International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 3, ISSN No. 2455-2143, Pages 519-529
Published Online July 2019 in IJEAST (http://www.ijeast.com)

ANALYZING WINE TYPES AND QUALITY USING MACHINE LEARNING TECHNIQUES

Sowmya D
Department of CSE
JNNCE, Shivamogga, Karnataka, India

Sayyed Johar
Department of CSE
JNNCE, Shivamogga, Karnataka, India

Ganavi M
Department of CSE
JNNCE, Shivamogga, India

Sankhya N Nayak
Department of CSE
JNNCE, Shivamogga, India

***Abstract*— This paper helps to solve the major problems by leveraging Machine Learning and data analysis on wine quality dataset by Training, Predicting & Evaluating Model using Decision Tree, Random Forests and predict if each wine sample is a red or white wine and predict the quality of each wine sample, which can be low, medium, or high Wine is a beverage from fermented grape and other fruit juices with a lower amount of alcohol content. Quality of wine is graded based on the taste of wine and vintage.. Tasting it is an ancient process as the wine itself is. When it comes to the quality of the wine, many other factors or attributes come into consideration other than the flavour. The dataset that to analyse 'Wine Quality', represents the quality of wines (white & red) based on different physiochemical attributes (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, and alcohol). The quality score for each wine combination in the dataset varies from 0 to 10 (ranging from least to highest). This analysis will uncover some important relationships between wine chemical contents like acidity and sugar levels versus its quality. The dataset exhibits a vast and distinct chemical and acidic combination of two types of wine (white & red). By employing smart data analysis techniques, a hand full of important and interesting insights that would be helpful in predicting wine quality and type that would also be prolific for the economic/financial sector and business sector of the production company can be unearthed**

***Keywords—data set, machine learing, decision tree,* physiochemical attributes**

I. INTRODUCTION

Wine (from Latin vinum) is an alcoholic beverage made from grapes, fermented without the addition of sugars, acids, enzymes, water, or other nutrients. Yeast consumes the sugar in the grapes and converts it to ethanol and carbon dioxide. 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, the terroir and the production process.

Wines not made from grapes include rice wine and fruit wines such as plum, cherry, pomegranate and elderberry. Wine has been produced for thousands of years.
Wine quality assessment is one of the key elements in this context and this assessment can be used for certification. Such type of quality certification helps to assure wine quality in market. Attributes of wine will decide the quality of the wines. In the past few years, with the availability of lot of wine brands it is difficult to identify the good quality wines. Good quality wine depends on the so many important factors such as chemical, scientific as well as technical factors. In the last few year's machine learning techniques caught lot of attention in every field. Most of the machines learning techniques are able to produce highly accurate result that compels most of the data scientist to implement it in case of predictive analytics. Here we considered only red wine and white wine analysis. Red wine is made from dark-colored grape varieties. The actual color of the wine can range from violet, typical of young wines, through red for mature wines, to brown for older red wines. The juice from most purple grapes is actually greenish-white; the red color comes from anthocyan pigments (also called anthocyanins) present in the skin of the grape; exceptions are the relatively uncommon teinturier varieties, which actually have red flesh and produce red juice. Fermentation of the non-colored grape pulp produces white wine. The grapes from which white wine is produced are typically green or yellow. Other white wines are blended from multiple varieties; Tokay, Sherry, and Sauternes are examples of these. Dark-skinned grapes may be used to produce white wine if the wine-maker is careful not to let the skin stain the wort during the separation of the pulp-juice.

A major challenge faced by analysing wine type and quality is getting perfect accuracy within less time in the final result. This can be achieved by using some machine learning techniques.

Wine Attributes and Properties
Different attributes which are present in the wine which determine the wine types and its quality. Properties and attributes we have considered are:

- **Fixed acidity:** Acids are one of the fundamental properties of wine and contribute greatly to the taste of the wine. Reducing acids significantly might lead to wines tasting flat. This variable is

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 3, ISSN No. 2455-2143, Pages 519-529
Published Online July 2019 in IJEAST (http://www.ijeast.com)

usually expressed in *g(tartaric acid)/dm3* in the dataset.

- **Volatile acidity:** These acids are to be distilled out from the wine before completing the production process. It is primarily constituted of acetic acid. The volatile acidity is expressed in *g(acetic acid)/dm3* in the dataset.
- **Citric acid:** This is one of the fixed acids which gives a wine its freshness. Usually most of it is consumed during the fermentation process and sometimes it is added separately to give the wine more freshness. It's usually expressed in *g/dm3* in the dataset.
- **Residual sugar:** This typically refers to the natural sugar from grapes which remains after the fermentation process stops, or is stopped. It's usually expressed in *g/dm3* in the dataset.
- **Chlorides:** This is usually a major contributor to saltiness in wine. It's usually expressed in *g(sodium chloride)/dm3* in the dataset.
- **Free sulfur dioxide:** This is the part of the sulphur dioxide that when added to a wine is said to be free after the remaining part binds. Winemakers will always try to get the highest proportion of free sulphur to bind. This variable is expressed in *mg/dm3* in the dataset.
- **Total sulfur dioxide:** This is the sum total of the bound and the free sulfur dioxide (*SO2*). Here, it's expressed in *mg/dm3*..
- **Density:** This can be represented as a comparison of the weight of a specific volume of wine to an equivalent volume of water. It is generally used as a measure of the conversion of sugar to alcohol. Here, it's expressed in *g/cm3*.
- **pH:** Also known as the potential of hydrogen, this is a numeric scale to specify the acidity or basicity the wine. Fixed acidity contributes the most towards the pH of wines. You might know, solutions with a pH less than 7 are acidic, while solutions with a pH greater than 7 are basic. With a pH of 7, pure water is neutral. Most wines have a pH between 2.9 and 3.9 and are therefore acidic.
- **sulphates:** These are mineral salts containing sulfur. Sulphates are to wine as gluten is to food. They are a regular part of the winemaking around the world and are considered essential. Here, it's expressed in *g(potassiumsulphate)/dm3* in the dataset.
- **Alcohol:** Wine is an alcoholic beverage. Alcohol is formed as a result of yeast converting sugar during the fermentation process. The percentage of alcohol can vary from wine to wine. Hence it is not a surprise for this attribute to be a part of this dataset. It's usually measured in % vol or alcohol by volume (ABV).
- **Quality:** Wine experts graded the wine quality between 0 (very bad) and 10 (very excellent). The eventual quality score is the median of at least three evaluations made by the same wine experts.
- **Wine_type:** Since we originally had two datasets for red and white wine, we introduced this attribute in the final merged dataset which indicates the type of wine for each data point. A wine can either be a 'red' or a 'white' wine. One of the predictive models we will build would be such that we can predict the type of wine by looking at other wine attributes.
- **Quality_label:** This is a derived attribute from the quality attribute. We bucket or group wine quality scores into three qualitative buckets namely low, medium and high.

Machine Learning

Machine Learning is extensively used in analysing wine type and quality. Machine learning is an application of artificial intelligence(AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Machine learning algorithms are often categorized as supervised, unsupervised and semi supervised algorithms:

- **Supervised machine learning algorithms** can apply what has been learned in the past to new data using labelled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

- In contrast, **unsupervised machine learning algorithms** are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabelled data.

- **Semi-supervised machine learning algorithms** fall somewhere in between supervised and unsupervised learning, since they use both labelled and unlabelled data for training – typically a small amount of labelled data and a large amount of unlabelled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labelled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabelled data generally doesn't require additional resources.

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 3, ISSN No. 2455-2143, Pages 519-529
Published Online July 2019 in IJEAST (http://www.ijeast.com)

II. LITERATURE SURVEY

Literature survey is the important part of report as it gives a direction in the area of research. It helps to set a goal for your analysis-thus giving problem statement. It is also a systematic and thorough search of all types of published literature as well as other sources including dissertation. Following are some of the research papers related to this project.

- P. Appalasamy et al. [1] discussed about modeling the complex human taste is an important focus in wine industries. The main purpose of this study was to predict wine quality based on physicochemical data. This study was also conducted to identify outlier or anomaly in sample wine set in order to detect adulteration of wine. In this project, two large separate datasets are used, which contains 1, 599 instances for red wine and 4, 989 instances for white wine with 11 attributes of physicochemical data such as alcohol, PH and sulphates. Two classification algorithms, Decision tree and Naïve Bayes are applied on the dataset and the performance of these two algorithms is compared. Results showed that Decision tree (ID3) outperformed Naïve Bayesian techniques particularly in red wine, which is the most common type. The study also showed that two attributes, alcohol and volatile-acidity contribute highly to wine quality. White wine is also more sensitive to changes in physico-chemistry as opposed to red wine, hence higher level of handling care is necessary. This research concludes that classification approach will give rooms for corrective measure to be taken in effort to increase the quality of wine during production.
  From the resulting classification accuracy, we found that accuracy rate for the white wine is influenced by a higher number of physicochemistry attribute, which are alcohol, density, free sulphur dioxide, chlorides, citric acid, and volatile acidity. Meanwhile, red wine quality is highly correlated to only four attributes, which are alcohol, sulphates, total sulphur dioxide, and volatile acidity. This shows white wine quality is affected by physicochemistry attributes that does not affect the red wine in general. Therefore, we suggest that white wine manufacturer should conduct wider range of test particularly towards density and chloride content since white wine quality is affected by such substances. Since, white wine is more sensitive to changes in physicochemistry properties as compared to red wine, we suggest a higher level of separation between white wine and red wine production line with particularly further customization to the white wine production. Attribute selection algorithm we conducted also ranked alcohol as the highest in both datasets, hence the alcohol level is the main attribute that determines the quality in both red and white wine. One suggestion is that wine manufacturer to focus in maintaining a suitable alcohol content, may be by longer fermentation period or higher yield fermenting yeast.
- Shen Yin *et al.* [2] evaluated that the quality prediction models are constructed based on multivariate statistical methods, including ordinary least squares regression(OLSR), principal component regression (PCR), partial least squares regression (PLSR), and modified partial least squares regression (MPLSR). The prediction model constructed by MPLSR achieves superior results, compared with the other three methods from both aspects of fitting efficiency and prediction ability. Based on it, further research is dedicated to selecting key variables to directly predict the product quality with satisfactory performance. The prediction models presented are more efficient than tradition ones and can be useful to support human experts in the evaluation and classification of the product quality. The effectiveness of the quality prediction models is finally illustrated and verified based on the practical data set of the red wine.
  Based on a data set of wine qualities and grape physicochemical indexes, wine quality prediction models are constructed with multivariate statistical methods. In order to obtain an efficient wine quality prediction model, a comparison amongst the models established by ordinary least squares regression, principal component regression, partial least squares regression, and a modified partial least square regression is made and the best model is selected out. The calibration set includes 50 grape physicochemical indexes. Dealing with so much data is a time consuming and complex task. With the correlation analysis, it has been found that there exists multicollinearity problems in some grape physicochemical indexes. Under the framework of wine quality prediction model, a suggestion regarding the use of fewer grape physicochemical indexes to predict the wine quality under the promise of prediction accuracy is proposed.
  Compared with the majority of the methods reported for wine quality analysis, the method discussed in this paper provides a simpler and more convenient way to predict the wine quality. Besides, what is most remarkable of the proposed method is the increasing possibility for wine makers to predict the wine quality before the complete wine making process and make appropriate decisions in advance, such asthe grape selection, wine classification, and target marketing. Multivariate statistical analysis is a powerful tool for solving the wine analysis problems which often involve large amounts of data and has been wildly applied in many relevant studies, such as analyzing the elements in wines by PLS regression investigating the relationship between wine composition and vintage via principal component analysis (PCA) and partial least squares (PLS) [13, 14], and utilizing the visible-near infrared spectroscopy and chemometrics to classify Riesling wines from different countries with the help of PCA.
  Four multivariate statistical methods are used: OLSR, PCR, PLSR, and MPLSR. Based on these methods and the real data wine quality prediction models have been

constructed with the superior fitting efficiency and better predicting ability. The efficiency of MPLSR model is essential to be validated on a larger data set. This is not robust wine quality prediction model and robust models to be proposed in the future work.

- Y. Subba Reddy *et al.*in [3] introduced a user-centric similarity framework in which the similarity of products is assessed by user preferences. A popular dataset named "Red wine quality "is considered in this work to assess the quality of Wine by grouping the individual products into clusters and then grade the groups based on preferences. The user centric approach provided quite different and interesting results than the conventional approaches have, that do not consider the preferences that the customers have expressed. It is observed that the query types introduced in this work need higher execution times as the number of users and the preferences increased. This may lead to a scalability problem of the proposed framework. This can be smoothened by introducing R-tree like data structures for searching and indexing purpose that can optimize the execution time of the proposed framework. The purpose of developing this kind of a system is to support and advise wine users for better selection and winemakers for providing a better quality.
  It presents a critical review of research trends on Wine quality and user centric similarity measures as well. A novel user centric similarity measure in product clustering is proposed to evaluate the popular Wine data set named Red Wine dataset. The experimental results obtained in this work are able to provide better recommendations to product buyers than the existing systems. The proposed approach is competent to group the Red wine dataset into ordered groups of preferred wine variants and can judge the wine quality based on these user preference groups.
  The proposed process groups the wine dataset records into priority based clusters. The clustered data using classification form a model to assign the test data records with a recommended voting label. Most of the previous research on wine data limited to normal clustering and classification approaches depending on the taster sensing data whereas, the proposed novel hybrid approach can recommend the user a better wine combination without depending on the taster sensing data.

- Dimitrija Angelkov*et al.* [4] introduced data mining technology for wine analysis. Data mining is an integral, interactive and iterative process of extracting and displaying useful implicitly and innovative knowledge from the data. Data analysis is a consequence of the development of science. This paper made an analysis of data from the measurements in order to find hidden laws and relationships between the data. For correctly determining the quality of wine requires that they have a huge base of measurement. They analyzed 1600 measurements of red wines and 4890 measurements of white wines. From the data analysis used the program Easy Data Mining and produced the following graphs of standard deviation of wine quality. Because of a big number of measurements manually can hardly get to the disclosure of regularities and relationships between the data. Therefore use a computer that contains programs that contain modern tools, algorithms and technique. This paper using a module for monitoring the grape fermentation process constructed of low cost sensors for temperature, wine acidity (pH), alcohol and carbon dioxide released gases. Sensor values are recorded over the wine fermentation process and are sent through wireless modules in real time to a server. Constituent part of the module is a microcontroller PIC16F877A. It processes data received from sensors and sends them to the server through Ethernet controller ENC28J60. The main advantage of this low cost prototype is the possibility to be used by small winemakers for control and monitoring of a grape fermentation process. The proposed system has been tested in a winery in the Tikves region and it fulfilled the initial expectations.

  From analysis of regression tree is done with method 'CHAIN' deep level 5 perception that wine quality will be higher if the level of the degree of acidity pH is below 3.29 for red wines. This is obtained by mechanical analysis from 1599 measurements. Another argument for this is the analysis of red wines where we see regression tree that if the value of the degree of acidity pH below 3.22 then the quality drops because deduce the value of degree of acidity should be in the range of (3.2 - 3.9) to obtain higher quality wines. Models of classification are used like system who gives as decision support on different wine production stages. This gives opportunity of manufactory for production of wine with high quality.

- Yesim Er*et al.* [5] explains that this study is to predict wine quality based on physicochemical data. Two large separate data sets which were taken from UC Irvine Machine Learning Repository were used. These data sets contain 1599 instances for red wine and 4898 instances for white wine with 11 features of physicochemical data such as alcohol, chlorides, density, total sulphur dioxide, free sulphur dioxide, residual sugar, and pH. First, the instances were successfully classified as red wine and white wine with the accuracy of 99.5229% by using Random Forests Algorithm. Then, the following three different data mining algorithms were used to classify the quality of both red wine and white wine: k-nearest-neighbourhood, random forests and support vector machines. There are 6 quality classes of red wine and 7 quality classes of white wine. The most successful classification was obtained by using Random Forests Algorithm.
  It is also observed that the use of principal component analysis in the feature selection increases the success

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 3, ISSN No. 2455-2143, Pages 519-529
Published Online July 2019 in IJEAST (http://www.ijeast.com)

rate of classification in Random Forests Algorithm. For each classification model, we analysed how the results vary whenever test mode is changed. The study includes the analysis of classifiers on both red and white wine data set. The results are described in percentage of correctly classified instances, precision, recall, F measure, and ROC after applying the cross-validation or percentage split mode. Three classifiers are used: k-nearest-neighborhood, random forests, and support vector machines are evaluated on datasets. Results from the experiments lead us to conclude that Random Forests Algorithm performs better in classification task as compared against the support vector machine, and k-nearest neighbourhood. After applying PCA, the success rate of quality classification for white wine has decreased from 70.3757% to 69.9061% for cross validation mode. The success rate of quality classification for white wine has decreased from 68.6735% to 67.449% for percentage split mode.
However in the previous study the researchers always focus on the subjective study to define the quality of wine. The result based on the subjective study takes much time as well as it is not effective compared to the objective study with the analytical methods. In the past few works related to wine data has been studied using different classifiers, however so far nobody has compared the performance metrics of the different classifiers with different feature sets to predict the quality of different type of wine by considering several factors.
The present work new approach has been proposed by considering the use of wine dataset for all the experiments. Wine dataset is a collection of white and red wines. White wine consists of 4898 samples and red wine contains 1599 samples. Each sample of both types of wine consists of 12 physiochemical variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol, and quality rating. The quality rating is based on a sensory test carried out by at least three sommeliers and scaled in 11 quality classes from 0 - very bad to 10 - very excellent. It is not possible to use both type of wine collections without pre-processing due to some deficiencies. One of the major deficiencies is the large amplitude of variable values e.g. sulphates (0.3–2) vs. sulphur dioxide (1–72).This analysis will help the wine experts to know the important factors to consider while selecting the good quality wine.

III. SYSTEM ARCHITECTURE

Very few systems use the available wine data for prediction purposes and even if they do, they are restricted by large number of association rules that apply. Disadvantages of using such systems are,

- Detection is not possible at an earlier stage.
- In the existing system, practical use of various collected data is time consuming.
- This practice leads to errors and less accuracy rate.

The present model allows to predict the wine type and quality. The model is fed with various physiochemical properties of wine. The main goal of this system is to analyse the wine type and quality using machine learning algorithm such as Logistic regression, decision tree andRandom forest. Wine data set is used and then pre-processed and transformed the data set. Then apply the decision tree algorithm on the transformed data set. After applying the algorithm, wine type and quality is analysed and then the result obtained based on the prediction of whether the wine quality is low, high or medium.

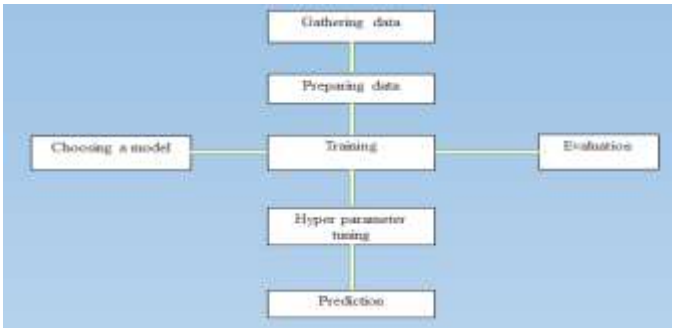

Fig1: System Architecture

Fig 3.1 shows the steps in analysis of wine which is discussed below,

1. Gathering data
2. Preparing that data
3. Choosing a model
4. Training
5. Evaluation
6. Hyper parameter tuning
7. Prediction.

1. Gathering Data:

This step is very crucial as the quality and quantity of data gathered will directly determine how good the predictive model will turn out to be. The data collected is then tabulated and called as Training Data or Data Set. The attribute values which determine the quality of and type of wine are given as data.

2. Data Preparation:

After the training data is gathered, next step is Data preparation, where the data is loaded into a suitable place and then prepared for use. Also, the data has to be split into two parts. The first part that is used in training our model, will be the majority of the dataset and the second will be used for the evaluation of the trained model's performance.

3. Choosing a model:

The next step that follows in the workflow is choosing a model.Choosing a model that is best suitable for analysis of wines. The model includes a Decision tree. It is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.
They provide a highly effective structure within which you can lay out options and investigate the possible outcomes of

choosing those options. They also help you to form a balanced picture of the risks and rewards associated with each possible course of action.

4. Training:

The training process involves initializing some random values for model, predict the output with those values, then compare it with the model's prediction and then adjust the values so that they match the predictions that were made previously.This process then repeats and each cycle of updating is called one training step.

5. Evaluation:

Evaluation allows the testing of the model against data that has never been seen and used for training. It is the representative of how the model will analysis in the real time.

6. Parameter Tuning:

Once the evaluation is over, any further improvement in training can be possible by tuning the parameters. There were a few parameters that were implicitly assumed when the training was done. Another parameter included is the learning rate that defines how far the line is shifted during each step, based on the information from the previous training step. These values all play a role in the accuracy of the training model, and how long the training will take.

7.Prediction: This is the point where the value of machine learning is realized. Previously predicted data will give the output.

IV. DESIGN AND IMPLEMENTATION

The redwine and whitewine data-frames and data setsare used here which are necessary for basic exploratory analysis and visualizations. The gathered data set will be used for data analysis and modellingthrough which output can be predicted. Data of different attributes and properties is considered for the analysis. The analysis of the data includes exploratory and predictive analysis.

- **Exploratory data analysis:** Standard Machine Learning and analytics workflow recommend processing, cleaning, analysing, and visualizing your data before moving on toward modelling data. The required packages and necessary dependencies for analysis can be imported by using these code.

```
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib as mpl
import numpy as np
import seaborn as sns
%matplotlib inline
```

The datasets (red and white wine) and add some additional variables are used that is required to predict in future sections. The first variable to be added is wine_type, which would be either red or white wine based on the dataset and the wine sample. The second variable to be added is quality_labelwhich is a qualitative measure of the quality of the wine samplebased on the qualityvariable score. The rules used formapping quality to quality_label are described as follows.

  - Wine quality scores of 3, 4, and 5 are mapped to low quality wines under the quality_label attribute.
  - Wine quality scores of 8 and 9 are mapped to high quality wines under the quality_label attribute.
- **Descriptive statistics:**Some descriptive statistics of various features of interest in the dataset should be computed. This involves computing aggregation metrics like mean, median, standard deviation, and so on. The primary objectives is to build a model that can correctly predict if a wine is a red or white wine based on its attributes.
- **Inferential Statistics:** The general notion of inferential statistics is to draw inferences and propositions of a population using a data sample. The idea is to use statistical methods and models to draw statistical inferences from a given hypotheses. Each hypothesis consists of a null hypothesis and an alternative hypothesis. Based on statistical test results, if the result is statistically significant based on pre-set significance levels (e.g., if obtained p-value is less than 5% significance level), the null hypothesis is rejected in favor of the alternative hypothesis. Otherwise, if the results is not statistically significant, concluded as that the null hypothesis was correct.
- **Univariate analysis:** Is the simplest form of analysing data. “Uni” means “one”, so in other words the data has only one variable. It doesn’t deal with causes or relationships and its major purpose is to describe: it takes data, summarizes that data and finds patterns in the data. A variable in univariate analysis is just a condition or subset that the data falls into. It can be considered as a category. For example the analysis might look at a variable age or it might look at height or weight.
  Univariate descriptive statistics: some ways patterns found in univariate data include central tendency (mean, mode and median) and dispersion can be described: range, variance, maximum, minimum, quartiles, and standard deviation. Several options are there for describing data.
- **Multivariate analysis:**Is used to study more complex sets of data than what univariate analysis methods can handle. There are more than 20 different ways to perform multivariate analysis. This type of analysis is almost always performed with software as working with even the smallest of data sets can be overwhelming by hand.

Predicting Wine Types

In the wine quality dataset, there are two variants or types of wine—red and white wine. The main task of classification system is to predict the wine type based on other features. To start with, necessary features should be selected and

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 3, ISSN No. 2455-2143, Pages 519-529
Published Online July 2019 in IJEAST (http://www.ijeast.com)

separate the prediction class labels and prepare train and test datasets. Prefix wtp_ is used in the variables to easily identify them as needed, where wtp depicts wine type prediction. The model is designed and implemented using the logistic regression algorithm.

Logistic Regression is one of the most used Machine Learning algorithms for binary classification. It is a simple Algorithm and use as a performance baseline, it is easy to implement and it will do well enough in many tasks. The building block concepts of Logistic Regression can also be helpful in deep learning while building neural networks. Here it is used to predict the wine type whether it is red or white. There are three types of logistic regression algorithm: Binary, Multi and ordinal. The flow chart of this algorithm can be return as

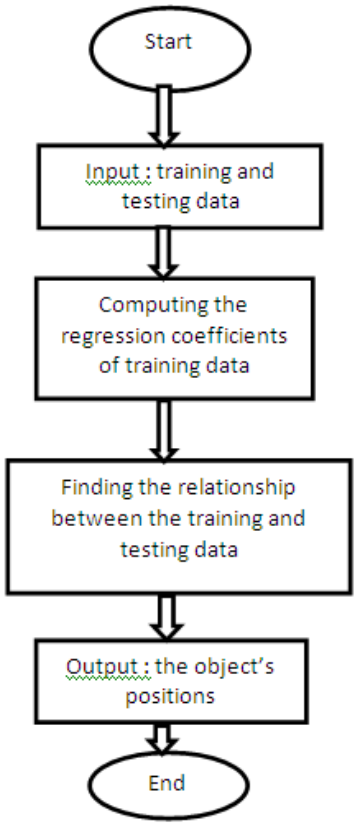

Fig 2: Flow chart of logistic regression

Predicting Wine Quality

In the wine quality dataset, there are several quality rating classes ranging from 3 to 9. Quality_label variable is focused that classifies wine into low, medium, and high ratings based on the underlying quality variable based on the mapping created in the "Exploratory Data Analysis". This is done because several rating scores have very few wine samples and hence similar quality ratings were clubbed together into one quality class rating. Prefix wqp_is used for all variables and models involved in prediction of wine quality to distinguish it from other analysis. The prefix wqp stands for wine quality prediction .The following algorithms are used for classifying quality.

- **Decision tree algorithm**

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables.

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification **problems** too. Here this algorithm is used to analyse the quality of wine.

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

A decision tree consists of three types of nodes:

1. Decision nodes – The node where a there is a requirement set that determines the outcome, typically represented by squares.
2. Chance nodes – A node were there isn't a set requirement to determine where to go in the split but something that just has a probability of happening or not Ex: 50% chance of success or failure of a business ,typically represented by circles
3. End nodes – The end of a split, so something feeds into this but nothing comes out of it. Usually some result, ex: Business is successful , typically represented by triangles

Algorithm :
INPUT : S, where S= set of classified instances
OUTPUT : Decision tree
Require : S=0, num_attributes>0
1: **procedure** BUILDTREE
2: **repeat**
3: maxGain← 0
4: splitA← null
5: e← Entropy(Attributes)
6: for all Attributes a in S do
7: gain←InformationGain(a,e)
8: if gain>maxGain then
9: maxGain← gain
10: splitA← a
11: end if
12: end for
13: Partition(S,splitA)
14: until all partitions processed
15: end procedure

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 3, ISSN No. 2455-2143, Pages 519-529
Published Online July 2019 in IJEAST (http://www.ijeast.com)

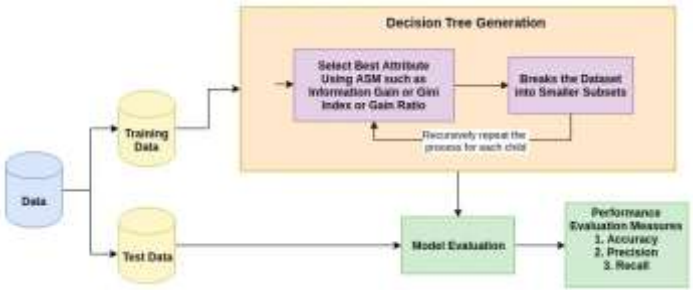

Fig 3: Workflow of decision tree algorithm

In this project decision tree algorithm is used to classify the wine data into its different quality and types. The leaf node will have the attributes with its value and represents its quality. One can easily understand by looking into the generated decision tree.

- **Random Forest Algorithm**

Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach.

The difference between Random Forest algorithm and the decision tree algorithm is that in Random Forest, the processes of finding the root node and splitting the feature nodes will run randomly. Some applications of using Random Forest algorithm: Banking, Medicine, Stock Market and E-commerce:

- For the application in banking, Random Forest algorithm is used to find loyal customers, which means customers who can take out plenty of loans and pay interest to the bank properly, and fraud customers, which means customers who have bad records like failure to pay back a loan on time or have dangerous actions.
- For the application in medicine, Random Forest algorithm can be used to both identify the correct combination of components in medicine, and to identify diseases by analyzing the patient's medical records.
- For the application in the stock market, Random Forest algorithm can be used to identify a stock's behavior and the expected loss or profit.
- For the application in e-commerce, Random Forest algorithm can be used for predicting whether the customer will like the recommend products, based on the experience of similar customers.

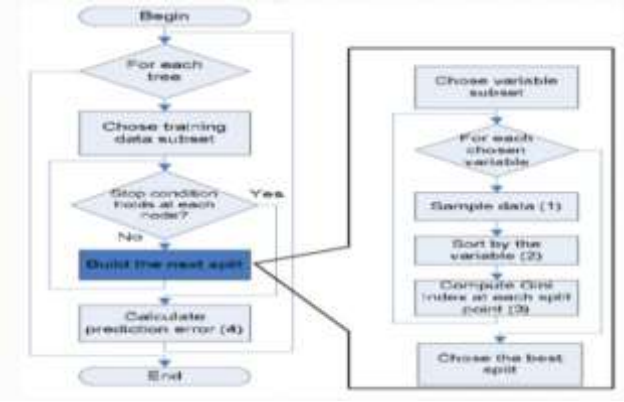

Fig 4: Flowchart of random forest algorithm

V. RESULTS AND SNAPSHOTS

Table 1: List of attributes and values with its types and quality

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density
0	7.0	0.17	0.74	12.8	0.045	24.0	126.0	0.99420
1	7.7	0.64	0.21	2.2	0.077	32.0	133.0	0.99560
2	6.8	0.39	0.34	7.4	0.020	38.0	133.0	0.99212
3	6.3	0.28	0.47	11.2	0.040	61.0	183.0	0.99592
4	7.4	0.35	0.20	13.9	0.054	63.0	229.0	0.99888

Table 1 consists of the result of loading and merging the two datasets. Based on the physicochemical properties type of wine and quality will be predicted. It gives both type and quality of wine.

Table 2 List of attributes of wine types with statistical values

	Red Wine Statistics						White Wine Statistics	
	residual sugar	total sulfur dioxide	sulphates	alcohol	volatile acidity	quality	residual sugar	total dioxi
count	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	4898.00	
mean	2.54	46.47	0.66	10.42	0.53	5.64	6.39	
std	1.41	32.90	0.17	1.07	0.18	0.81	5.07	
min	0.90	6.00	0.33	8.40	0.12	3.00	0.60	
25%	1.90	22.00	0.55	9.50	0.39	5.00	1.70	
50%	2.20	38.00	0.62	10.20	0.52	6.00	5.20	
75%	2.60	62.00	0.73	11.10	0.64	6.00	9.90	
max	15.50	289.00	2.00	14.90	1.58	8.00	65.80	

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 3, ISSN No. 2455-2143, Pages 519-529
Published Online July 2019 in IJEAST (http://www.ijeast.com)

Table 3 List of attributes of wine quality with statistical values

	Low Quality Wine				Medium Quality Wine			
	alcohol	volatile acidity	pH	quality	alcohol	volatile acidity	pH	
count	2384.00	2384.00	2384.00	2384.00	3915.00	3915.00	3915.00	
mean	9.87	0.40	3.21	4.88	10.81	0.31	3.22	
std	0.84	0.19	0.16	0.36	1.20	0.14	0.16	
min	8.00	0.10	2.74	3.00	8.40	0.08	2.72	6.00
25%	9.30	0.26	3.11	5.00	9.80	0.21	3.11	6.00
50%	9.60	0.34	3.20	5.00	10.80	0.27	3.21	6.00
75%	10.40	0.50	3.31	5.00	11.70	0.36		
max	14.90	1.58	3.90	5.00	14.20	1.04		

Table .2 and .3 gives the statistical values of some of the attributes values from which its type and quality are determined. Table 5.2 has values for attributes such as residual sugar, total sulfur dioxide, sulphates and volatile acidity with the quality values. This analysis is made for two types of wine that is red and white. Table 5.3 has values for attributes such as alcohol, volatile acidity, ph. and quality values. Analysis is made for quality types that is low, medium and high.

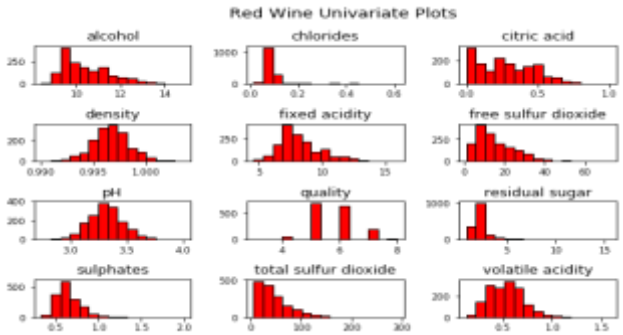

Fig 5: Red Wine Univariate Plots

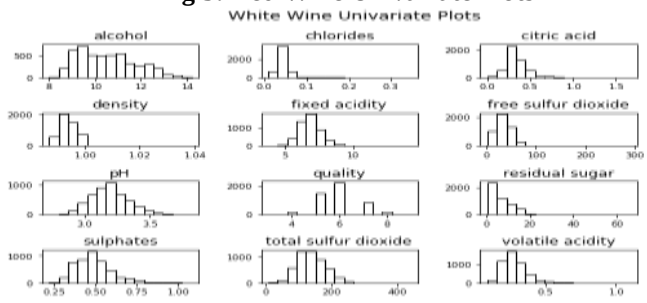

Fig 6: White Wine Univariate Plots

Univariate plots depicting feature distributions for the wine quality dataset. The power of packages like matplotlib and pandas enable to easily plot variable distributions as depicted in Figure 3.2 and 3.3 using minimal code.

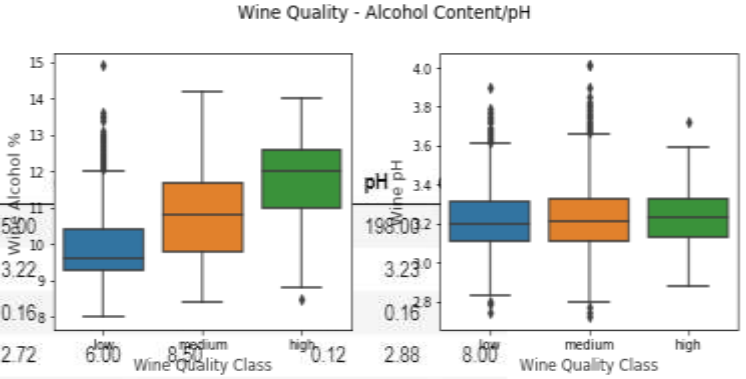

Fig 7: Wine Alcohol/PH content

Visualizing wine alcohol content and pH level distributions based on quality ratings

The boxplots depicted in Figure5.3 show the stark differences in wine alcohol content distributions based on wine quality as compared to pH levels, which look to be between 3.1 - 3.3 and in fact if you look at the mean and median values for pH levels across the three groups, it is approximately 3.2 across the three groups as compared to alcohol %, which varies significantly.

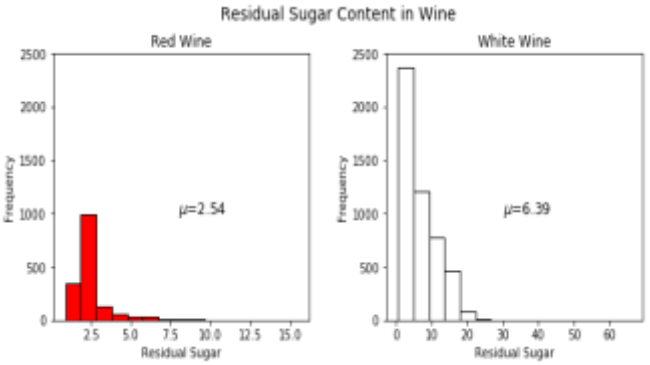

Fig 8: Residual sugar contents

Fig 8 depicts Residual sugar distribution for red and white wine samples, residual sugar content in white wine samples seems to be more as compared to red wine samples.

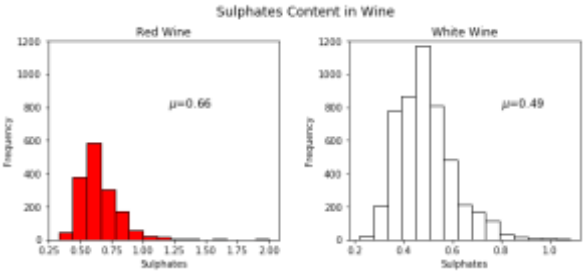

Fig 9: Sulphates content

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 3, ISSN No. 2455-2143, Pages 519-529
Published Online July 2019 in IJEAST (http://www.ijeast.com)

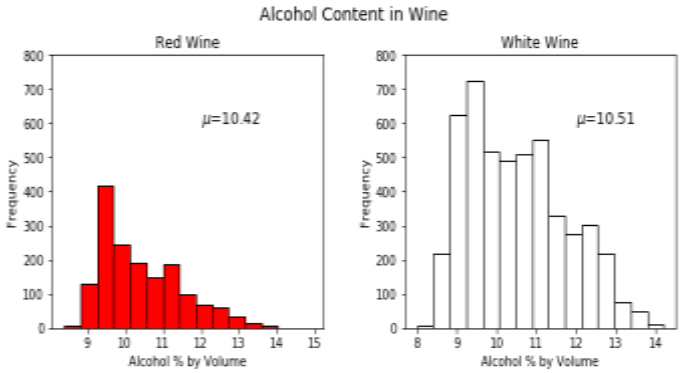

Fig 10: Alcohol content

Fig 9 and 10 depicts the distributions for sulphate content and alcohol content for red and white wine samples. The plots depicted shows that the sulphate content is slightly more in red wine samples as compared to white wine samples and alcohol content is almost similar in both types on an average.

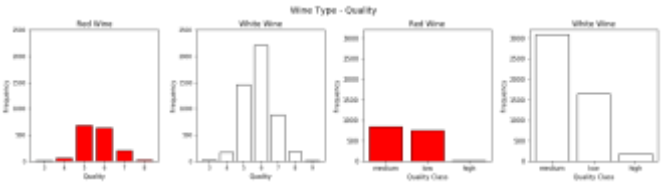

Fig 11: Wine types and quality

The bar plots in fig 11 depicted shows the distribution of wine samples based on type and quality. It is quite evident that high quality wine samples are far less as compared to low and medium quality wine samples.

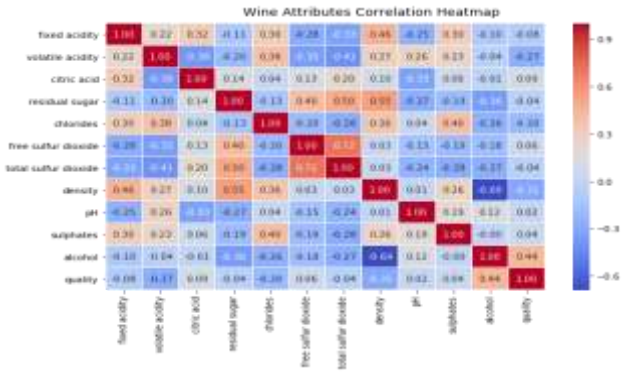

Fig 12: Wine Attributes Correlation Heatmap

Fig 12 represents a heat map. A heat map is a two dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colors. The seaborn python package allows the creation of annotated heatmaps which can be tweaked using matplotlib tools as per the creators requirement.

```
Model Performance metrics:
--------------------------------
Accuracy: 0.9923
Precision: 0.9923
Recall: 0.9923
F1 Score: 0.9923

Model Classification report:
--------------------------------
             precision    recall  f1-score   support

        red       0.98      0.99      0.98       470
      white       1.00      0.99      0.99      1480

avg / total       0.99      0.99      0.99      1950

Prediction Confusion Matrix:
--------------------------------
                Predicted:
                      red white
Actual: red           463     7
        white           8  1472
```

Fig 13: Performance metrics, Classification report, Prediction Confusion Matrix

Fig 13 gives the performance metrics, classification report and prediction confusion matrix of the wine data using Logistic Regression for both red and white wines. To analyse wine types using Logistic regression algorithm is used.

```
Model Performance metrics:
--------------------------------
Accuracy: 0.7282
Precision: 0.7283
Recall: 0.7282
F1 Score: 0.7282

Model Classification report:
--------------------------------
             precision    recall  f1-score   support

        low       0.68      0.68      0.68       718
     medium       0.78      0.78      0.78      1178
       high       0.30      0.30      0.30        54

avg / total       0.73      0.73      0.73      1950

Prediction Confusion Matrix:
--------------------------------
                Predicted:
                      low medium high
Actual: low           488    225    5
        medium        229    916   33
        high            2     36   16
```

Fig 14: Performance metrics, Classification report, Prediction Confusion Matrix

Fig 14 gives the performance metrics, classification report and prediction confusion matrix of the wine data using decision tree classifier. Which classifies the quality of the wine based on the properties and give output as high, low and medium.

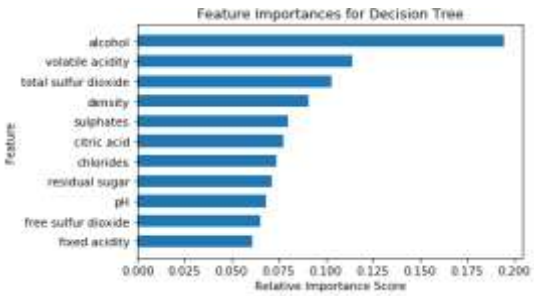

Fig 15: Feature importance from decision tree model

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 3, ISSN No. 2455-2143, Pages 519-529
Published Online July 2019 in IJEAST (http://www.ijeast.com)

Fig 15 depicts different attributes or features and their relative importance score. The quality of wine are determined by scores of the features and their importance. The model is trained, predicted and evaluated before this.

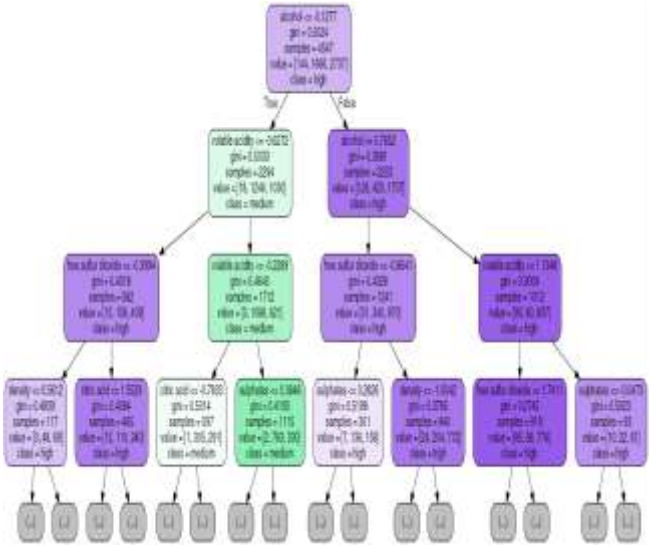

Fig 16: Decision tree structure for wine analysis

Figure 16 where the starting split is determined by the rule of alcohol <= -0.1277 and with each yes\no decision branch split, we have further decision nodes as we descend into the tree at each depth level. The class variable is what we are trying to predict, i.e. wine quality being low, medium, or high and value determines the total number of samples at each class present in the current decision node at each instance. The gini parameter is basically the criterion which is used to determine and measure the quality of the split at each decision node. Best splits can be determined by metrics like gini impurity\gini index or information gain. Just to give you some context, the gini impurity is a metric that helps in minimizing the probability of misclassification.

```
Model Performance metrics:
------------------------------
Accuracy: 0.7815
Precision: 0.7809
Recall: 0.7815
F1 Score: 0.7789

Model Classification report:
------------------------------
             precision    recall  f1-score   support

        low       0.73      0.74      0.73       718
     medium       0.81      0.83      0.82      1178
       high       0.75      0.33      0.46        54

avg / total       0.78      0.78      0.78      1950

Prediction Confusion Matrix:
------------------------------
                Predicted:
                     low medium high
Actual: low          528    189    1
        medium       195    978    5
        high           2     34   18
```

Fig 17: Performance metrics, Classification report, Prediction Confusion Matrix

Fig 17 gives the performance metrics, classification report and prediction confusion matrix of the wine data using random forest classifier which is also used to classify the quality of the wine.

VI. CONCLUSION

The specific objective of this study is to analyse how physicochemical properties like alcohol percentage, chlorides, sulphates contents etc., varies the quality of wine. This study analyse the wine types and quality with the various physicochemical variables. Two datasets were created, using red and white wine samples. Out of thirteen attributes, the statically significant attribute that influence the quality of wine is an essential finding. The model that highlights the significant attribute in both sets. This result helpful in production and in quality prediction by studying those attributes. Analyse the wine type using logistic regression and quality by three machine learning algorithms such as decision tree, random forest and extreme gradient boosting. The results obtained are more accurate than previous techniques.

VII. REFERENCES

1. P.Appalasamy and A. Mustapha,(2012) "Classification-based Data Mining Approach for Quality Control in Wine Production", Volume 12 (6), Journal of Applied Sciences.
2. Shen Yin,(2013) "Research Article: Quality Evaluation Based on Multivariate Statistical Methods", Article ID 639652, 10 pages.
3. DimitrijaAngelkov,(2016) "Data mining of wine quality", Volume 45, August 31.
4. Y. Subba Reddy and Prof. P. Govindarajulu,(2017),"An Efficient User Centric Clustering Approach for Product Recommendation Based on Majority Voting: A Case Study on Wine Data Set", IJCSNS International Journal of Computer Science and Network Security, VOL.17 No.10.
5. Bernad Chen,(2016) "Wine Informatics-Association rule based classification", 11th December 2016. Miami, FL, USA.
6. Yesim Er and AytenAtasoy,(2016) "The Classification of White Wine and Red Wine According to Their Physicochemical Qualities".
7. Zhang Lingfeng and Feng Feng,(2017) "Wine quality identification based on data mining research",12th International Conference on Computer Science and Education (ICCSE).
8. Nao Wariishi and Brendan Flanagan,(2015) "Sentiment Analysis of Wine Aroma", 4th International Congress on Advanced Applied Informatics.
9. Yogesh Gupta,(2017) "Selection of important features and predicting wine quality using machine learning techniques", Procedia Computer Science, vol.125.
10. AytenAtasoy,(2017) "Classification of white wine and red wine according to their Physicochemical Qualities".