Wine Quality and Taste Classification Using Machine Learning Model

Anurag Sinha¹, Atul Kumar²

¹²Department of computer science and IT, Student, Amity University Jharkhand
Ranchi, Jharkhand(India), 834001
Email: anuragsinha257@gmail.com¹, atulkto08@gmail.com²

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 atasoy1,2,3 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. Applasammy8 proposed an algorithm to predict wine quality during wine production in warehouses. Beltran9 proposed an algorithm to classify wine data sets based on genetic algorithms and aroma chromatograms. Thakkar10, 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.
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 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:

![Machine Learning Algorithm Diagram]

#### 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 data set 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:

```python
# Import libraries necessary for this project
import numpy as np
import pandas as pd
from sklearn import time
from Python.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
import nbformat inline

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

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

Table 1: 12 different features of wine data set

<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed acidity</td>
<td>(g/tartaric acid)/dm³</td>
<td>4.600</td>
<td>15.90</td>
<td>8.320</td>
<td>1.741</td>
</tr>
<tr>
<td>Volatile acidity</td>
<td>(g/acetate acid)/dm³</td>
<td>0.120</td>
<td>1.580</td>
<td>0.528</td>
<td>0.179</td>
</tr>
<tr>
<td>Citric acid</td>
<td>(g/dm³)</td>
<td>0.000</td>
<td>1.000</td>
<td>0.271</td>
<td>0.195</td>
</tr>
<tr>
<td>Residual sugar</td>
<td>(g/dm³)</td>
<td>0.900</td>
<td>15.50</td>
<td>2.539</td>
<td>1.410</td>
</tr>
<tr>
<td>Chlorides</td>
<td>(g/sodium chloride)/dm³</td>
<td>0.012</td>
<td>0.611</td>
<td>0.087</td>
<td>0.047</td>
</tr>
<tr>
<td>Free sulfur dioxide</td>
<td>(mg/dm³)</td>
<td>1.000</td>
<td>72.00</td>
<td>15.87</td>
<td>10.46</td>
</tr>
<tr>
<td>Total sulfur dioxide</td>
<td>(mg/dm³)</td>
<td>6.000</td>
<td>289.0</td>
<td>46.47</td>
<td>32.89</td>
</tr>
<tr>
<td>Density</td>
<td>(g/cm³)</td>
<td>0.990</td>
<td>1.044</td>
<td>0.997</td>
<td>0.002</td>
</tr>
<tr>
<td>pH</td>
<td></td>
<td>2.740</td>
<td>4.010</td>
<td>3.311</td>
<td>0.154</td>
</tr>
<tr>
<td>Sulphates</td>
<td>(g/potassium sulphate)/dm³</td>
<td>0.330</td>
<td>2.000</td>
<td>0.658</td>
<td>0.170</td>
</tr>
<tr>
<td>Alcohol (vol)</td>
<td>(%)</td>
<td>8.400</td>
<td>14.90</td>
<td>10.42</td>
<td>1.066</td>
</tr>
</tbody>
</table>
```
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.

```python
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.

```python
% Visualize skewed continuous features of original data
```

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.
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:-

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

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 the correlations between features to get more information.

<table>
<thead>
<tr>
<th>Table 2: The physicochemical data statistics of white wine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>------------------------------------------</td>
</tr>
<tr>
<td>Fixed acidity (g/tartaric acid)/dm^3</td>
</tr>
<tr>
<td>Volatile acidity (g/acetic acid)/dm^3</td>
</tr>
<tr>
<td>Citric acid (g/dm^3)</td>
</tr>
<tr>
<td>Residual sugar (g/dm^3)</td>
</tr>
<tr>
<td>Chlorides (g/sodium chloride)/dm^3</td>
</tr>
<tr>
<td>Free sulfur dioxide (mg/dm^3)</td>
</tr>
<tr>
<td>Total sulfur dioxide (mg/dm^3)</td>
</tr>
<tr>
<td>Density (g/cm^3)</td>
</tr>
<tr>
<td>pH</td>
</tr>
<tr>
<td>Sulphates (g/potassium sulphate)/dm^3</td>
</tr>
<tr>
<td>Alcohol (%vol)</td>
</tr>
</tbody>
</table>

Figure 4: scattered plot

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 zero correlation has little dependence between the two functions.

VI. RESULTS

The pH versus fixed acidity programs it stone below:

```R
# Visualize the correlation between pH and fixed acidity.
# Create a new dataframe containing only pH and fixed acidity columns to visualize their co-relationships.
fixedAcidity_pH <- data[, c('pH', 'fixed acidity')]

# Initialize a grid with the dataframe, using seaborn library.
grid = sns.jointplot(x='fixed acidity', y='pH', data=fixedAcidity_pH, size=6)

# Draw regression plot in the grid.
grid = grid.plot_joint(sns.regplot, scatter_kws={'s': 10})

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

**FIGURE 6: pH V/S ACIDITY**

This scatter plot demonstrates how the values of the page are changing with the fixed acidity levels also, see that as the fixed acidity levels increase the pH levels get down.

**Fix acidity versus citric acid is given below--1**

```R
fixedAcidityCitricAcid = data[, c('citric acid', 'fixed acidity')]
g = sns.jointplot(x='fixed acidity', y='citric acid', data=fixedAcidityCitricAcid, 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**
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.

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.78</td>
<td>0.66</td>
<td>0.99</td>
</tr>
<tr>
<td>1</td>
<td>0.84</td>
<td>0.45</td>
<td>0.34</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.87</td>
<td>0.77</td>
<td>0.66</td>
</tr>
<tr>
<td>1</td>
<td>0.33</td>
<td>0.77</td>
<td>0.34</td>
</tr>
</tbody>
</table>

Support Vector Classifier provided accurateness of 85%.

Performance matrix of SVC:

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.99</td>
<td>0.77</td>
<td>0.97</td>
</tr>
<tr>
<td>1</td>
<td>0.67</td>
<td>0.99</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Random Forest has accuracy of 87.33%
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.


