AIES MINI PROJECT

GROUP MEMBERS:
DEV SHARMA(PE_18)
ANKIT KUMAR(PE_21)
RIJUL (PE_27)
ABHAY NAYAK (PE_28)

Wine Quality predication analysis

Abstract*:* Wine classification is a difficult task and also we do not know on what basis taste can be identified and considered to be a good wine. Predicting the quality of wine can help in certification phase, at present sensory analysis is performed by food tasters being clearly a subjective approach. Furthermore, a feature selection process can help to analyze the impact of the analytical tests. For our work, we collected the dataset of various red and white variants of the Portuguese "Vinho Verde" wine from Kaggle, this includes various physicochemical properties. We used Google Colab to work on this dataset. Machine learning algorithms are used to detect few excellent or poor wine qualities. We preprocessed the data by identifying and handling the missing values. One Hot encoder is used to convert the categorical values into numerical values. The Feature Scaling step is used to normalize the range of independent variables. We have many machine learning algorithms for prediction among them we used Logistic Regression, Decision Tree Classifier, Random Forest Classifier and Extra Trees Classifier. We trained the dataset by all the four models and compared the accuracy and precision to choose the best machine learning algorithm. In turn, this helps us to predict the quality of wine on a range of 0–10 given a set of features.

Index Terms **– Wine Quality, Machine Learning, Logistic Regression, Random Forest Classifier, Decision Tree, Extra Trees Classifier.**

I. INTRODUCTION

Consumption of wine has increased dramatically over the years not only for its fun and pleasure but also because of its welfare to the human heart. Many industries are adopting and applying new techniques and implementing the same to increase the production and making the whole process effective. The production of wine is increasing over time also their demands. Wine consumption has various purposes, but the chemicals used in them are related, but the chemical components need to be examined hence, we adopt these techniques to verify. Once wine consumption was considered as a royal commodity, but today it is liked by wide range of customers. Portugal is the 11th largest wine producer in the world.

Machine Learning is a sub-field of Artificial Intelligence (AI). Now a day's machine learning is very important. Machine Learning gives the ability to learn and improve from experience without being explicitly programmed. Machine learning is used to make the systems learn from data by identifying patterns and making decisions with minimal human intervention. Traditionally data analysis was a trial and error-based approach of the data. By using efficient algorithms and fast analyses machine learning algorithms produce accurate results. The value of machine learning is recognized by the industries working on large amount of data. Machine learning can be used in various fields like food, finance, health, government, retail, transportation, etc. Machine Learning is used to understand the data and fit the data in to several machine learning models that can be used by people for future decisions. We have known that the lifeblood of a business is data. For future decisions and competition of the business will depend on the data. In healthcare and life science, machine learning is used for disease identification and risk satisfaction. We can easily build the machine learning model without any advanced statistics. As we input more data into the machine, this helps the algorithms to teach the computer. Computations and Transactions of the previous help the applications to learn. We cannot apply the machine learning model directly to real-world data. Real-world data is preprocessed before the application of the model. The data set is divided into train and test data. To teach the computer machine learning algorithms use training data. We predict unknown data using machine learning algorithms.In our model, we used a machine learning algorithm to predict the wine quality. These technologies are also helpful to enhance the production and making the whole process smooth. Machine learning is used in predicting quality in food industry, powerful data tools are needed to extract useful information from the huge amount of data. Classification algorithms use several attributes for finding the quality. Machine learning will predict the values accurately. We can apply many machine learning algorithms among them we will choose which one will give us the best accuracy.

II. LITERATURE SURVEY

Few researchers worked on Wine Quality Prediction using Machine Learning. In recent years some of them are described below.

Three authors namely Paulo Cortez, Juliana Teixeira, António Cerdeira Fernando Almeida Telmo Matos José Reis worked on a paper using Data Mining techniques by using Support vector machine(SVM) and Neural Network(NN) on wine quality assessment. In this paper, wine preferences is predicted based on easily available analytical tests at the certification step, using a data mining approach. A large dataset was considered with red and white vinhos verdes samples from the Minho region of Portugal. To preserve the order of grades, they used regression approach to model wine quality. Promising results were achieved, with the SVM model providing the best performances, outperforming the NN and MR techniques. The overall accuracies are 64.3% and 86.8%.

Chen et al put forward an approach using the human savory reviews [3] to predict wine quality. They got an accuracy of

73.79~82.58% to predict wine quality using Hierarchical clustering approach and association rule algorithm.

Thakkar et al. to rank the attributes primarily they used analytical hierarchy process followed by machine learning classifiers such as random forest and support vector machine and they found accuracy of 70.33% using random forest and 66.54% using support vector machine [4].

Wine quality testing is one of the most important in this context and this test can be used for validation. Such type of quality certification helps to ensure the quality of wine in the market. Wine has various properties such as density, pH value, alcohol, acidity, variable acidity, and acids such as citric, sulfate, chlorides, free sulfur dioxide, and total sulfur dioxide content. The quality of the wine can be assessed by two types of tests; firstly physicochemical tests and secondly nerve tests. Physicochemical tests can be determined by laboratory tests and no human specialist is required but for neurological testing, a human specialist is needed. In addition, wine quality testing is very difficult as the relationship between physicochemical and sensory analysis is complex and not yet fully understood.

III. METHODOLOGY

3.1 Workflow

- **Data Collection:** The two datasets are related to red and white variants of the Portugal "Vinho Verde"wine. These datasets can be viewed as classification or regression tasks. The Wine Quality dataset contains 13 columns and 6498 rows.

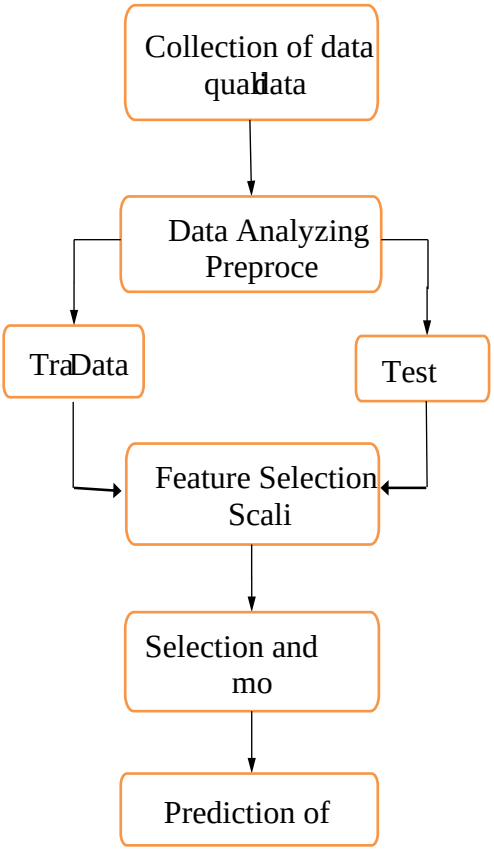

Fig 1 work flow of wine

- **Data Pre-Processing:** In real-world data there can be missing values or noisy and inconsistent data. If data quality is low then no quality results may be found. It is necessary to pre-process the data to achieve quality results.
- **Training/test sets:** we can evaluate a dataset using a train-test split. First, the loaded dataset must be split into input and output components.
- **Feature Selection and Feature Scaling:** Feature selection is the process of selecting a subset of relevant features (variables, predictors) for use in model construction and it is desirable to reduce the number of input variables to both reduce the computational cost of modeling and also to improve the performance of the model. Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. In our Wine Quality Dataset we have 13 features but we make use of 12 features.
- **Selection of model:** It is the process of selecting final machine learning model from among a collection of training dataset. It is the process of choosing one of the models such as Logistic Regression, Decision Tree Classifier, Random Forest Classifier, andExtraTreeClassifier for a training dataset and the technique is capable of performing both regression and classification tasks.

Prediction: By predicting the test set results, machine learning model is using the dataset to answer questions and the checking of accuracy is done in this step of prediction or inference.

3.2 Data Exploration

The red wine and white wine datasets have been used in this paper which is obtained from the Kaggle it contains a large collection of datasets that have been used for the machine learning community. The dataset contains two excel files, related to red and white wine varients. The red wine dataset consists of 1599 instances and the white wine dataset consist of 4898 instances. Both datasets have 11 input variables and 1 output variable (based on sensory data): quality. Sensory data is scaled in 11 quality classes from 0 to 10 (0very bad - 10-very good). We had used the Google Colab tool and python as a programming language for our model. Colab is opensource software and it contains live code, equations, visualization, it can be used in carrying various ML Process.

In the Data Exploration step first, we import the libraries use to build our model. For our model, we used mat plot lib, pandas, and NumPy libraries. For data analyzing and numerical plotting, we used Matplotlib. Pandas are a very important library, it helps us to work on data structures and it also supports sorting, Re-indexing, iteration. NumPy is used for complex mathematical implementations. Secondly, we read the dataset using panda's library.

3.3 Data Preprocessing

It is very difficult to work on real-world data. Real-world data is inconsistent, incomplete and lacking from certain too many errors. If we format the data into an understandable format it will be useful to predict the high accuracy and efficiency of our model. The data preprocessing step enhances the quality of data and in this step, we will extract meaningful insights from the data. Data Preprocessing contains many steps which will help to fill the missing value in data, as we have known some models will not work on the Null values. All dependent and independent variables should be numerical, so in preprocessing step we can convert categorical data into numerical data.

3.3.1 Handling Missing Values

Missing values can be deleted but doing like this we lose the information so we have various methods to fill the missing values. First, we can fill it manually but it can't be implemented for big datasets. Secondly, it can be filled by a mean value of that attribute when the data is normally distributed, In the case of non-normal distribution median value of the attribute can be used. In the dataset, we used the mean of the fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, sulfates and pH attribute to fill the missing values.

Table 3.1 No of Missing Values in Attributes

Attributes Name	No of Missing Values
fixed acidity	10
volatile acidity	8
citric acid	3
residual sugar	2
Chlorides	2
free sulfur dioxide	0
total sulfur dioxide	0
Density	0
pH	9
Sulfates	4
Alcohol	0
Quality	0

.

Referencing to the Table 1 we can conclude that, the fields namely fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, pH and sulfates have 10, 8, 3, 2, 2, 9, and 4 lost values.

3.3.2 Handling Categorical Values

Dataset consists of two types of wine (White, Red). We have to convert all categorical attribute to numerical attribute. There are many methods for handling categorical values. The first method is ordinal encoding, where each of the unique value is replaced by integer value but it is a natural encoding for ordinal values. This can cause problems for nominal values. One Hot Encoding is the best way to treat the categorical attributes. The variable which is integer encoded is removed and a new binary variable is added.

3.4 Data Visualization

In the data visualization process, representing data or information on a graph, chart or other visual format communication between data and information can be accessed through data display. We can easily understand the data with illustrated representation. In machine learning we have many types of graph representations such as histograms, congestion sites, box sites, connection matrix sites, dispersing matrix sites. Fig3.2 represents the sites of the box. The box layout is also used to find the outside of the data set. It captures a snapshot of the data well with a simple box and mustache and allows us to easily compare across groups. In our modeling properties such as citric acid, residual sugar, free sulfur-dioxide and total-sulfur-dioxide have a high concentration of foreign substances; by removing foreign substances we can improve efficiency by 0.2 - 0.4%.

Figure 3.3 represents a remote structure. It is basically used in a fixed set of views and visualizes it with a histogram i.e. only one view and that is why we choose one column of the database i.e., overcrowding. From the graph we can say that all the values are in the normal distribution and the value range is small, the free sulfur dioxide is in the main range so we use log conversion.

Figure 3.4 represents Log Modification. Log modification reduces or eliminates the inclination of our original data in the free sulfur dioxide free attribute. Figure 3.5 represents the number of wines namely, white and red wines. We have almost 4900 white wine database numbers and almost 1800 red wine database values.

Figure 3.6 represents the range of wine quality (0-10), but in the wine database it is in the range (3-9), and the majority category is between 5 and 7, so the model will be biased towards these three classes as the average quality, but we need to make the model more flexible. Standard test.

Fig. 3.7 represents the correlation between attributes, from the correlation matrix we can say that alcohol is highly correlated i.e., 0.44 in the output variability i.e., quality. Throughout the matrix there is a negative correlation of alcohol and congestion and 0.75 is best correlated between free sulfur dioxide and total sulfur dioxide content.

3.4 Data Visualization

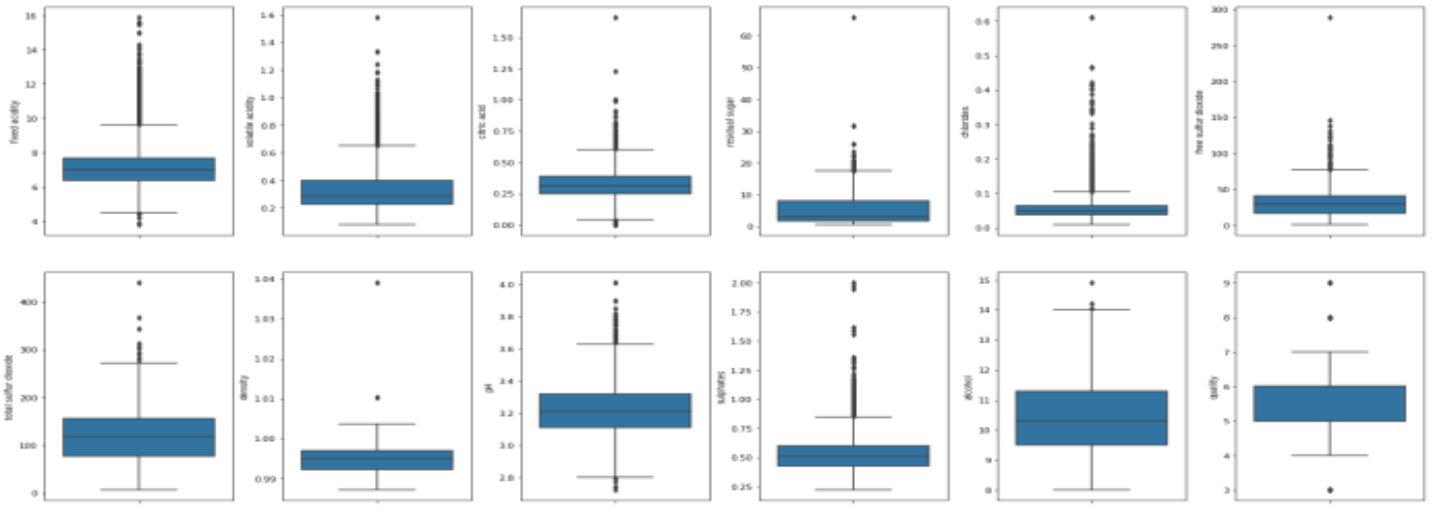

Fig-3.2 Box-Plot

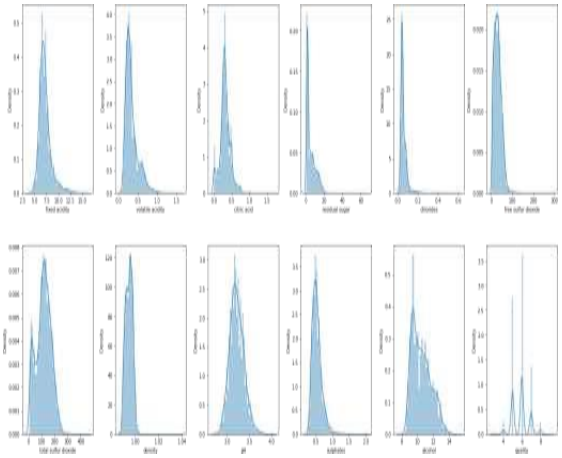

Fig 3.3 Dist Plot

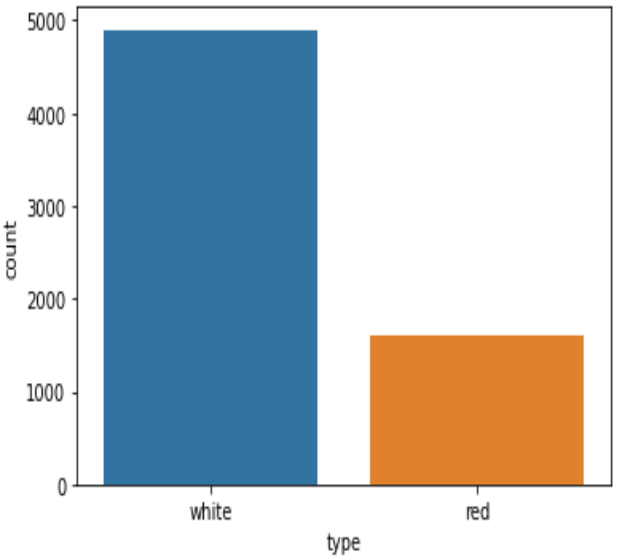

Fig 3.5 Red and White wine count

======

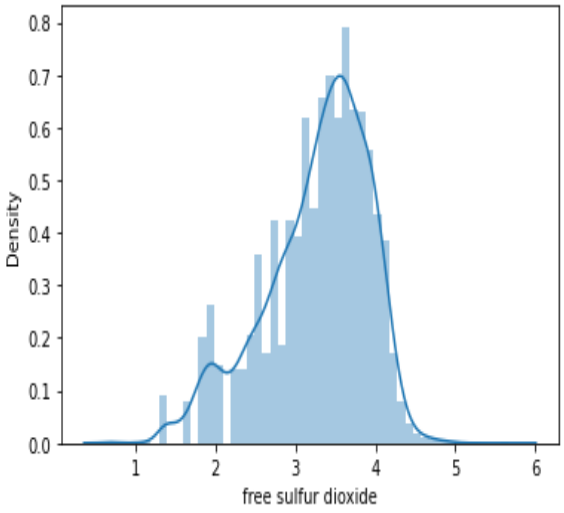

Fig 3.4 Log transformation

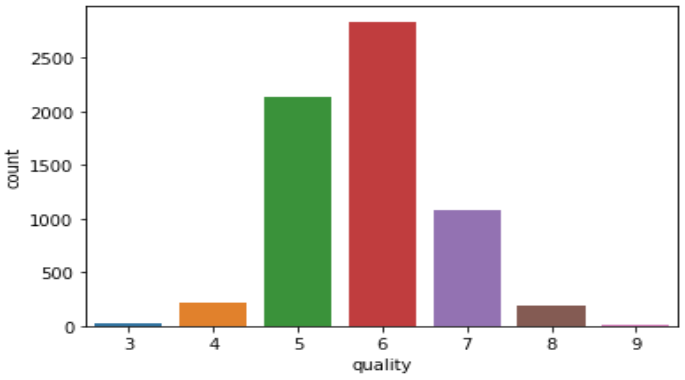

Fig 3.6 Quality Range

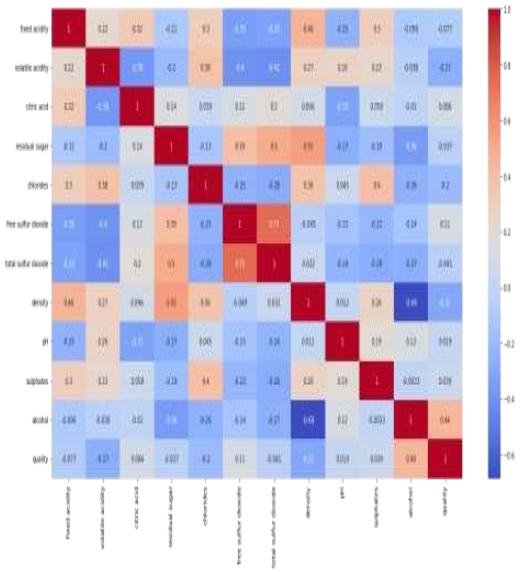

Fig 3.7 Correlation Matrix

3.5 Feature Selection

Feature selection is the process in which we choose appropriate features as train dataset in constructing the model by reducing the number of feature variables so it can reduce computing cost and improve the performance of the model.

3.6 Feature Scaling

Feature Scaling is a technique tostandardize the input variables present in the data in a fixed range. In our Wine Quality Dataset we have 13 features but we make use of 12 features.

3.7 Model Selection

Machine Learning is divided into three types of Supervised, Unsupervised and Reinforcement. Supervised Reading is used when data is provided and output is known. In Supervised Reading, we have two types: - Regression and Classification. Regression is used to predict continuous values, Classification is used to identify a unique class, unsupervised learning is used when we have no idea about our exit, and Reinforcement is used to find the best behavior in a particular situation.

It is the process of selecting the final machine learning model from the training database collection. It isprocess of selecting one of the models such as Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and Extra Tree Classifier for training database and the process is capable of performing both regression and classification functions.

3.7.1 Random Forest

The three major components of Random forest algorithms are size of the nodes, the number of trees and the number of elements taken and this components need to be set before training. From there, a random forest divider can be used to solve classification or regression problem.

The random forest algorithm is combined by a set of decision trees, and each tree in the compound is made up of a data sample taken from a flexible training set, called a bootstrap sample. In that training sample, one-third of it is set aside as test data, known as a sample from the bag (oob), which will be returned later. Another random example is then injected by inserting feature bags, adding more variability to the database and reducing the correlation between decision trees. Depending on the nature of the problem, the determination of the forecast will vary. For a retreat, the trees for each decision will be weighed, and for the casting process, the majority vote — viz. the most common category variables — will produce the predicted category. Finally, the oob sample was then used to confirm the opposite, completing that prediction.

The random forest algorithm determines the result based on the prediction of decision trees. It predicts taking a measure or rate of output of various trees and is constructed of cut trees. Increasing the number of trees increases the accuracy of the result. By examining the table, it has been shown that the model accurately predicted 991 values, which means that the model has correctly classified 262 values. In addition, it found that the accuracy of the model is 88.19%.

The importance for each feature on a decision tree is then calculated using equation 1 and 2:-

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k}$$

---------- (1)

fi sub(i)= the importance of feature i. ni sub(j)= the importance of node j. These can then be normalized to a value between 0 and 1 by dividing by the sum of values of the entire important feature:

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j}$$

---------- (2)

3.7.2 Extra Trees Classifier

The Extremely Randomized Trees Classifier (Extra Trees Classifier) is an integrated learning approach that combines the effects of multiple deforestation trees collected in the "forest" to extract their subdivision effect. In a sense, it is very similar to the Random Forest Classifier and differs only in that it is the construction of decision trees in the forest. Each Decision Tree in the Extra Tree Forest is made up of the first sample of training. Then, at each test site, each tree is given a random sample of k features from the element set where each decision tree should select the best data classification based on specific computational terms. This random sample of traits leads to the formation of many unrelated decision trees. Making the selection of the feature using the above forest structure, during the formation of the forest, each element, the total standard deviation of the mathematical process used in the determination of the subdivision factor (Gini Index when Gini Index is used in forest formation) is calculated. This value is called the Gini Value feature. To select a feature, each element is ordered in a dynamic order depending on the Gini value of each element so that the end-user chooses the higher k value.

3.8 Prediction

After applying all the machine learning algorithms, we will calculate the accuracy of each algorithm and predict the new values. Prediction refers to the output of an algorithm.

3.8.1 Result

The results of our applied models are analyzed by accuracy and CV score. Confusion Matrix is used to summarize the performance of the algorithm.

Accuracy is used to determine which algorithm is best at identifying relationships. Precision is used to calculate the positive prediction made by machine learning algorithms. The Recall is used to measure how accurately our algorithms can identify therelevantdata.f1_score is used to find both poor precision and poor recall. Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set. Referring to the table 3.2 Accuracy can be calculated using below mentioned formulae.

Accuracy = TP + TN / (TP + TN + FP + FN)
---------- (1)

The simplest way to use cross-validation is to call the cross_val_score helper function on Wine quality dataset.

Table 3.3:- Comparison of the performance of machine learning algorithms.

Machine Learning Algorithms	**Accuracy Score**	**CV Score**
Logistic Regression	36.147	31.925
Decision Tree Classifier	81.059	74.76
Random Forest Classifier	88.192	82.465
Extra Trees Classifier	88.797	83.215

From Table 3.3 We can conclude that among our applied machine learning algorithms we got the highest accuracy in both Random Forest Classifier of 88.19% and Extra Trees Classifier of 88.79%.

IV Conclusion and Future Work

In this paper, Quality of the wine is accurately predicted. Four classifier's such as Logistic Regression, Decision Tree classifier, Random Forest Classifier and Extra Trees Classifier are used for the prediction of the quality. The contribution of this paper is collecting the dataset of the wine and prepared that using Machine learning algorithms. The classification model is based on the 6497 records. From the analysis, we can conclude that Random Forest Classifier and Extra Trees Classifier are better with an accuracy of 88.19% and 88.79% respectively than the other methods. This model can help people with the accurate prediction of wine quality.

In the future, this system can be implemented further using IOT to get the real time values of the wine. In the manufacturing stage, the sensors can be installed to collect information about the chemical components, temperature and the systems can therefore increase the accuracy of correctness of the results. Hence, Wine quality assessment can be done in a smart way.

V References

[1] https://www.kaggle.com

[2] Paulo Cortezl, Juliana Teixeiral, Ant'oniocerdeira. "using data mining for wine quality assessment"-2019.

[3] B Chen, C. Rhodes, A. Crawford and L.Hambuchen, "Wine in formatics: applying data mining on wine sensory reviews processed by the computational wine wheel," IEEE International conference on Data Mining workshop, 2014.

[4] K.Thakkar.Ajoshi "AHP and MACHINE LEARNING TECHINQUES for wine recommendations", International journal of computer science and information technologies, 2016.

[5]https://towardsdatascience.com/predicting-wine-quality-with-severalclassification-techniques-179038ea6434

[6] https://www.youtube.com/watch?v=W25TEa93T_I&t=13s

[7] https://scikit-learn.org

[8] https://www.geeksforgeeks.org

[9] https://vitalflux.com/python-draw-confusion-matrix-matplotlib/

[10] https://medium.com

[11] https://www.analyticsvidhya.com

[12] https://www.ibm.com/cloud/learn/random-forest