International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

Quality Prediction of Red Wine based on Different Feature Sets Using Machine Learning Techniques

Nikita Sharma

Student, Department of Computer Science & Engineering, IMS Engineering College, Ghaziabad, U.P, 201009, India

Abstract: ***We propose a data mining approach to predict human wine taste preferences that is based on easily available analytical tests at the certification step. A large dataset (when compared to other studies in this domain) is considered, with red vinho verde samples (from Portugal). Three regression techniques were applied, under a computationally efficient procedure that performs simultaneous variable and model selection. The support vector machine achieved promising results, outperforming the multiple regression and neural network methods. Such model is useful to support the oenologist wine tasting evaluations and improve wine production. Furthermore, similar techniques can help in target marketing by modeSling consumer tastes from niche markets. In this quality prediction testing is done on the 20 percent of the data and the training is done on the 80 percent of the data. All the results are according to training -0.8 & testing -0.2 Red Wine Dataset.***

Keywords: Regression, Classification, Support Vector Classifier, Gradient Boosting Classifiers, Random Forest, K-nn, Naïve Baye's, Decision Tree.

1. Advantages of Moderate Consumption

There is no standard measure of a moderate amount of alcohol a person needs to take. However, several factors determine the amount of alcohol that individuals can take. In her book Dakota Stevens (2010) notes that the influence of red wine of a person depends on a variety of factors; in particular such aspects like age, sex, experience, genetics and body weight determine an individual's moderate consumption rate.

There is a significant difference between when individuals drink alcohol while taking other food substances and when one is drinking alcohol without any food supplement. According to Dakota Stevens (2010) food substances reduce absorption rate of alcohol. In addition, children have a lower consumption rate compared to adults. Research has established that men have higher body water mass compared to women.

This explains why women are quickly intoxicated compared to men. However, most medical researchers assume that taking two glasses of wine a day is moderate drinking. The most abundant chemical substances in red wine are polyphenols and resveratrol. These substances protect body cells against damages caused by pathogens. This is one of the advantages of drinking red wine.

Attribute Information
Input variables (based on physicochemical tests):
1) Fixed acidity
2) Volatile acidity
3) Citric acid
4) Residual sugar
5) Chlorides
6) Free sulphur dioxide
7) Total sulphur dioxide
8) Density
9) pH
10) Sulphates
11) Alcohol

Output variable (based on sensory data):
12) Quality (score between 0 and 10)

2. Introduction

In recent years there is a modest increase in the wine consumption as it has been found that wine consumption has a positive correlation to the heart rate variability. With the increase in the consumption wine industries are looking for alternatives to produce good quality wine at less cost. Different wines have different purposes. Although most of the chemicals are same for different type of wine based on the chemical tests, the quantity of each chemical have different level of concentration for different type of wine. These days it is really important to classify different wine for quality assurance. In the past due to lack of technological resources it become difficult for most of the industries to classify the wines based on the chemical analyses as it takes lot of time and also need more money. These days with the advent of the machine learning techniques it is possible to classify the wines as well as it is possible to figure out the importance of each chemical analyses parameters in the wine and which one to ignore for reduction of cost. The performance comparison with different feature sets will also help to classify it in a more distinctive way. In this paper machine learning approach is proposed by considering based feature selection considering the classifiers, linear classifiers and probabilistic classifiers to predict the quality in red wine.

The structure of the paper is organized as follows: Section presents the past work related to this field. Section 3 describes about the methodologies used for this research work. Section 4 describes about the result of feature selection as well.

3. Related Works

In the past few attempts have been made to use different machine learning approaches and feature selection techniques to the wine dataset. Er and Atasoy proposed a

method to classify the quality of wines using three different classifier such as support vector machines, Random forest and k-nearest neighbourhood. They have used principal component analysis for feature selection and they found good result using Random forest algorithm. Chen et al proposed an approach that will predict the grade of wine using the human savory reviews. They have used hierarchical clustering approach and association rule algorithm to process the reviews and predict the wine grade and they found an accuracy of 85.25% while predicting the grade [4]. Thakkar et al., used analytical hierarchy process (ahp) to rank the attributes and then used different machine learning classifiers such as support vector machine and random forest and they found accuracy of 70.33% using random forest and 66.54% using SVM [5]. Reddy and Govndarajulu used a user centric clustering approach to recommend the product. They have used red wine data set for the survey purpose. They have allocated relative voting to the attributes based on the literature review. Then they assigned weight to the attributes using Gaussian Distribution Process. They judged the quality based on the user preference group [6]. The above past work motivated us to try different feature selection algorithm as well as different classifiers to compare the performance metrics. This paper proposed implementation of different machine learning techniques on the red wine dataset to analyse the quality.

4. Methodologies

Data Set Information

The dataset is related to red variants of the Portuguese "Vinho Verde" wine. 2009. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables available.

This dataset can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

Steps followed while testing quality:

1) Data Collection of Red Wine from public datasets.
2) Data preparation for building model.
3) Feature selection
4) Implementing machine learning techniques
5) Comparison of performance.
6) Interpretation of results

5. Results and Discussion

We have divided the data into two groups such as train data and test data. We trained each classifier based on the trained data and predict the power of classifier on the test data. So, each classifier able to show all the performance metrics such as accuracy. We build few graphs so that it will be easy to understand it in a better way. We also observed the data types present in the dataset and also analysed the data for null values.

The keys present in the dataset are-

```
Index (['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
       'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',
       'pH', 'sulphates', 'alcohol', 'quality'],
      dtype='object')
```

Figure 1: Keys in the datasets

On the basis of quality we build a bar plot.

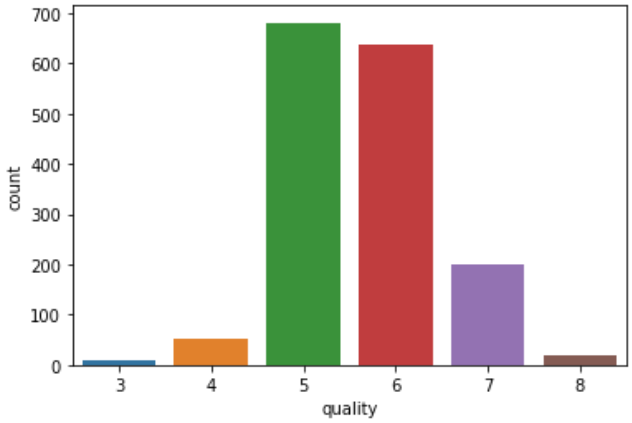


Figure 2: Quality vs Count plot

We made 3 categories of quality-Fine, Superior & Inferior

Fine	1319
Superior	217
Inferior	63

Name: rating, dtype: int64

We build a graph which tells the correlation between the features of the dataset.

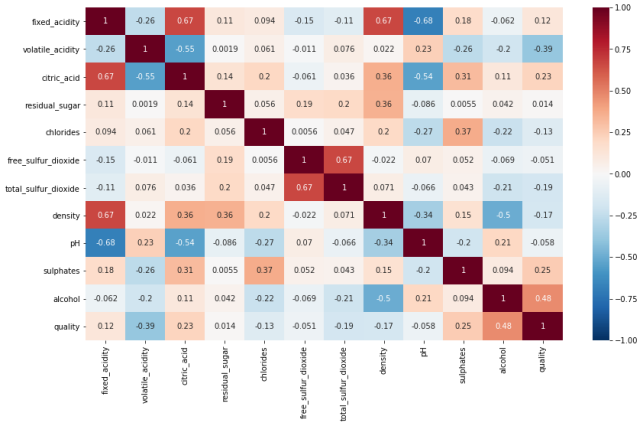


Figure 3: Correlation graph between features of Dataset

The correlation count is between -1.0 to 1.0 The highest correlation is higher or nearly 1.0 and the lowest correlated features have correlation nearly -1.0

The average glass of wine contains around 11% to 13% alcohol, but bottles range from as little as 5.5% alcohol by volume to as much as around 20% ABV. When tasting a wine, you'll notice alcohol comes through as heat in your back of your mouth or throat. A higher ABV wine will taste warmer and bolder; almost like a slight burning sensation on your palate. This graph shows the alcohol percent with respect to quality.

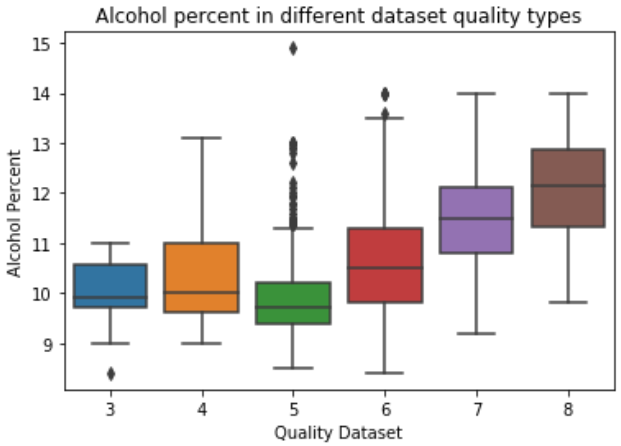


Figure 4: Alcohol % in quality of wine

Wine ranges from about 5 mg/L (5 parts per million) to about 200 mg/L. The maximum legal limit in the United States is 350 mg/l. A well made dry red wine typically has about 50 mg/l sulphites.

Wines with lower acidity need more sulphites than higher acidity wines. At pH 3.6 and above, wines are much less stable, and sulphites are necessary for shelf-life.

Wines with more colour (i.e, red wines) tend to need less sulphites than clear wines (i.e, white wines). A typical dry white wine may have around 100 mg/L whereas a typical dry red wine will have around 50–75 mg/L.

Wines with higher sugar content tend to need more sulphites to prevent secondary fermentation of the remaining sugar.

We tried to observe the sulphates present in the Red Wine and we noticed that there were many outliers present.

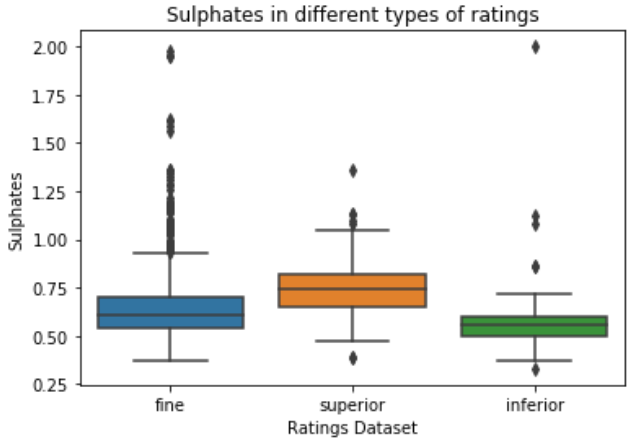


Figure 5: Sulphates present in different categories of Red Wine

PH is a measure of the acidity of wine. All wines ideally have a pH level between 2.9 and 4.2. The lower the pH, the more acidic the wine is, the higher the pH, the less acidic the wine. Each point of the pH scale is a factor of 10. This means a wine with a pH of 3 is 10 times more acidic than a wine with a pH of 4.

The pH of wine is typically adjusted by a process called malolactic fermentation. During this fermentation, malic acid is turned into lactic acid and carbon dioxide. This process serves to lower the pH of the wine to the desired level.

Of all the factors that affect the qualities of wine, pH is among the most important. The pH of a wine affects the flavour, texture, colour and smell of the wine. Different wines typically stay within certain pH values. Most white wines are between 3 and 3.3 pH. Most red wines are between 3.3 and 3.5 pH.

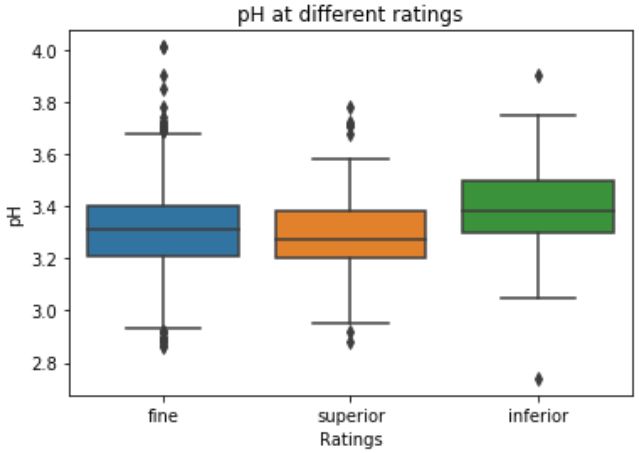


Figure 6: pH level in different ratings of wine

Wines range from 0–220 grams per litre sugar (g/L), depending on the style. in case you didn't know, dry-tasting wines contain up to 10 grams of sugar per bottle.

1) Bone-Dry <1 sugar calories per glass
2) Dry 0-6 sugar calories per glass
3) Off-Dry 6–21 sugar calories per glass
4) Sweet 21–72 sugar calories per glass
5) Very Sweet 72–130 sugar calories per glass

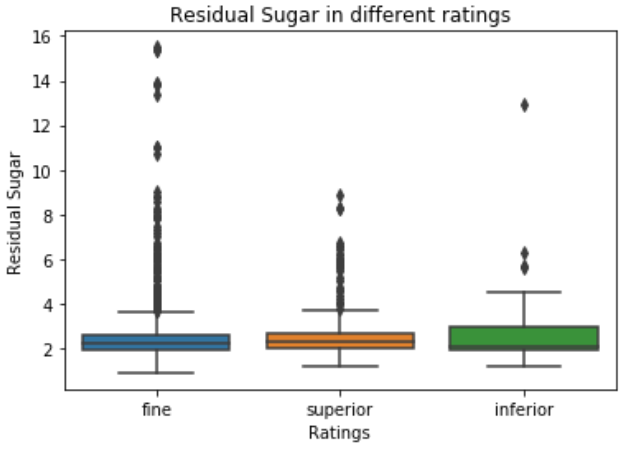


Figure 7: Residual Sugar in different ratings

A scatter matrix is a estimation of covariance matrix when covariance cannot be calculated or costly to calculate. The scatter matrix is also used in lot of dimensionality reduction exercises. If there are k variables, scatter matrix will have k rows and k columns .e. k X k matrix.

The scatter matrix is computed by the following equation:

$$S = \sum_{k=1}^{n} (x_k - m)\,(x_k - m)^T$$

where m is the mean vector

$$m = \frac{1}{n} \sum_{k=1}^{n} x_k$$

We tried to build the scatter matrix of the red wine dataset to visualize the trends in our data.wth respect to each feature present in the dataset.

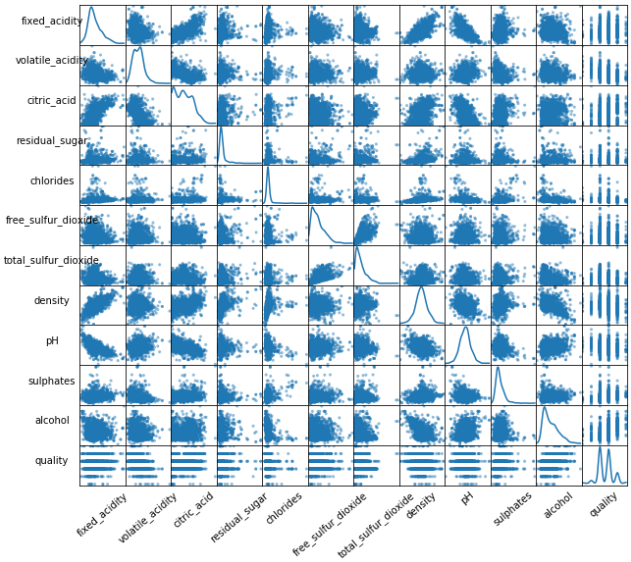


Figure 8: Scatter Matrix of the Red Wine Dataset

A density plot shows the **distribution** of a numerical variable. It takes only set of numeric values as input. It is really close to a histogram. To check the multiple distributions of data we visualized the density.

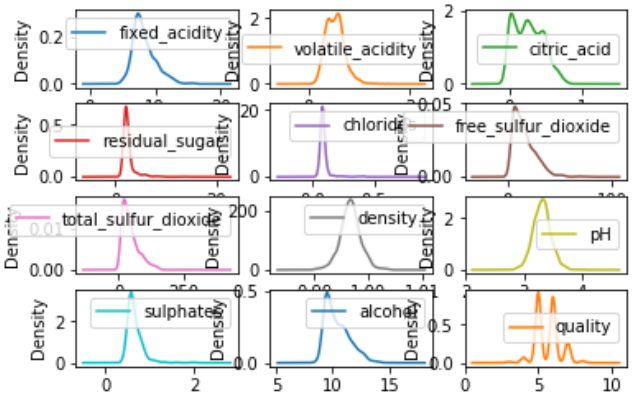


Figure 9: Density Plot of the data

Volatile acidity (VA) is a measure of the wine's volatile (or gaseous) acids. The primary volatile acid in wine is acetic acid, which is also the primary acid associated with the smell and taste of vinegar.

In lower-VA wines (less than 0.7 g/L acetic acid), wine makers can blend with a non-contaminated and lower-VA wine. Ensure that the high-VA wine is sterile filtered before blending.

In higher-VA wines (greater than or equal to 0.7 g/L acetic acid), wine makers can use reverse osmosis (RO) to lower the acetic acid concentration.
It is possible to use the kernel density estimation procedure described above to visualize a bivariate distribution.

In statistics, kernel density estimation is a non-parametric way to estimate the probability density function of a random variable.

We tried to build the kernel density joint plot of quality and volatile acidity.

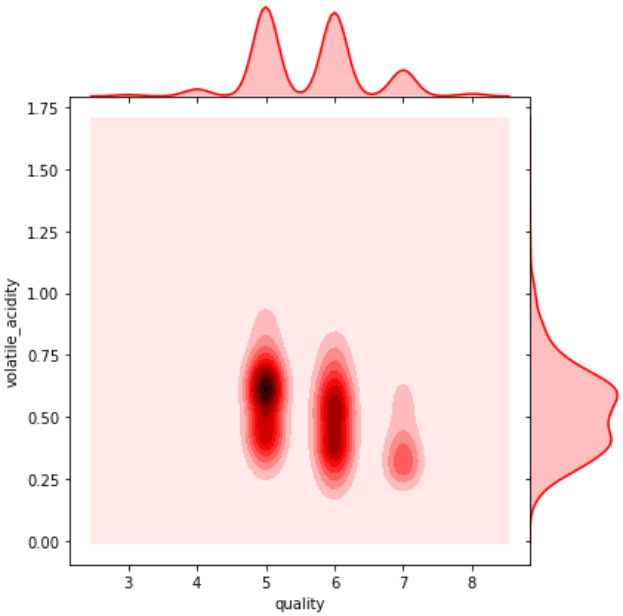


Figure 10: Joint Plot of Quality vs Volatile Acidity in the Wine

Histograms are good for showing general distributional features of dataset variables. You can see roughly where the peaks of the distribution are, whether the distribution is skewed or symmetric and if there are any outliers.

To check skewness and symmetry all the features are visualized with the help of Histogram Bins plot.

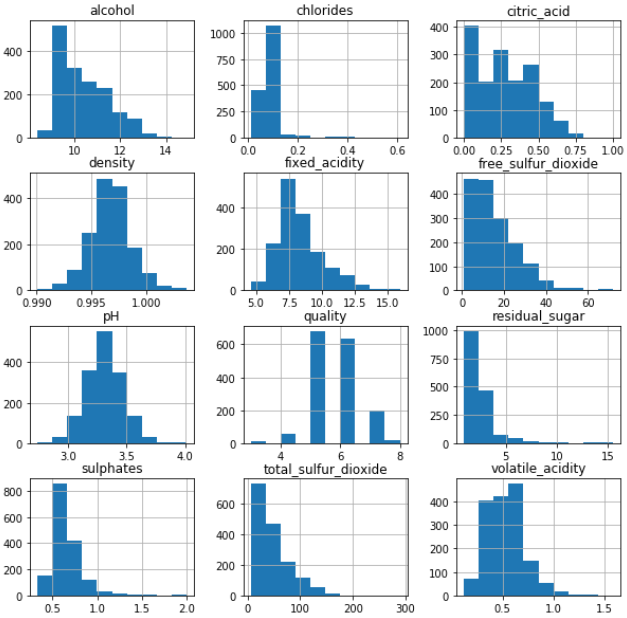


Figure 11: Histogram Bins of features in data

From interpreting the Regression Joint plot above, we can see a STRONG Negative correlation between pH level & fixed acidity. In other words, as one of these increases, the other decreases. According to the Heat map, these two features have a correlation co-efficient of -0.71

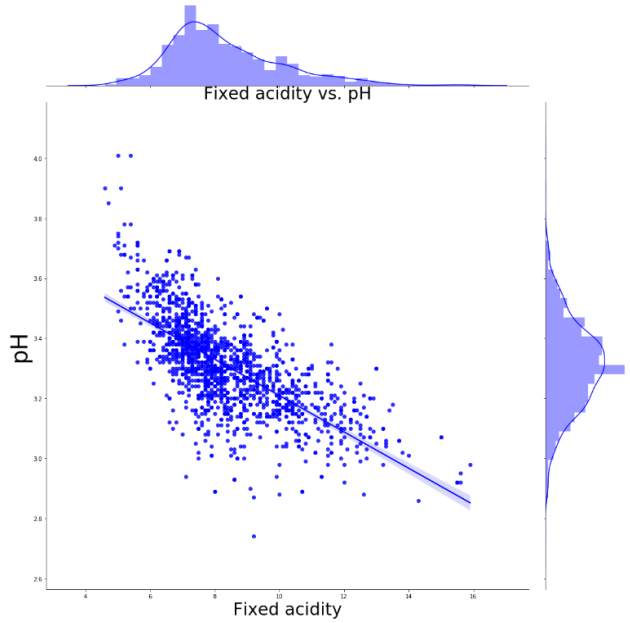


Figure 12: Correlation plot - Fixed Acidity vs pH

In the below Violin plot we observe that quality level 4,5 & 6 contains the outliers among which quality level 5 consists of maximum outliers with respect to the percent alcohol content in Red Wine Data.

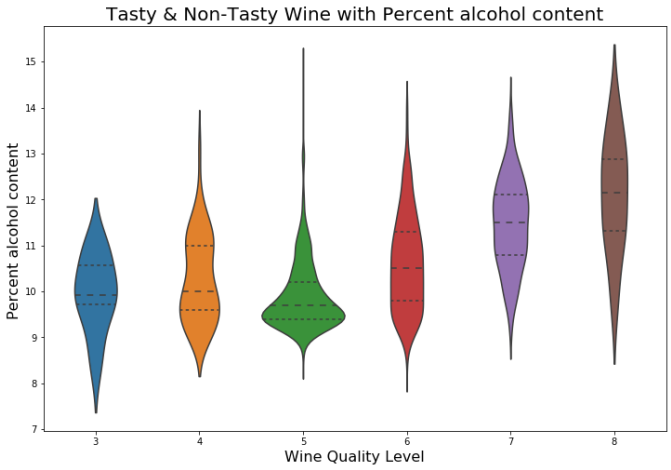


Figure 13: Violin Plot to observe outliers in Wine Quality and Percent Alcohol content

Techniques Description and Their Results:

1- Logistic Regression
2- K-nn
3- SVM
4- Random Forest
5- Decision Tree
6- XGB
7- Nave Bayes

Regression

Regression technique predicts a single output value using training data.

Classification

Classification means to group the output inside a class. If the algorithm tries to label input into two distinct classes, it is called binary classification. Selecting between more than two classes is referred to as multiclass classification.

Logistic Regression-

Logistic regression is one of the types of regression model where the regression analysis is executed when the dependent variable is binary. This regression method is used to explain the data and the relationship between the independent binary variable and one or more nominal, ratio-level independent variables.

The logistic regression analyses is widely used to determine the behavioural pattern of the individuals where they are able to predict an outcome. Most of the time the outcome can be ether yes or no. This regression analysis is widely used in each and every organization and based on the prediction made the decisions vary. Thus impacting the overall execution of the day to day process in an organization. It can also be used as a risk analysis tool.

$$y= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_n X_n$$

Where

y: is the dependent variable or target variable
$\mathbf{X_1, X_2}\ldots\ldots\mathbf{X_n}$ independent variables.
$\boldsymbol{\beta_0}$: the y – intercept.
$\boldsymbol{\beta_1}$: the slope.

The sigmoid function is given by- $\mathbf{p=1/1+e^{-y}}$

$$p = 1 / 1 + e^{-\beta 0 + \beta 1 X1 + \beta 2 X2 + \ldots\ldots + \beta n Xn}$$

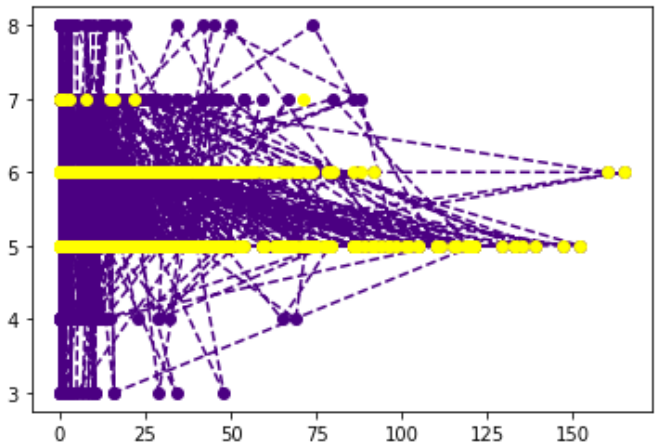


Figure 14: Scatter Graph in Log. Reg. (Actual Vs Predicted)

The observation in Logistic Regression is as follows:

Logistic Regression				
	Precision	Recall	F1-Score	Support
	0.00	0.00	0.00	2
	0.00	0.00	0.00	9
	0.68	0.79	0.73	134
	0.53	0.66	0.59	128
	0.00	0.00	0.00	43
	0.00	0.00	0.00	4
	Accuracy		0.60	320
	Macro avg	0.20	0.24	0.22	320
	Weighted avg	0.50	0.60	0.54	320

The exact accuracy is
Logistic Regression: 0.596875

K-Nearest-Neighbour:
K-nearest neighbours (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is manly used for classification predictive problems in industry. The following two properties would define KNN well −

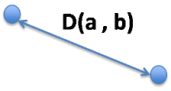
- Lazy learning algorithm − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
- Non-parametric learning algorithm − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

Working of KNN Algorithm:
K-nearest neighbours (KNN) algorithm uses 'feature similarity' to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. We can understand its working with the help of following steps −

Step 1 − For implementing any algorithm, we need dataset. So during the first step of KNN, we must load the training as well as test data.
Step 2 − Next, we need to choose the value of K i.e. the nearest data points. K can be any integer.
Step 3 − For each point in the test data do the following −
Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.
Now, based on the distance value, sort them in ascending order.
3.3 − Next, it will choose the top K rows from the sorted array.
3.4 − Now, it will assign a class to the test point based on most frequent class of these rows.
Step 4 – End

Euclidean Distance Formula:

$$D(a,b) = \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2}$$



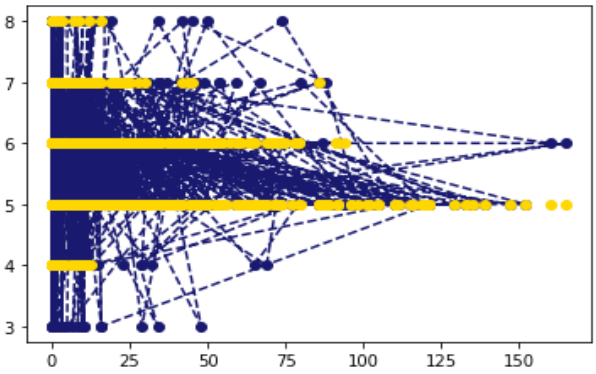


Figure 15: Scatter Graph in K-nn (Actual vs Predicted)

Observations in K-nn are as follows:

K-nn				
	Precision	Recall	F1-Score	Support
3	0.00	0.00	0.00	2
4	0.00	0.00	0.00	9
5	0.53	0.63	0.57	134
6	0.45	0.45	0.45	128
7	0.47	0.33	0.38	43
8	0.00	0.00	0.00	4
	Accuracy		0.49	320
	Macro avg	0.24	0.23	0.23	320
	Weighted avg	0.46	0.49	0.47	320

The exact accuracy is
K-nn: 0.4875

Support Vector Machine:
An SVM is implemented in a slightly different way than other machine learning algorithms. It is capable of performing classification, regression and outlier detection.

Support Vector Machine is a discriminative classifier that is formally designed by a separative hyperplane. It is a representation of examples as points in space that are mapped so that the points of different categories are separated by a gap as wide as possible. In addition to this, an SVM can also perform non-linear classification. Let us take a look at how the Support Vector Machine works. The main objective of a support vector machine is to segregate the given data in the best possible way. When the segregation is done, the distance between the nearest points is known as the margin. The approach is to select a hyperplane with the maximum possible margin between the support vectors in the given data-sets.

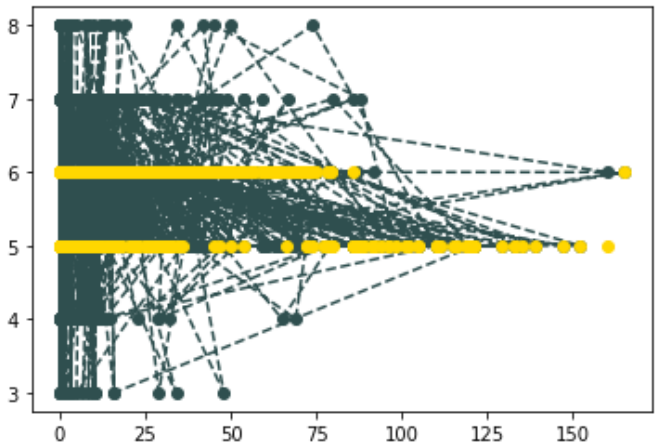


Figure 16: Scatter Graph in SVM (Actual vs Predicted)

The observations in SVM is as follows:

SVM				
	Precision	Recall	F1-Score	Support
3	0.00	0.00	0.00	2
4	0.00	0.00	0.00	9
5	0.74	0.31	0.44	134
6	0.44	0.91	0.60	128
7	0.47	0.00	0.00	43
8	0.00	0.00	0.00	4
	Accuracy		0.5	320
	Macro avg	0.20	0.20	0.17	320
	Weighted avg	0.49	0.50	0.42	320

The exact accuracy is
SVM: 0.496875

Random Forest:
Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting. Random Forest works in two-phase first is to create the random forest by combing N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps:
Step-1: Select random K data points from the training set.
Step-2: Build the decision trees associated with the selected data points (Subsets).
Step-3: Choose the number N for decision trees that you want to build.
Step-4: Repeat Step 1 & 2.
Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

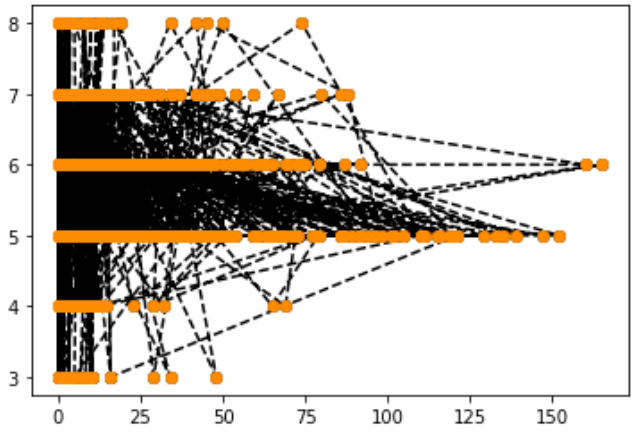


Figure 17: Scatter Graph in Random Forest (Actual vs Predicted)

The observations in Random Forest are as follows:

Random Forest				
	Precision	Recall	F1-Score	Support
3	1.00	1.00	1.00	2
4	1.00	1.00	1.00	9
5	1.00	1.00	1.00	134
6	1.00	1.00	1.00	128
7	1.00	1.00	1.00	43
8	1.00	1.00	1.00	4
	Accuracy			1.00 320
	Macro avg	1.00	1.00	1.00 320
	Weighted avg	1.00	1.00	1.00 320

The exact accuracy is
Random Forest: 1.0

Decision Tree
A decision tree is a map of the possible outcomes of a series of related choices. It allows an individual or organization to weigh possible actions against one another based on their costs, probabilities, and benefits. They can be used ether to drive informal discussion or to map out an algorithm that predicts the best choice mathematically.

A decision tree typically starts with a single node, which branches into possible outcomes. Each of those outcomes leads to additional nodes, which branch off into other possibilities. This gives it a treelike shape.

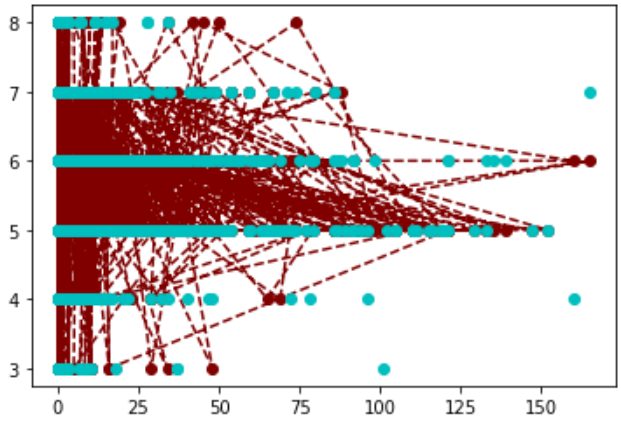


Figure 18: Scatter Graph in Decision Tree (Actual vs Predicted)

The observations of Decision Tree are as follows:

Decision Tree				
	Precision	Recall	F1-Score	Support
3	0.00	0.00	0.00	2
4	0.00	0.00	0.00	9
5	0.68	0.71	0.69	134
6	0.67	0.64	0.66	128
7	0.60	0.588	0.59	43
8	0.00	0.00	0.00	4
	Accuracy			0.63 320
	Macro avg	0.32	0.32	0.32 320
	Weighted avg	0.63	0.63	0.63 320

The exact accuracy is
Decision Tree: 0.63125

XGB:
The XG Boost library implements the gradient boosting decision tree algorithm. This algorithm goes by lots of different names such as gradient boosting, multiple additive regression trees, stochastic gradient boosting or gradient boosting machines. Boosting is an ensemble technique where new models are added to correct the errors made by improvements can be made. A popular example is the AdaBoost algorithm that weights data points that are hard to predict.

Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is

called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

This approach supports both regression and classification predictive modelling problems.
Salient features of XG Boost which make it different from other gradient boosting algorithms include:

- Clever penalization of trees
- A proportional shrinking of leaf nodes
- Newton Boosting
- Extra randomization parameter

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$$

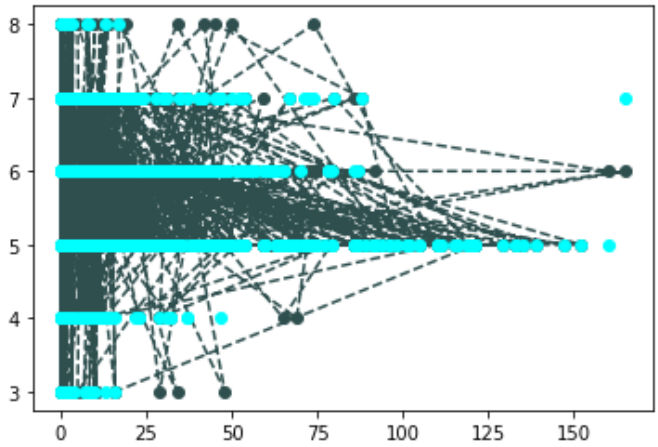


Figure 19: Scatter Graph in XGB (Actual vs Predicted)

The observations of XGB are as follows:

XGB				
	Precision	Recall	F1-Score	Support
3	0.00	0.00	0.00	2
4	0.00	0.00	0.00	9
5	0.78	0.78	0.78	134
6	0.63	0.70	0.66	128
7	0.60	0.49	0.54	43
8	0.00	0.00	0.00	4
	Accuracy		0.67	320
	Macro avg	0.33	0.33	0.33	320
	Weighted avg	0.66	0.66	0.66	320

The exact accuracy is
XGB: 0.66875

Naïve Bayes:
Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is manly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the bass of the probability of an object.

Baye's Theorem:
Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of hypothesis with prior knowledge. It depends on the conditional probability:

The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,
P(A|B) is Posterior probability: Probability of hypotheses A on the observed event B.
P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypotheses is true.
P(A) is Prior probability: Probability of hypotheses before observing the evidence.
P(B) is Marginal Probability: Probability of Evidence.

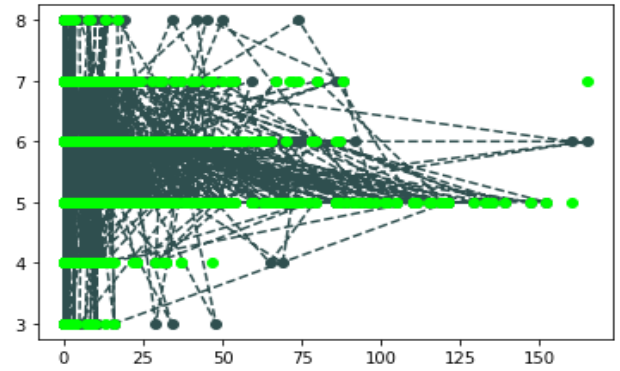


Figure 20: Naïve Baye's Graph

The observations of Naïve Baye's technique are as follows

Naïve Baye's				
	Precision	Recall	F1-Score	Support
3	0.00	0.00	0.00	2
4	0.00	0.00	0.00	9
5	0.72	0.62	0.67	134
6	0.51	0.59	0.55	128
7	0.33	0.28	0.30	43
8	0.00	0.00	0.00	4
	Accuracy		0.53	320
	Macro avg	0.26	0.25	0.25	320
	Weighted avg	0.55	0.53	0.54	320

The Exact Accuracy is
Naive Baye's: 0.534375
Based on the all techniques we have build a graph which will give use the rough idea about the accuracy and performance of each and every technique used for predicting the Red Wine Quality in the dataset.

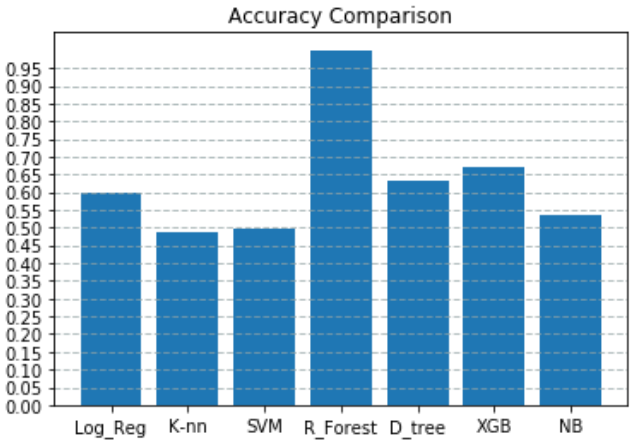


Figure 21: Accuracy Comparison with different techniques

The result shows that the accuracy ranges from 48.75% to 100% in the techniques. Specially it is performing better in Random Forest. Although it is easy to say based on our result that Random Forest is better algorithm for feature selection compared to the rest of the techniques, however the result could be different for other datasets as well as it could be different for bigger datasets. Similarly based on our result we can say that Random Forest classifiers best, but in practical lot of other parameters also come into picture that could change the scenario completely. This analysis will give a clear idea about the important attributes for the prediction of quality as well as it saves lot of time and money for the industries.

6. Conclusion and Future Work

This paper mentioned about predicting the quality of Red Wine using various machine learning techniques. The feature select on algorithm provided a clear idea about the importance of the attributes for prediction of quality, which was time consuming and expensive when done in the traditional way. We have also compared the accuracy of each technique used in prediction of quality and it was found that these classifiers performed well. We have found that the Random Forest based feature sets performed better than others. We have also found that the Random Forest classifier performed better compared to all other classifiers for red w ne data set.

In future we can try other performance measures and other mach ne learning techniques for better comparison on results. This analysis will help the industries to predict the quality of the different type of wines based on certain attributes and also it will helpful for them to make good product in the future.

References

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modelling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009
[2] https://archive.ics.uci.edu/ml/datasets/wine+quality
[3] https://www.researchgate.net/publication/221612614_Using_Data_Mining_for_Wine_Quality_Assessment
[4] B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen,"Wine informatics: applying data mining on wine sensory reviews processed by the computational wine wheel," *IEEE International Conference on Data Mining Workshop*, pp. 142-149, Dec. 2014.
[5] K.Thakkar,J.Shah,R.Prabhakar,A.Narayan,A.Joshi, "AHP and MACHINE LEARNING TECHNIQUES for Wine Recommendations,"*International Journal of Computer Science and Information Technologies*,7(5),pp. 2349-2352 ,2016
[6] Reddy, Y. S., & Govindarajulu, P. (2017). An Efficient User Centric Clustering Approach for Product Recommendation Based on Majority Voting: A Case Study on Wine Data Set. *IJCSNS*, 17(10), 103.

Author Profile

Nikta Sharma is a final year student in the Department of Computer Science & Engineering, MS Engineering College, Ghaziabad, Uttar Pradesh, 201009. Her area of interest is programming in python, Machine Learning, Deep Learning. Her research interest is Machine Learning.