ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 4, DOI:10.29027/IJIRASE.v4.i4.2020.715-721, October 2020

Wine Quality and Taste Classification Using Machine Learning Model

Anurag Sinha[1], Atul Kumar[2]
[1,2]Department of computer science and IT, Student, Amity University Jharkhand
Ranchi, Jharkhand(India), 834001
Email:anuragsinha257@gmail.com[1], atulkt008@gmail.com[2]

Abstract- In recent years the product quality has been one of the crucial parts for every single industry. The conventional methods for assessing the product quality are very time consuming and also not having the optimal result however with the resultant in dynamic Technology movement. We have the concepts of machine learning and data science through this technique it's become more efficient to assess or to predict any kind of thing efficiently. In this paper, we have explored several machine learning techniques for evaluating wine quality based on different metrics and properties related to wine quality. In this paper, we have also used several machine learning algorithms to rank the quality of wines and investigate why people make the wine taste more interesting. We have selected the features using the most popular machine learning techniques. We used different types of datasets for this particular study.

Keywords— Data science, machine learning, wine dataset, Logistic Regression, Stochastic gradient descent, Support Vector Classifier, Random Forest

I. INTRODUCTION

In recent years the data science has taken great popularity in the field of computer science with the rise of data generated throughout the internet the domain of data science is getting so much popularity. It's one of the most demanding fields of the 21st century. Every single industry is to try and adapting its workflow. Data science follows the data-driven approach which helps the organization or different company to understand the need of their customer more effectively. we are living in the age of the information revolution where data is become becoming one of the most precious parts of Information Technology. Different companies collect different types of data based on their business approach

example for a retail store the data could be the kind of a product they are using and they and their selling the products to their customer. Similarly for Netflix, it can be their customer's interest in watching. Purpose of data science, the purpose of deciding is to get the most recent from the data that could go undetected, and how they are expecting the features from the customer decisions and their reviews and employing the Mining in sentiment analysis for effective business decisions.

As the demand for wine has increased in recent years, wine consumption has also increased. With increasing demand, the wine industry is looking for alternatives to quality wines with minimal cost. Different types of wine are produced depending on the conditions and purpose. Since the chemical composition of most wines are similar to each other and have different concentration levels than other types of wines, it has become more important over the years to classify different types of wines for quality assurance. Although it has not been possible to classify efficiently based on wind properties due to a lack of technology in the industry, the invention of a machine learning approach can now extract feature selections and classify wines based on characteristics. This is because it is also possible to grasp the importance of the chemical analysis parameters in wines, which can be neglected due to cost-effectiveness. Additionally, the structure of this white paper focuses on analyzing why the factors that make wine good for customers are by analyzing these other parameters.

II. LITERATURE SURVEY

In the past few years, several attempts have been chosen to make it more beneficial by using machine learning approaches and feature selection techniques on wine dataset. Er, and atasoy[1,2,3] proposed an approach based on support vector machine classifier to classify the quality of a wine It uses three classification techniques. Chen et al. [4, 5, 6] Based on customer reviews, we proposed an algorithm to predict the grade or condition of the vine. The third of this proposed feature selection was based on a hierarchical clustering approach and association rule mining. Applasammy [8] proposed an algorithm to predict wine quality during wine production in warehouses. Beltran [9] proposed an algorithm to classify wine data sets based on genetic algorithms and aroma chromatograms. Thakkar [10, 11] proposed an algorithm based on the analytical layer clustering procedure to classify wines according to their classification and properties. [12] Proposed an algorithm that uses a support vector machine and a random forest prediction model based on a central clustering method to recommend the best wine products to customers. For this study, they used datasets on white and red wines.

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 4, DOI:10.29027/IJIRASE.v4.i4.2020.715-721, October 2020

III. DATA COLLECTION

This wine data set is publicly available for access in the database of UCI. it is an open-source data provided by the ***Kaggle*** community which is an active participation part of the data science community. this data set has been collected from the UCI machine learning repository having two wine data sets. In which the one data set contains information about the red wines and another contains the information about the white wines.

IV. Classification of Machine Learning Algorithm

Machine learning classification is shown below:-

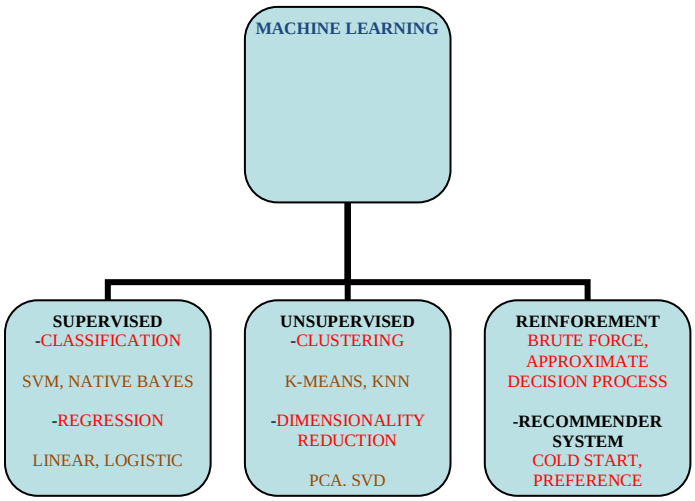


FIGURE 1: Machine Learning Algorithm

a. Supervised learning: in this type of learning system the datasets are labelled and planned. if data is not labelled then well-equipped dataset can be used.

b. Semi-supervised learning: Here, learning is based on a set of imprecise in a sequence which is generally underprovided.

c. Reinforcement learning: The best example of driving a vehicle.

c. Active learning: For this learning type, a PC can obtain for a confined diversion sketch of cases. accurately when used impulsively, this information can be shown to the buyer.

d.Unsupervised learning: No pathway, labelling or classification are given to the knowledge calculation. It isolates the information to establish the collection in its data.

V. METHODOLOGY

5.1 Assessing the taste feature of wine

we have imported some libraries from the data analysis some library of the Python programming is given below:-

```
# Import libraries necessary for this project
import numpy as np
import pandas as pd
from time import time
from IPython.display import display # Allows the use of display() for displaying DataFrames

import matplotlib.pyplot as plt
import seaborn as sns

# Import supplementary visualization code visuals.py from project root folder
import visuals as vs

# Pretty display for notebooks
%matplotlib inline
```

then we load our data set Into The Notebook of jupyter notebook and type the same code in the notebook and Run and have the output. in this table we can see there are 12 different features of wine is given and the last column is given with the quality of a metric and attributes which specify the wine ranking between 1 to 10 the output of the following is given below:-

it will print this output -

```
# Load the Red Wines dataset
data = pd.read_csv("data/winequality-red.csv", sep=';')

# Display the first five records
display(data.head(n=5))
```

Table 1: 12 different features of wine data set

Fixed acidity (g(tartaric acid)/dm^3)	4.600	15.90	8.320	1.741
Volatile acidity (g(acetic acid)/dm^3)	0.120	1.580	0.528	0.179
Citric acid (g/dm^3)	0.000	1.000	0.271	0.195
Residual sugar (g/dm^3)	0.900	15.50	2.539	1.410
Chlorides (g(sodium chloride)/dm^3)	0.012	0.611	0.087	0.047
Free sulfur dioxide (mg/dm^3)	1.000	72.00	15.87	10.46
Total sulfur dioxide (mg/dm^3)	6.000	289.0	46.47	32.89
Density (g/cm^3)	0.990	1.004	0.997	0.002
pH	2.740	4.010	3.311	0.154
Sulphates(g(potassium sulphate)/dm^3)	0.330	2.000	0.658	0.170
Alcohol (%vol)	8.400	14.90	10.42	1.066

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 4, DOI:10.29027/IJIRASE.v4.i4.2020.715-721, October 2020

data.isnull().any()

Let's see if there is any missing information in these

columns. In the cellblock, type:
This output shows that there are no empty columns.

```
data.isnull().any()
```

```
fixed acidity           False
volatile acidity        False
citric acid             False
residual sugar          False
chlorides               False
free sulfur dioxide     False
total sulfur dioxide    False
density                 False
pH                      False
sulphates               False
alcohol                 False
quality                 False
dtype: bool
```

We can acquire several supplementary added information on our data-set by running:

data.info()

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
fixed acidity           1599 non-null float64
volatile acidity        1599 non-null float64
citric acid             1599 non-null float64
residual sugar          1599 non-null float64
chlorides               1599 non-null float64
free sulfur dioxide     1599 non-null float64
total sulfur dioxide    1599 non-null float64
density                 1599 non-null float64
pH                      1599 non-null float64
sulphates               1599 non-null float64
alcohol                 1599 non-null float64
quality                 1599 non-null int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

Figure 2: The output of extra added information

Now we start examining the initial analysis of the wine taste classification. this method is considered all wines with

rating 7. we are measuring the good quality wine. for every quality wines where assigning the rating 6 and for the lesser than that and we are assigning it 5.

```
n_wines = data.shape[0]

# Number of wines with quality rating above 6
quality_above_6 = data.loc[(data['quality'] > 6)]
n_above_6 = quality_above_6.shape[0]

# Number of wines with quality rating below 5
quality_below_5 = data.loc[(data['quality'] < 5)]
n_below_5 = quality_below_5.shape[0]

# Number of wines with quality rating between 5 to 6
quality_between_5 = data.loc[(data['quality'] >= 5) & (data['quality'] <= 6)]
n_between_5 = quality_between_5.shape[0]

# Percentage of wines with quality rating above 6
greater_percent = n_above_6*100/n_wines

# Print the results
print("Total number of wine data: {}".format(n_wines))
print("Wines with rating 7 and above: {}".format(n_above_6))
print("Wines with rating less than 5: {}".format(n_below_5))
print("Wines with rating 5 and 6: {}".format(n_between_5))
print("Percentage of wines with quality 7 and above: {:.2f}%".format(greater_percent))

# Some more additional data analysis
display(np.round(data.describe()))
```

```
Total number of wine data: 1599
Wines with rating 7 and above: 217
Wines with rating less than 5: 63
Wines with rating 5 and 6: 1319
Percentage of wines with quality 7 and above: 13.57%
```

Quality distribution using graph visualization

Visualize skewed continuous features of original data

vs.distribution(data, "quality")

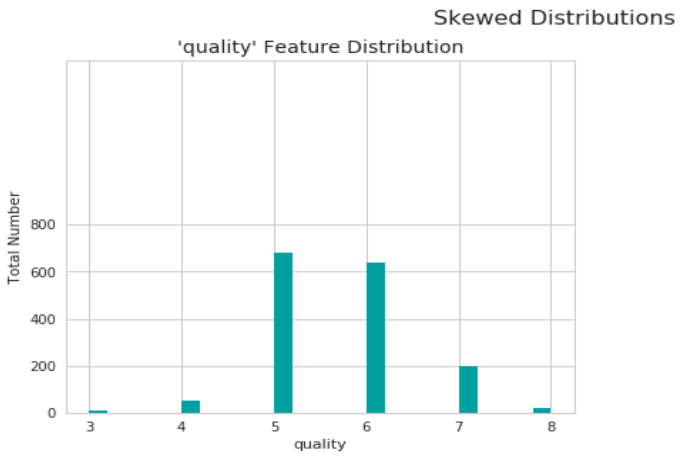


Figure 3: feature distribution

In the graph given, the quality of the wine is classified as 5 6 and 7. There are fuel wines which are given wines of good quality, good taste and not very good. Describes how to get useful statistical operations using Pandas.

Some of the statistics used in the analysis relate to the median mean, mode, range, and standard deviation.

The next step is to extract the characteristics of the story in more detail from the dataset. As we all know, the quality of a wine depends on several properties, which are the chemical composition and the chemical properties of which the wine is composed. The quality of the wine depends on the ampoule of various chemical properties that affect the test and taste. It can be said that we are using this chemical property for feature selection and feature extraction.

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 4, DOI:10.29027/IJIRASE.v4.i4.2020.715-721, October 2020

Terminologies

Wine contains vitamin cells and other types of properties such as sugar alcohol organic acid cells made up of volatile aromatic compounds such as sulfur dioxide and other pigments. In the field of wind testing, the term acidity defines the freshness of wine for its properties. Basic wines free from tartaric acid, malic acid and citric acid, this property defines the balance between acidity.

The sweet and bitter components of wine, the properties of this functional selection, are defined below.

• **Fixed Acidity** - The titrated acidity X is called the fixed acidity, which is the total measure of hydrogen ions present in wine, regardless of the total concentration of titrated acid.

• **Volatile acidity**: The bacteria in wine mainly produce acetic acid, an acid that gives vinegar its distinct flavour and aroma.

• **Citric Acid**: found every minute in wine grapes, it acts as a preservative and is added to wine to increase acidity and complement certain flavours.

• **Residual sugar** -R refers to natural glucose which stops after fermentation due to insufficient residual sugar.

• **Density**: also called specific gravity, it can be used to determine the alcohol content of the wine. In the slicing process, the sugar is branched into juice and converted to ethanol with carbon dioxide.

• • **Ph**- indicates the power of hydrogen, which is the total measure of the concentration of hydrogen ions in a solution. In general, if the pH is below 7, it is considered a date with the strongest acid close to a solution greater than 7 and is considered alkaline or basic.

We have now seen the relationship between functionality and display. From a dataset that contains many characteristics such as alcohol level, residual sugar and PH value. Most of these features are independent of each other and some can even affect the quality range. For example, the

PH value affects the acidity level, the volatile acidity level is related to quality, and those with high alcohol content wines taste better and have better quality. Various visualizations were shown using a Python library suitable for curve fitting on a set of trained data models. I plotted a scatter plot with one dataset and looked at the results.

Visualization is given below:-

pd.plotting.scatter_matrix(data, alpha = 0.3, figsize = (40,40), diagonal = 'kde');

Table 2: The physicochemical data statistics of white wine

Fixed acidity (g(tartaric acid)/dm^3)	3.800	14.20	6.855	0.844
Volatile acidity (g(acetic acid)/dm^3)	0.080	1.100	0.278	0.101
Citric acid (g/dm^3)	0.000	1.660	0.334	0.121
Residual sugar (g/dm^3)	0.600	65.80	6.391	5.072
Chlorides (g(sodium chloride)/dm^3)	0.009	0.346	0.046	0.022
Free sulfur dioxide (mg/dm^3)	2.000	289.0	35.31	17.01
Total sulfur dioxide (mg/dm^3)	9.000	440.0	138.4	42.50
Density (g/cm^3)	0.987	1.039	0.994	0.003
pH	2.720	3.820	3.188	0.151
Sulphates (g(potassium sulphate)/dm^3)	0.220	1.080	0.490	0.114
Alcohol (%vol)	8.000	14.20	10.51	1.231

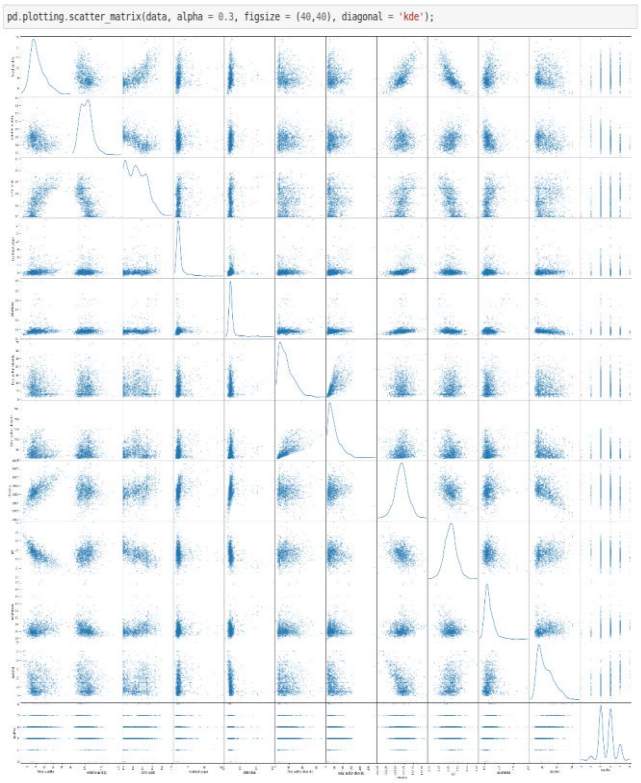


Figure 4 : scattered plot

You can get interesting details from the scatter plot above. For some features, the distribution appears to be fairly straightforward. In some other cases, the distribution appears to be negatively skewed. So this confirms our first doubts: there are indeed interesting interdependencies between some characteristics. You can draw a heat map of t he correlations between features to get more information.

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 4, DOI:10.29027/IJIRASE.v4.i4.2020.715-721, October 2020

```
correlation = data.corr()
# display(correlation)
plt.figure(figsize=(14, 12))
heatmap = sns.heatmap(correlation, annot=True, linewidths=0, vmin=-1, cmap="RdBu_r")
```

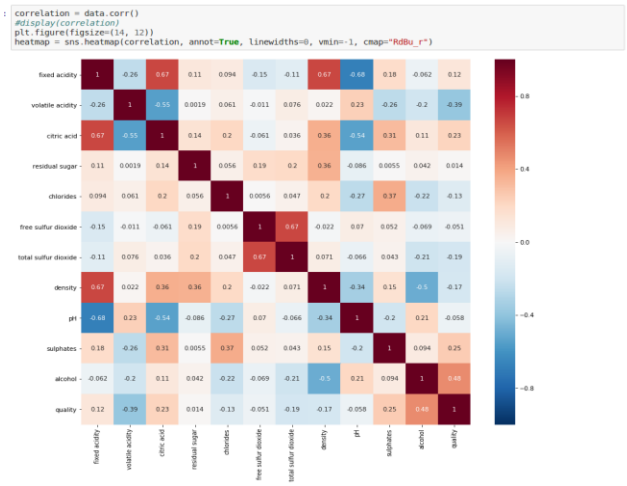


Figure 5:co-relation matrix

We can see a square with positive values that represent a direct relationship between the features. This means that the higher the value, the stronger the ratio that represents the reddish color in the diagram. That is, if one function increases, the other corresponding function increases and vice versa. Squares with negative values indicate an inverse correlation. The more negative value it will have, the more inversely proportional to the values. Finally, we can say that this is choir has little dependence between the two functions.

VI. RESULTS

The pH versus fix acidity programs it stone below;

```
#Visualize the co-relation between pH and fixed Acidity

#Create a new dataframe containing only pH and fixed acidity columns to visualize their co-relations
fixedAcidity_pH = data[['pH', 'fixed acidity']]

#Initialize a joint-grid with the dataframe, using seaborn library
gridA = sns.JointGrid(x="fixed acidity", y="pH", data=fixedAcidity_pH, size=6)

#Draws a regression plot in the grid
gridA = gridA.plot_joint(sns.regplot, scatter_kws={"s": 10})

#Draws a distribution plot in the same grid
gridA = gridA.plot_marginals(sns.distplot)
```

```
#Create a new dataframe containing only pH and fixed acidity columns to visualize their co-relations
fixedAcidity_pH = data[['pH', 'fixed acidity']]

#Initialize a joint-grid with the dataframe, using seaborn library
gridA = sns.JointGrid(x="fixed acidity", y="pH", data=fixedAcidity_pH, size=6)

#Draws a regression plot in the grid
gridA = gridA.plot_joint(sns.regplot, scatter_kws={"s": 10})

#Draws a distribution plot in the same grid
gridA = gridA.plot_marginals(sns.distplot)
```

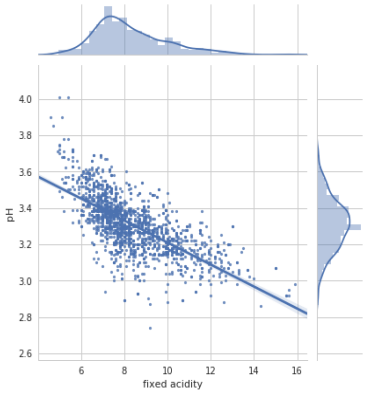


FIGURE 6: PH V/S ACIDITY

this scatter points demonstrate how the values of the page are changing with the fix acidity levels also, see that as the fixed acidity levels increase the pH levels get down.

Fix acidity versus citric acid is given below--l

```
fixedAcidity_citricAcid = data[['citric acid', 'fixed acidity']]
g = sns.JointGrid(x="fixed acidity", y="citric acid", data=fixedAcidity_citricAcid, size=6)
g = g.plot_joint(sns.regplot, scatter_kws={"s": 10})
g = g.plot_marginals(sns.distplot)
```

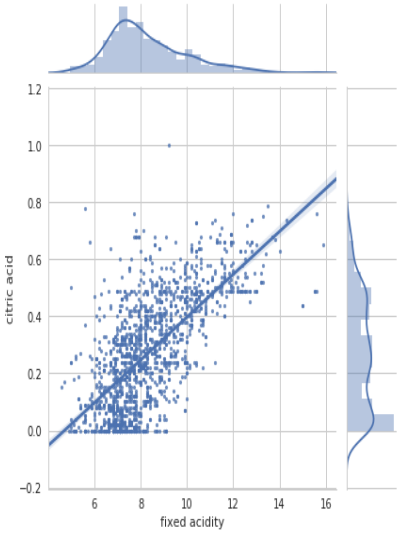


FIGURE 7 - Fix acidity versus citric acid

ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 4, DOI:10.29027/IJIRASE.v4.i4.2020.715-721, October 2020

Volatile Acidity vs Quality

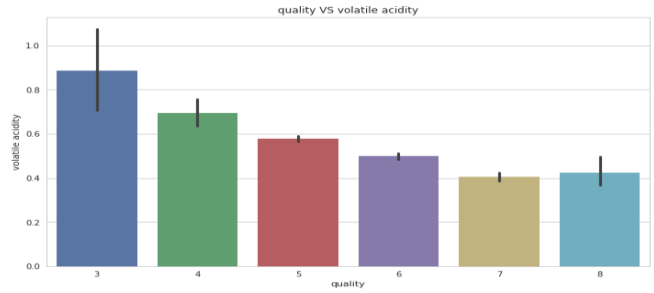
```
fig, axs = plt.subplots(ncols=1,figsize=(10,6))
sns.barplot(x='quality', y='volatile acidity', data=volatileAcidity_quality, ax=axs)
plt.title('quality VS volatile acidity')

plt.tight_layout()
plt.show()
plt.gcf().clear()
```

```
#We can visualize relationships of discreet values better with a bar plot

fig, axs = plt.subplots(ncols=1,figsize=(10,6))
sns.barplot(x='quality', y='volatile acidity', data=volatileAcidity_quality, ax=axs)
plt.title('quality VS volatile acidity')

plt.tight_layout()
plt.show()
plt.gcf().clear()
```



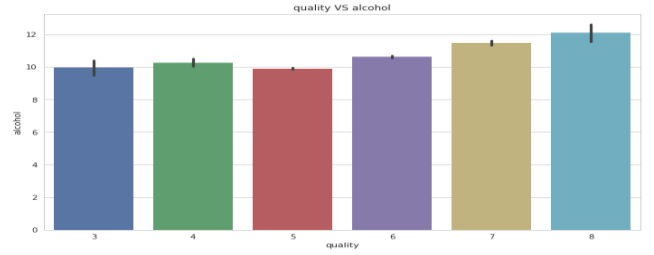
Alcohol vs. Quality

```
fig, axs = plt.subplots(ncols=1,figsize=(10,6))
sns.barplot(x='quality', y='alcohol', data=quality_alcohol, ax=axs)
plt.title('quality VS alcohol')

plt.tight_layout()
plt.show()
plt.gcf().clear()
```

```
fig, axs = plt.subplots(ncols=1,figsize=(10,6))
sns.barplot(x='quality', y='alcohol', data=quality_alcohol, ax=axs)
plt.title('quality VS alcohol')

plt.tight_layout()
plt.show()
plt.gcf().clear()
```



VII. CONCLUSION

So we might think that most people generally like wines with high alcohol content and are too monotonous and enthusiastic. High quality is usually associated with low levels of volatile acidity. This means that volatile acidity is a sign of a spoiler and can cause an unpleasant fragrance.

Results and discussion: Algorithm used for classification we are:

1) logistic regression
2) Stochastic descent of the gradient
3) support Vector classifier
4) Random forest

- • ***Logistic regression provided 86% accuracy in the logistic regression performance matrix.***

	Precision	Recall	F1-Score	Support
0	0.78	0.66	0.99	456
1	0.84	0.45	0.34	65

• The stochastic gradient descent was able to provide an average accuracy of 81%.Performance matrix of SGD:

	Precision	Recall	F1-Score	Support
0	0.87	0.77	0.66	453
1	0.33	0.77	0.34	88

Support Vector Classifier provided accurateness of 85%.

Performance matrix of SVC:

	Precision	Recall	F1-Score	Support
0	0.99	0.77	0.97	678
1	0.67	0.99	0.39	56

Random Forest has accuracy of 87.33%

	Precision	Recall	F1-Score	Support
0	0.89	0.77	0.45	888
1	0.98	0.64	0.54	87

VIII. 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 Qualities",ISSN 2147-67992147-6799,3rd September 2016

[[4] B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen, "Wineinformatics: 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] P.Appalasamy, A.Mustapha, N.D.Rizal, F.Johari, and A.F.Mansor, "Classification-based Data Mining Approach for Quality Control in Wine Production," Journal of Applied Sciences, 12(6), pp.598-601,2012

[6] N. H. Beltran, M. A. Duarte- MErmound, V. A. S. Vicencio, S. A. Salah, and M. A. Bustos, "Chilean wine classification using volatile organic compounds data obtained with a fast GC analyzer," Instrum. Measurement, IEEE Trans., 57: 2421-2436, 2008.

[7] 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

[8] 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.

[9] M.Forina, R. Leardi, C. Armanino, and S. Lanteri, "PARVUS An Extendible Package for Data Exploration," Classification and Correla,1988.

[10] Bledsoe, W. W. (1961). The use of biological concepts in the analytical study of systems. In the ORSA-TIMS National Meeting

[11] Holland, J. H. (1992). Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press.

[12] Jeong, I. S., Kim, H. K., Kim, T. H., Lee, D. H., Kim, K. J., & Kang, S. H. (2018). A Feature Selection Approach Based on Simulated Annealing for Detecting Various Denial of Service Attacks. Software Networking, 2018(1), 173-190