Wine Quality Analysis

Atharva Gadad
SRN: PES1UG20CS088
Computer Science and Engineering
PES University
Bangalore, India
gadadatharva123@gmail.com

Dhruv Garodia
SRN: PES1UG20CS527
Computer Science and Engineering
PES University
Bangalore, India
dhruvgarodia2002@gmail.com

Aaditya Vikram
SRN: PES1UG20CS528
Computer Science and Engineering
PES University
Bangalore, India
aaditya.vikram1903@gmail.com

Alan S Paul
SRN: PES1UG20CS624
Computer Science and Engineering
PES University
Bangalore, India
alanpaul303@gmail.com

***Abstract* — The wine industry is a fast-growing one, and has shown exponential growth over the past few decades. With this consumption growth comes the need to deliver the best quality wine. Understanding how the variation in one of the factors associated with the wine affects the overall quality of the wine is the primary objective of this study. We try to do Exploratory Data Analysis along with visualizations, and literature review.**

***Index Terms*—Wine Quality, Exploratory Data Analysis**

I. INTRODUCTION, AND BACKGROUND

The consumption of wine has been on a rise for a long time. So, wine companies are focused on improving the quality of their wines to sustain this growth. Because of this, there has been an increase in demand for wine tasters. However, this is a dying vocation because of the reluctance of younger generations to join this trade. Hence, a need has arisen for automation of this aspect of production to fill the labor gap with models to analyze the quality of wines. This also helps reduce the operating costs of these companies, because fewer wine tasters will be required once these models become accurate and hence can help them increase their profits. As a result, companies have been pouring billions into the research and development of new technology to improve existing techniques. The quality of wines can be analyzed using various physicochemical and sensory attributes. The attributes we are using for modeling include fixed acidity, volatile acidity, citric acid, residual sugar, chloride (which indicates salt content), free sulfur dioxide, the density of the liquid, pH, sulfates, and alcohol content.

II. RELATED WORK/LITERATURE REVIEW

Our data has been obtained from Kaggle, and lists both white wines and red wines which have 11 attributes affecting the quality of wines and determining what type of wine it is. It is these 11 attributes that form the basis of the model we will be building for classification. In 1991, a wine dataset containing 178 instances was pushed into the UCI repository to classify 3 cultivars from Italy [1]. A work on wine classification depends on the physicochemical information, one among them being wine aroma-chromatograms, measured using a Fast GC Analyzer [2].

In the literature we have gone through so far, many types of classification algorithms have been used to varying degrees of success so as classifying different qualities of wines based on the various physicochemical and sensory factors listed above. The classification models include Linear Discriminant Analysis, Radial Basis Function of Neural Networks, and Support Vector Machines (SVM), comparing their results and performances in a two-staged architecture. Cortez *et al.* [3] have proposed a taste prediction technique for wine quality assessment. Their prediction technique involved the use of Support Vector Machines, Multiple Regressions, and Neural Networks for analyzing the chemical structure of wines. Shanmuganathan's technique was about predicting the seasonal and climatic effects on wine yields and wine quality [4]. The wine informatics system according to Chen *et al.* [5] has collected the reviews from wine consumers, processing the data achieved using Natural Language Processing (NLP). They also used hierarchical clustering and association rules. Yesim *et al.* [6] have used k-Nearest Neighbor Classification, Random Forests, and Support Vector Machines, for their modeling of the problem. Random Forest method has given them the best results for classifying white and red wines. For evaluation, two test models were used, namely k-fold cross-validation (k-fold cv) mode, and percentage split (holdout method) mode. In Sunny Kumar *et al.* [7] Naïve Bayes algorithm has been used in addition to the above-mentioned ones for the same task. They too have obtained the best results with the help of Random Forest technique. We hope to be able to use similar models to achieve good accuracy scores and prepare a usable model.

III. PROPOSED SOLUTION

For the first step we will be importing the necessary libraries required for analyzing the dataset.In this case, we will be using the NumPy , Pandas, and Seaborn for analyzing the dataset.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
```

For the first step, we will load the wine dataset.csv file by using the data_frame.shape command we can check the number of rows and columns present in the dataset.
The shape of the dataset:

```
(6497, 13)
```

For the second step we want to do a summary analysis for each of the columns to find the median values, mean, standard deviation, and more. From the given dataset we can see that there is a huge difference between the 75 percentile and the max value for residual sugar , free sulfur dioxide, and total sulfur dioxide. Indicating the skewness and how the outliers are affecting each column.

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates
count	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000
mean	7.215307	0.339666	0.318633	5.443235	0.056034	30.525319	115.744574	0.994697	3.218501	0.531268
std	1.296434	0.164636	0.145318	4.757804	0.035034	17.749400	56.521855	0.002999	0.160787	0.148806
min	3.800000	0.080000	0.000000	0.600000	0.009000	1.000000	6.000000	0.987110	2.720000	0.220000
25%	6.400000	0.230000	0.250000	1.800000	0.038000	17.000000	77.000000	0.992340	3.110000	0.430000
50%	7.000000	0.290000	0.310000	3.000000	0.047000	29.000000	118.000000	0.994890	3.210000	0.510000
75%	7.700000	0.400000	0.390000	8.100000	0.065000	41.000000	156.000000	0.996990	3.320000	0.600000
max	15.900000	1.580000	1.660000	65.800000	0.611000	289.000000	440.000000	1.038980	4.010000	2.000000

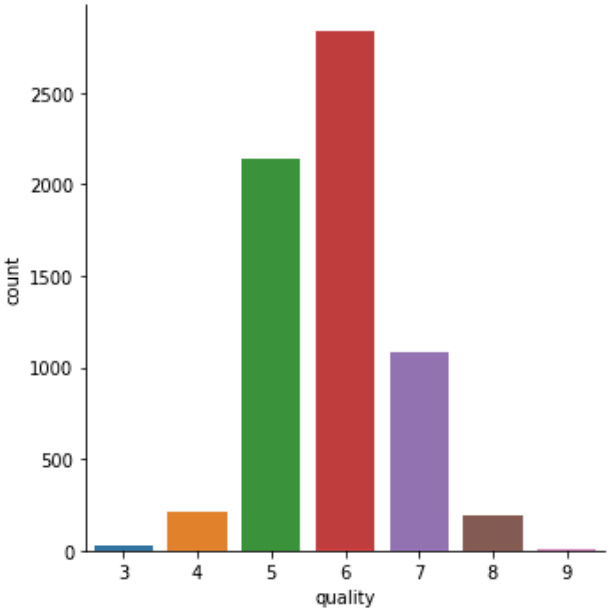
alcohol	quality
6497.000000	6497.000000
10.491801	5.818378
1.192712	0.873255
8.000000	3.000000
9.500000	5.000000
10.300000	6.000000
11.300000	6.000000
14.900000	9.000000

For the third step, we are checking if there are any null values. In the given dataset we do not have the presence of any null values.

```
df.isnull().any().any()#no null values
```

```
False
```

For the given dataset we only have one column which is of categorical data type. So we can visualize this with the help of a bar chart. From this, we can see that a lot of wines belonged to the category 5,6,7 when compared to 3 and 9.



We are going to try to do a comparative study, trying to understand how various models compare in terms of their performance on the classification of wines as red and white. We will also do a comparative study on the ability of these models to classify wines as good or bad. We set a threshold of 7 for the goodness of wine. Wines with a rating above 7 are considered good, while those below are considered to be of suboptimal quality

Our data has 6497 entries and 13 attributes. These 13 include the labels we created to help with the modeling process. One of them is the label ‘Label’, which is a binary variable to mimic the type of wine, which has values of red or white was categorical and couldn’t be used. We also ended up creating another label called ‘goodquality’ to classify wines with a rating above 7 as good (1) and the others as poor (0).

First, we try to select relevant features, discarding the irrelevant ones. We are going to implement this based on information gain (mutual information), random forests, etc. Then we are going to be performing a modal analysis to find the most important features.

Then we will be using multiple classifiers like Artificial Neural Networks (using Adam and SGD optimizers), Support Vector Machines, Logistic Regression, XGBoost, etc., and compare their accuracy to determine their performance.

The next phase will involve the classification of the wines into good and poor quality. We plan on using Random Forest Classifiers, Support Vector Machine with Grid Search, Artificial Neural Networks (using Adam, SGD, Adagrad optimizers, Hyperparameter tuned SGD), XGBoost, Decision Tree, Logistic Regression, etc.

IV. THEORETICAL BACKGROUND AND EXPERIMENTAL RESULTS

The theoretical background behind some of the models we have used is as follows:

Backward Elimination-based Feature Selection

In backward elimination, initially a significance level is set. In most cases, it is 0.05.
Then, we fit a selected ML model to the dataset and identify the feature with the highest p-values.
Now, if this p-value is greater than the significance level we had set, then we drop the feature.
Then, we repeat the same steps until we're left with the features that have p-values less than
the significance values.

Random Forest Feature Selection

Random forest is a model where decision trees and bagging are used. The training dataset is resampled using bootstrapping. Each tree of the random forest is used to analyze the importance of a feature by measuring its ability to increase the pureness of leaves. Higher purity means higher importance of the feature. This is done for all trees, then averaging the trees and normalizing them to 1. We try to do this for all features, remove less relevant features and only keep those that maximize performance.

Feature Selection by Backward Elimination:

Mutual Information based Feature Selection makes use of information gain. In this technique, the information gain about each feature concerning the target variable is calculated. After this, the features with the highest information gains are retained. The other features are eliminated.

Random Forest regressor

Random Forest is an ensemble learning technique. Here we make use of many basic decision trees, which act as stubs. Then we take the predictions from all the trees and use majority voting for classification and averaging for regression.

Decision Tree Regressor:

Decision Trees are a supervised learning algorithm, which makes use of a heuristic like Information Gain, Gini, etc., to find the features which can help partition the data in the most general ways and ultimately end up with leaf nodes.

ANN

The next model we used was the ANN which was a deep learning model which we tested out in our dataset. In a sense, they can make adjustments or learn as they receive new inputs. Due to this ability, they can uncover hidden patterns and make very accurate predictions for our dataset. We have used RELU as the activation is the non-saturation of its gradient, which greatly accelerates the convergence of stochastic gradient descent when compared to others.

SVM

Support vector machines were utilized in our instance to classify data into multiple categories. The dataset's total number of features, N, which divides data points into various classes or categories, is found to exist as a hyperplane in the vector space of N dimensions. The SVM can select from a wide variety of hyperplanes to divide the classes. It ends up on a plane that is farthest away possible from the data points for either of the classes or categories. To reduce classification error when test data points are received, the marginal distance should be maximized.

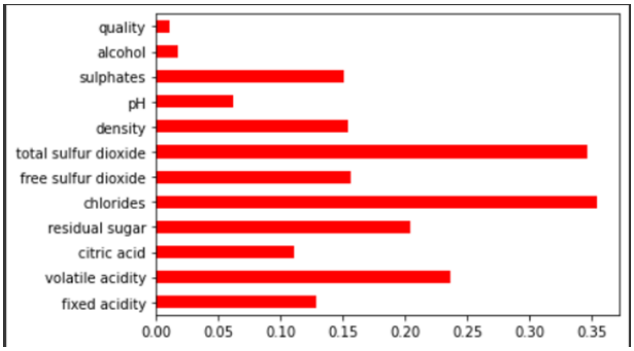
XGBoost

Each decision tree's data is considered by XGBoost as it constructs one tree at a time, and any missing data is filled in. This made it easier to combine the decision tree algorithm and gradient algorithms for better outcomes.

1) Classification of wines into red and white:

Our first task was Feature Reduction. We tried various different types of models:

The first technique used was based on Information Gain. Based on Information Gain, we tried identifying the features with the highest Information Gain. The results were as follows:



Another technique we used was Random Forest Feature Selection. We tried to calculate the relative importance of each feature in terms of ensuring the purity of leaves and ended up with the following:

Importances	Features
0.044511	fixed acidity
0.120775	volatile acidity
0.016145	citric acid
0.044068	residual sugar
0.270884	chlorides
0.049333	free sulfur dioxide
0.308913	total sulfur dioxide
0.059421	density
0.021777	pH
0.051965	sulphates
0.009736	alcohol
0.002472	quality

Finally, modal analysis was performed to select the following features: volatile acidity, residual sugar, chlorides, total sulfur dioxide, density sulfates, alcohol. Since the type of alcohol (Red/White) is categorical, we created a new label called 'Label,' with classification based on the type of wine. Then we split the dataset into a Training and Testing Set. We have used an 80:20 split. We have implemented many models, to understand how they stack up against one another. First, we implemented RELU-based ANN. Using the Adam optimizer, we obtained 98.1% accuracy. On using the SGD optimizer, we obtained 94.47% accuracy

For the SVM, we obtained an accuracy of 92.84%. We implemented the Logistic Regression-based classifier, which gave us an accuracy of 97.69%. On implementing XGBoost, we obtained an accuracy of 99.38%

Model	Accuracy
ANN (Adam)	0.981
ANN (SGD)	0.9447
SVM	0.9284
Logistic Regression	0.9769
XGBoost	0.9938

2) Classifying wine as good or poor:

We created a new label called 'goodquality', as explained earlier. We split the dataset into a training and testing set in an 80:20 ratio. ANN with Adam optimizer gives an accuracy of 80.615%. ANN with SGD optimizer gives an accuracy of 80.538%. ANN with the Adagrad optimizer gives an accuracy of 81.846%. Random Forest Classifier gives an accuracy of 87.839%. Random Forest Classifier with grid search gives an accuracy of 80.538%. Decision Tree Classifier gives an accuracy of 82.3846%. Logistic Regression Classifier gives an accuracy of 82%. XGBoost Classifier gives an accuracy of 83.538%. Support Vector Classifier gives an accuracy of 80.538%. SVC with grid search cv gives an accuracy of 84%.

Model	Accuracy
ANN (Adam)	0.80651
ANN (SGD)	0.80538
ANN (Adagrad)	0.81846
Random Forest	0.87839

Random Forest with Grid Search	0.80538
Decision Tree	0.823846
Logistic Regression	0.82
XGBoost	0.83538
SVC	0.80538
SVC with Grid Search CV	0.84

We also convert the numerical ratings of the wines into a relative rating. This is done using min-max standardization.

Insights

Our observations from this work have matched previous work we had gone through. We found Random Forest Models to be the best at classification when compared to other models. Upon performing feature selection using Information Gain and Backward Feature Elimination method, we have found out that the methods mutually agree on five features, namely [Volatile Acidity, Residual Sugar, Chlorides, Total Sulphur Dioxide, and Density]. Then from literature surveys and an industrial point of view, we found that Sulphate is very important to enhance the taste and appearance of wine. We choose alcohol as our next feature because the higher alcohol is desirable to enhance the quality and give an oaky flavor to the wine.

V. CONCLUSIONS

In this paper we have worked on different models to predict the quality and type of wine. We have used various models such as ANN, SVM, XGBoost, etc. Most of the models are giving accuracy averaging around 82%.

Contributions

Atharva S Gadad (PES1UG20CS088) and Dhruv Jyoti Garodia (PES1UG20CS527) worked on the Literature Survey, Report, and classification of wines based on their type. Alan S Paul (PES1UG20CS624) and Aditya Vikram (PES1UG20CS528) worked on the slides, EDA, and classification of wines based on their quality.

REFERENCES

[1] A. Asuncion, and D. Newman (2007), UCI Machine Learning Repository, University of California, Irvine,[Online].

[2] 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.

[3] P. Cortez, A. Cerderia, F. Almeida, T. Matos, and J. Reis, "Modelling wine preferences by data mining from physicochemical properties," In Decision Support Systems, Elsevier, 47 (4): 547-553. ISSN: 0167-9236.

[4] S. Shanmuganathan, P. Sallis, and A. Narayanan, "Data mining techniques for modeling seasonal climate effects on grapevine yield and wine quality," IEEE International Conference on Computational Intelligence Communication Systems and Networks, pp. 82-89, July 2010

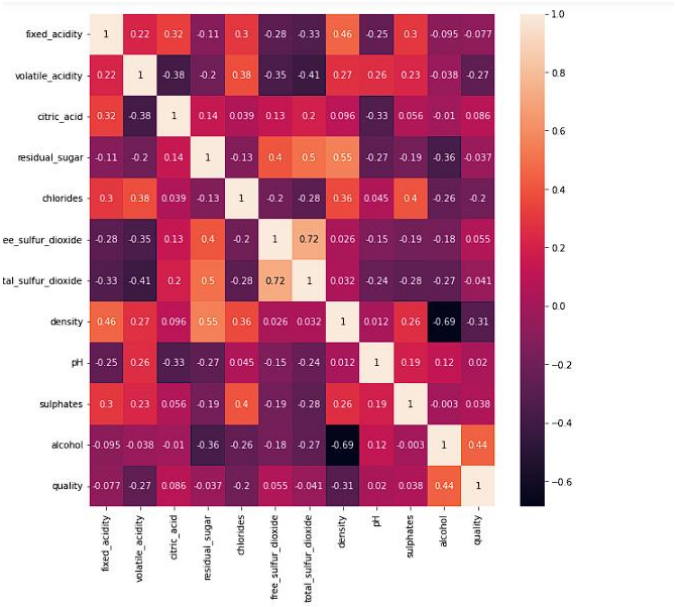
[5] B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen, "Wineinformatics: applying data mining on wine sensory reviews processed by the computational wine wheel," IEEE International Conference on Data Mining Workshop, pp. 142-149, Dec. 2014.

[6] Yesi Er, and Ayten Atasoy, " The Classification of White Wine and Red Wine According to Their Physiochemical Qualities", International Journal of Intelligent Systems and Applications in Engineering, pp. 23-26, Sept. 2016.

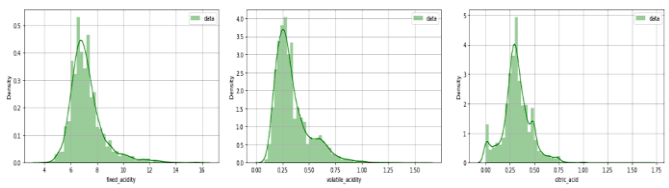
[7] Sunny Kumar, Kanika Agarwal, and Nelshan Mandan, "Red Wine Quality Prediction Using Machine Learning Techniques", 2020 International Conference on Computer Communication and Informatics (ICCCI -2020), Jan. 2020.

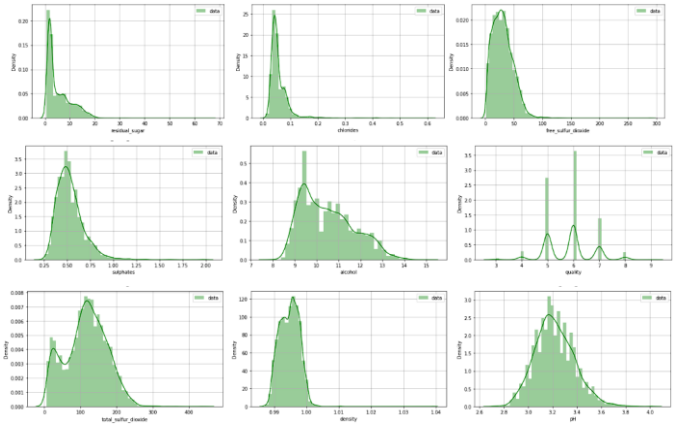
APPENDIX

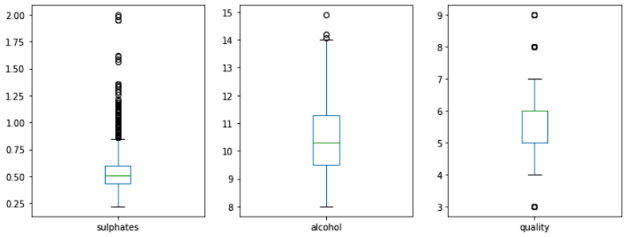
A pairwise heatmap plot is created to see how each of the features is related to each other.



Next, we check whether the data is normally distributed with respect to each column, so a distribution plot is created over each column to analyze this. The pH feature is approximately normally distributed; all the remaining features are positively skewed.







With the help of a box plot, we can find the five-number summary of a set of data which is the minimum , first quartile, median, third quartile, and maximum. From the given box plots we can see that except for alcohol all the other box plots have outliers.