalez-Weller, Gutierrez, Marino, Camean, Gonzalez and Hardisson. (2007) "Differentiation of two Canary DO red wines according to their metal content from inductively coupled plasma optical emission spectrometry and graphite furnace atomic absorption spectrometry by using Probabilistic Neural Networks". Talanta 72 263–268.

[9] Yu, Lin, Xu, Ying, Li and Pan. (2008) "Prediction of Enological Parameters and Discrimination of Rice Wine Age Using Least-Squares Support Vector Machines and Near Infrared Spectroscopy". Agricultural and Food Chemistry 56 307–313.

[10] Beltran, Duarte-Mermoud, Soto Vicencio, Salah and Bustos. (2008) "Chilean Wine Classification Using Volatile Organic Compounds Data Obtained With a Fast GC Analyzer". IEEE Transactions on Instrumentation and Measurement 57 2421-2436.

[11] Cortez, Cerdeira, Almeida, Matos and Reis. (2009) "Modeling wine preferences by data mining from physicochemical properties". Decision Support Systems 47 547-553.

[12] Jambhulkar and Baporikar. (2015) "Review on Prediction of Heart Disease Using Data Mining Technique with Wireless Sensor Network". International Journal of Computer Science and Applications 8 (1) 55-59.

[13] Zaveri, and Joshi. (2017) "Comparative Study of Data Analysis Techniques in the domain of medicative care for Disease Predication". International Journal of Advanced Research in Computer Science 8 (3) 564-566.

[14] Portuguese Wine - Vinho Verde. Comissa˜o de Viticultura da Regia˜o dos Vinhos Verdes (CVRVV), http://www.vinhoverde.pt, July 2008