ISSN No:-2456-2165

VinQCheck: An Intelligent Wine Quality Assessment

[1]Dr. M.S. Chaudhari, [2]Kiran A. Ande, [3]Hitanshu Shahare, [4]Vaishnavi Helwatkar,
[5]Sejal Shinde, [6]Divya Janbandhu, [7]Sahil Rangari.
[1]Head of Department, Department of Information Technology, Priyadarshini Bhagwati College of Engineering, Nagpur, Maharashtra, India
[2]Assistant Professor, Department of Information Technology, Priyadarshini Bhagwati College of Engineering, Nagpur, Maharashtra, India
[3,4,5,6,7]UG Student, Department of Information Technology, Priyadarshini Bhagwati College of Engineering, Nagpur, Maharashtra, India.

Abstract:- Wine is the most popularly consumed beverage in the world and its values are considered important in society. Wine quality is always important to consumers. If the quality is not good, you have to do a different procedure from the beginning, which is very expensive. As technology has evolved, manufacturers have relied on various devices to test during the development stage. Thus, they can get a better idea about the quality of the wine, which of course saves a lot of money and time. Predicting wine quality through machine learning involves using algorithms to analyze various factors that contribute to wine quality. Therefore, for our research, we used a dataset of the Portuguese red wine grape variety "Vinho Verde" from Kaggle, which has different input variables based on physicochemical tests. We use several machine learning algorithms including Logistic Regression, SVC, Random Forest, K-Neighbor Classifier, and Decision Tree. We trained the dataset on these selected models and compared the accuracy and precision to select the best machine-learning algorithm, and we found out that the Random Forest algorithm gave the best result out of the six models respectively. Thus, helping us to predict the quality of the wine on a scale of 0-10, considering a set of characteristics. In addition, through feature selection process we observed that alcohol content greatly affects the wine quality, which was calculated using Random forest's Feature Importance attribute. We will use ANN to build the model. Furthermore, training the model on an unbalanced database leads to underestimation, especially for minority classes. Therefore, we used SMOTE to oversample the minority class in the target variable. Our research explores the potential of these key machine learning techniques to effectively predict wine quality, providing insights for wine enthusiasts and the wine industry to improve the selection and production of quality wines.

Keywords:- *Quality of Wine, Machine Learning, Random Forest Classifier, Decision Tree, Neural Network, Accuracy, Precision, Recall, F-1 Score.*

I. INTRODUCTION

There is a high increase in alcohol consumption in the world. Therefore, it has become an important issue to analyze the quality of red wine before consumption to protect human health. Given how ambitious the wine market is, the wine industry is investing in creative technologies to produce and sell wine. Technology has empowered businesses to provide consumers with high-quality wine by disclosing machine learning algorithms and data mining techniques to predict the quality of wine. Wine quality certification is important for the wine industry. Wine industry professionals and consumers are concerned about wine quality. In addition, consumers and wine professionals can benefit from predictive models that provide insight into wine quality before making purchasing decisions.

With a rich history spanning thousands of years, wine has been embraced by civilizations around the world. Wine is a complex beverage and its quality is influenced by various chemical properties such as acidity, sugar content, alcohol level, and more. The effect of these components in the wine-making process determines the final taste, aroma, and overall quality of the wine. Winemakers often look for ways to optimize these factors to produce quality wines.

As a branch of artificial intelligence (AI), machine learning plays an important role in predicting wine quality using various chemical characteristics. The ability of machine learning models to analyze and identify patterns in databases allows winemakers to improve the quality of their products. By inputting information about key chemical components such as acidity, sugar content, and alcohol level, machine learning algorithms can explore and establish the relationship between these variables and the final quality of the wine. This predictive ability allows winemakers to make informed decisions during production and optimize the composition of their wine.

Various algorithms such as decision trees, support vector machines, and neural networks have been studied in the context of machine learning models to predict wine quality. The effectiveness of this model depends on the characteristics of the database and the complexity of the relationship between chemical properties and wine quality. Random forests, a type of supervised machine learning algorithm, have shown great success in this area. By combining several decision trees, random forest can capture complex patterns and provide more accurate predictions, making it a promising option for predicting wine quality.

Ultimately, incorporating machine learning into the prediction of wine quality not only makes winemaking easier but also allows for continuous improvement. As more data becomes available and models improve, the accuracy and efficiency of predicting wine quality will continue to improve, providing winemakers with valuable insights to create superior and consistent products.

II. LITERATURE SURVEY

Title of Paper	Methods/Techniques used	Analysis and Observation
"Prediction of wine quality using machine learning", Journal of Emerging Technologies and Innovative Research, November 2021, Volume 8, Issue 11	Four classifiers such as Logistic Regression, Decision Tree classifier, Random Forest Classifier, and Extra Trees Classifier are used for the prediction of the quality.	The results of our applied models are analyzed by accuracy and CV score. Confusion Matrix is used to summarize the performance of the algorithm.
"A Machine Learning Based Mechanism For Wine Quality Prediction", Ilkogretim Online - Elementary Education Online, 2021, Vol 20 (Issue 3)	Used a handful of libraries for the proposed mechanism for better visualization and to increase accuracy to provide a well-predicted model. Then, used Random Forest to train the machine to a Model using training data and test it using test data	It implemented the mechanism via the random forest machine learning model. We have achieved 89% accuracy in the best case along with the 0.93 F1 score.
"Wine Quality Prediction using Machine Learning Algorithms", International Journal of Computer Applications Technology and Research Volume 8–Issue 09	Algorithms used for classification are: 1) Logistic Regression 2) Stochastic gradient descent 3) Support Vector Classifier 4) Random Forest.	We can see that as the quality increases of residual sugar is moderate and does not change drastically. This feature is not so essential as compared to others like alcohol and citric acid, so we can drop this feature while feature selection
"Smart Agri Wine: An Artificial Intelligence Approach to Predict Wine Quality", Journal of Computer Science 2021	In this data, there are 1599 records. Each record contains 11 different features of wine. The next step is to check for null values, duplicate values, etc. After data cleaning, data visualization helps in clearly explaining each feature in the wine dataset. A random forest classifier is used to identify patterns and relationships in features. Lastly, we train and test the model.	It is observed that the accuracy of automated wine quality prediction is more than 90% accurate and also it is in agreement with human experts 90%. To test it more rigorously, more test data will be generated by producing different varieties of homemade local fruit and flower wines.
"A data mining approach to wine quality prediction", INTERNATIONAL SCIENTIFIC CONFERENCE 15 – 16 November 2019, GABROVO	The WEKA open-source software was used for data processing (preprocessing and classification). The following classification algorithms were used to solve the classification tasks: Decision Tree, Random Forest, algorithm K star, Support Vector Machine, Multilayer perceptron, and Naïve Bayes Classifier.	In our study, 6 classification algorithms were used for wine sample classification. The model was built using each of the methods and is applied to:1. a complete wine dataset 2. white wines data set 3. red wine data set. Based on all of the above, it can be concluded that RF is the most appropriate algorithm for wine classification, and methods of wine categorization is used.

III. PROPOSED METHODOLOGY

The Outline of the methodology involves: -

- Import Datasets
- Missing Value's Imputation
- Exploratory Data Analysis
- Feature Engineering
- Model Building
- Model Evaluation

❖ ***Procedure:***

A. Dataset and Analysis:

For this research, we collected the red wine dataset from Kaggle. The red wine dataset includes 1599 samples, and each sample consists of 11 different Input variables (based on physicochemical tests) such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and Quality as the output variable, divided into 11 levels from score between 0 and 10, based on sensory data.

➢ *Import Datasets*: -

We imported the red wine dataset from Kaggle into our workspace such as Google Collab and Jupyterlab and right away, read the dataset using panda's library which is a .csv file.

B. Data Pre-processing:

Generally, Datasets from any sources are noisy, irrelevant and inconsistent and usually with lots of missing values making it difficult to generate the desired outcome. Therefore, it is extremely essential to clean the data and preprocess it to achieve quality results,

In this step, we observed our dataset very keenly and found out the shape of dataset to be (1599, 12) which consist of 1599 rows and 12 columns along with quality, then we checked the column datatypes which came out to be float except for quality which was int datatype. Thus, all input features were numerical, so we didn't need to encode anything.

➢ *Missing Value's Imputations: -*

Handling missing values is a crucial step in preparing any dataset for machine learning, as many algorithms cannot handle them directly. We Started by identifying which features in our dataset have missing values. This can be done using descriptive statistics or visualizations.

If the proportion of missing values in a particular feature or row is small, we choose to simply remove those rows or columns. However, this might result in loss of valuable information. Certainly, we find no missing values.

Table 1 Missing Values in a Table

Attribute Name	No. of missing values
volatile acidity	0
citric acid	0
residual sugar	0
Chlorides	0
fixed acidity	0
free sulfur dioxide	0
total sulfur dioxide	0
Density	0
PH	0
Sulfates	0
Alcohol	0
Quality	0

C. Data Exploration:

The red wine dataset that has been used in this paper is obtained from the Kaggle it contains a large collection of datasets that have been used for the machine learning techniques. The red wine dataset consists of 1599 instances. The datasets have 11 input variables and 1 output variable, quality (based on sensory data). Sensory data is classified in 11 quality classes from 0 to 10 (0- very bad - 10- excellent). We had used the Google Collab tool or the anaconda's Jupyterlab tool for performing the prediction and python as a programming language for our model. Collab is open-source software and it contains live code, equations, visualization, it can be used in carrying various ML techniques.

JupyterLab is an open-source, interactive web-based platform that provides a flexible environment for data science, scientific computing, and machine learning.

In the Data Exploration step, initially, we imported various python modules and libraries to build our model, such as pandas, numpy, sklearn, matplotlib and seaborn.

- Pandas is a powerful and widely used open-source data manipulation and analysis library in Python.
- NumPy libraries are used for data analyzing and numerical plotting, complex mathematical implementation.
- Scikit-learn, often abbreviated as sklearn, is a popular machine learning library in Python that provides simple and efficient tools for data analysis and modeling.
- We will be using matplotlib and seaborn as data visualization libraries.
- Seaborn is a library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

➢ *Exploratory Data Analysis: -*

Exploratory Data Analysis (EDA) is a crucial step in understanding the characteristics of your dataset before applying machine learning techniques for wine quality prediction.

We checked the distribution of each Input Features using Histplot which is used to Plot univariate or bivariate histograms to show distributions of a variable. A histogram is a classic visualization tool that represents the distribution of one or more variables by counting the number of observations that fall within discrete bins.

ISSN No:-2456-2165

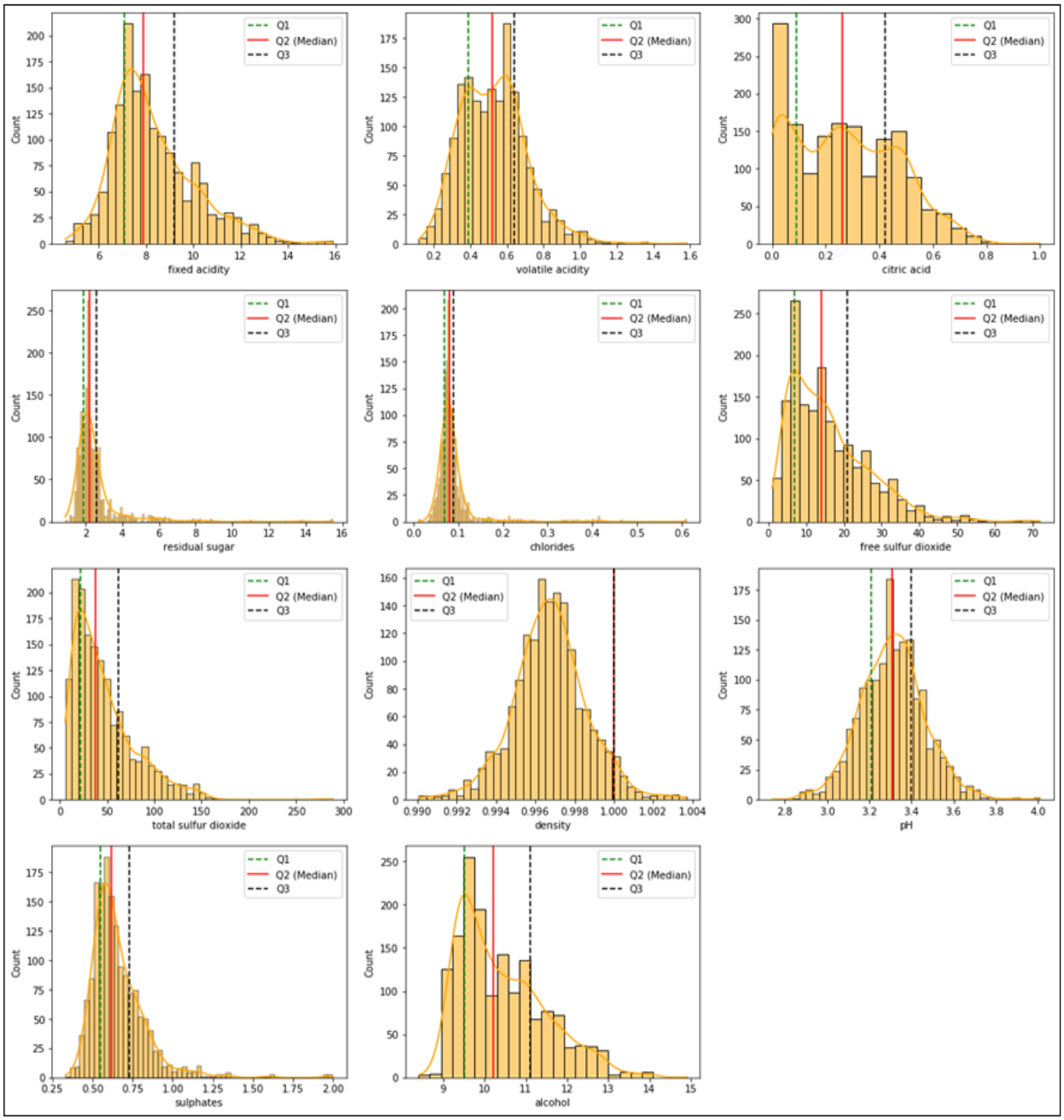

Fig 1 Univariate Analysis in Histplot

In univariate analysis, it was observed that a input features like 'chlorides', 'residual sugar', 'total sulfur dioxide', 'sulphates' are skewed due to the presence of outliers.

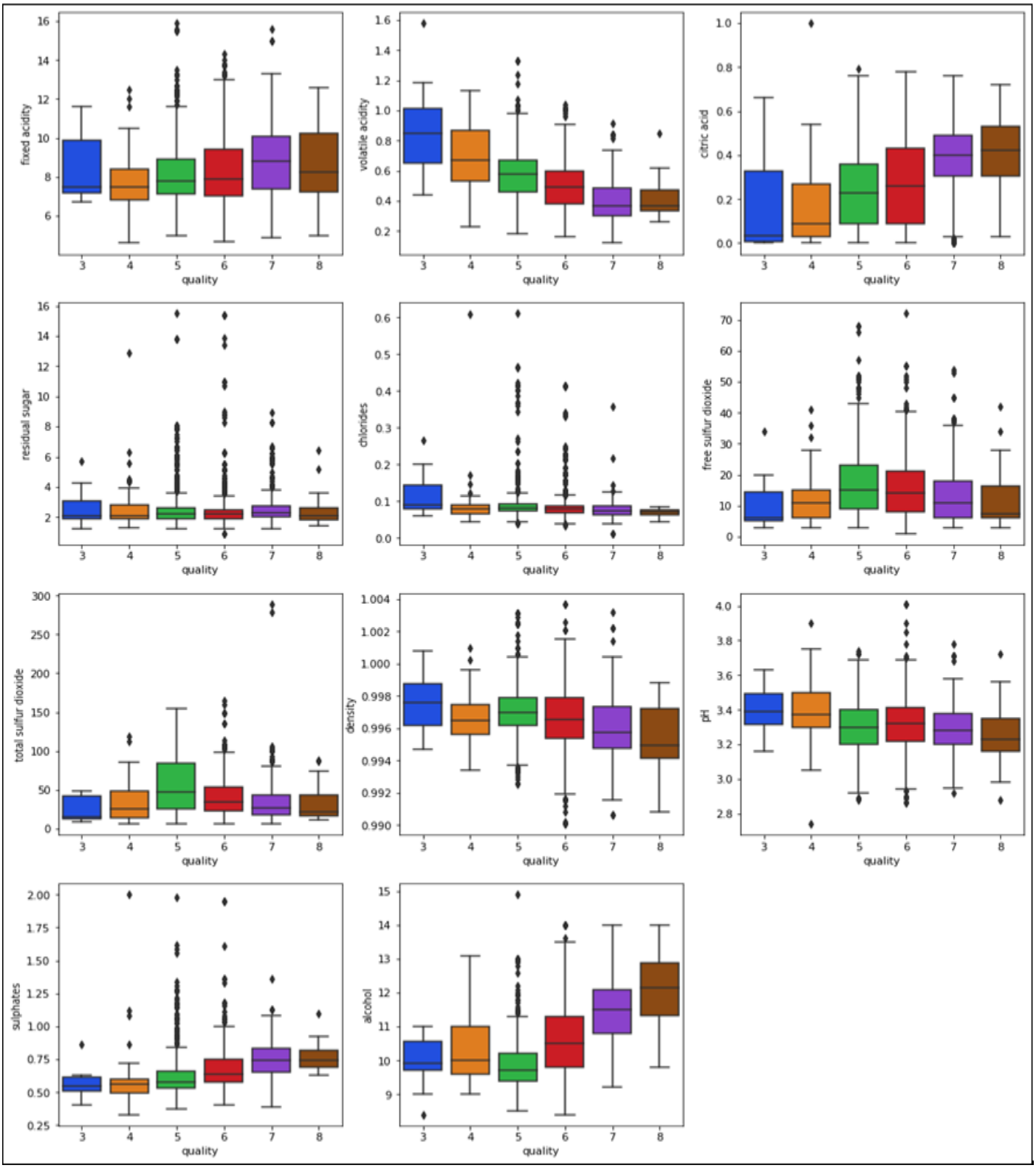

Fig 2 Bivariate Analysis in Histplot

No major pattern observed in bivariate analysis

- The distribution of 'sulphates', 'alcohol' and 'citric acid' tend to increase with increasing wine quality. So, we can say that they are positively related.
- The distribution of 'volatile acidity', 'density' and 'ph.' tend to decrease with increasing wine quality. So, we can say that they are negatively related.

➢ *Box Plots for Outliers*: -

We used box plots to identify outliers in numerical features. In descriptive statistics, a box-plot or boxplot (also known as box and whisker plot) is a type of chart often used in explanatory data analysis. Box plots visually show the distribution of numerical data and skewness through displaying the data quartiles (first, second (median), third), Maximum and Minimum values.

ISSN No:-2456-2165

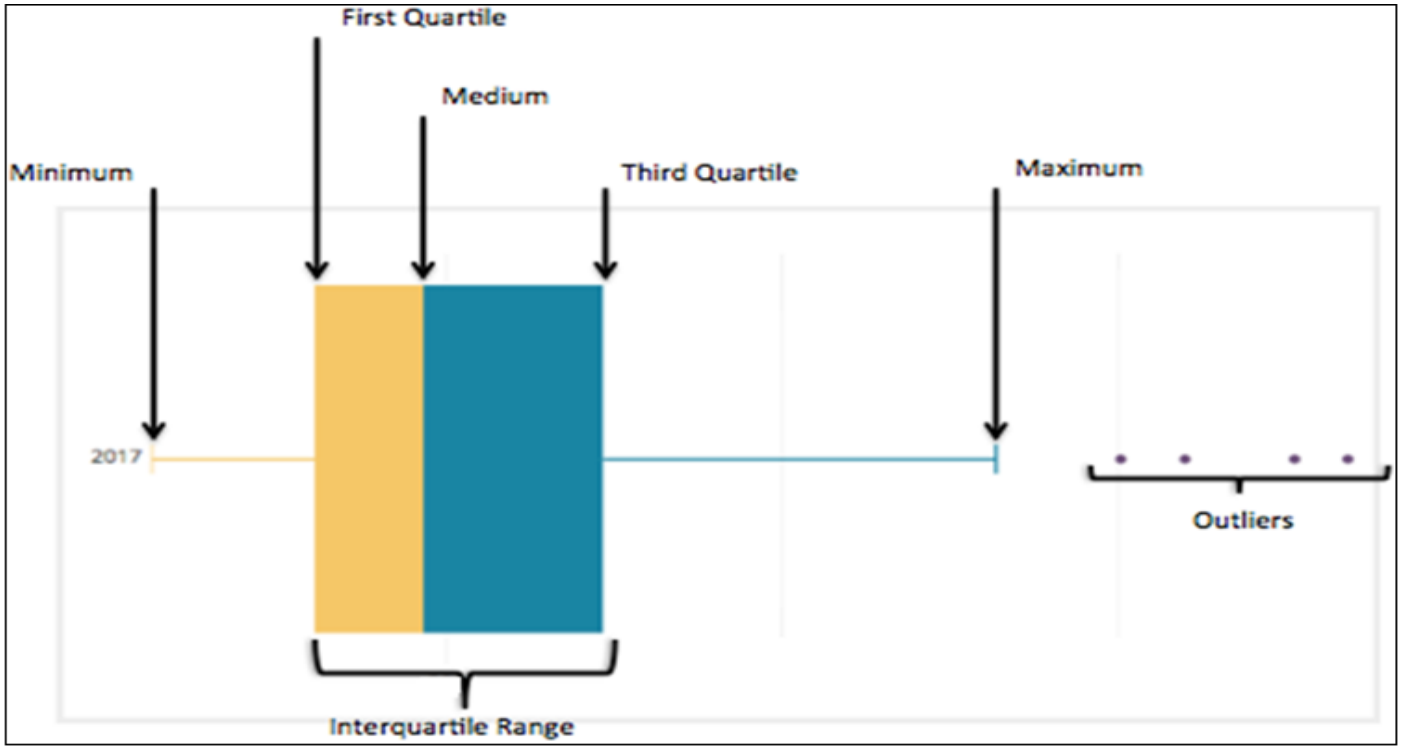

Fig 3 Box Plot

D. Feature Engineering and Feature Selection:

Feature engineering involves transforming raw data into a format that is better suited for machine learning models, improving their performance and accuracy. In the context of wine quality prediction, effective feature engineering can enhance the predictive power of your models

Feature selection is the process of selecting a subset of relevant features for use in model building, to improve the performance of the model. There are 13 features in our wine quality database, but we use 12 features.

Through this we calculated the feature importance using RandomForest's Feature Importance attribute, and found out that the Alcohol content greatly affects the wine quality. Sulphates, volatile acidity and sulfur dioxide are also important for wine quality prediction.

As the importance of citric acid and free sulfur dioxide is very low, we can drop and avoid them in model building.

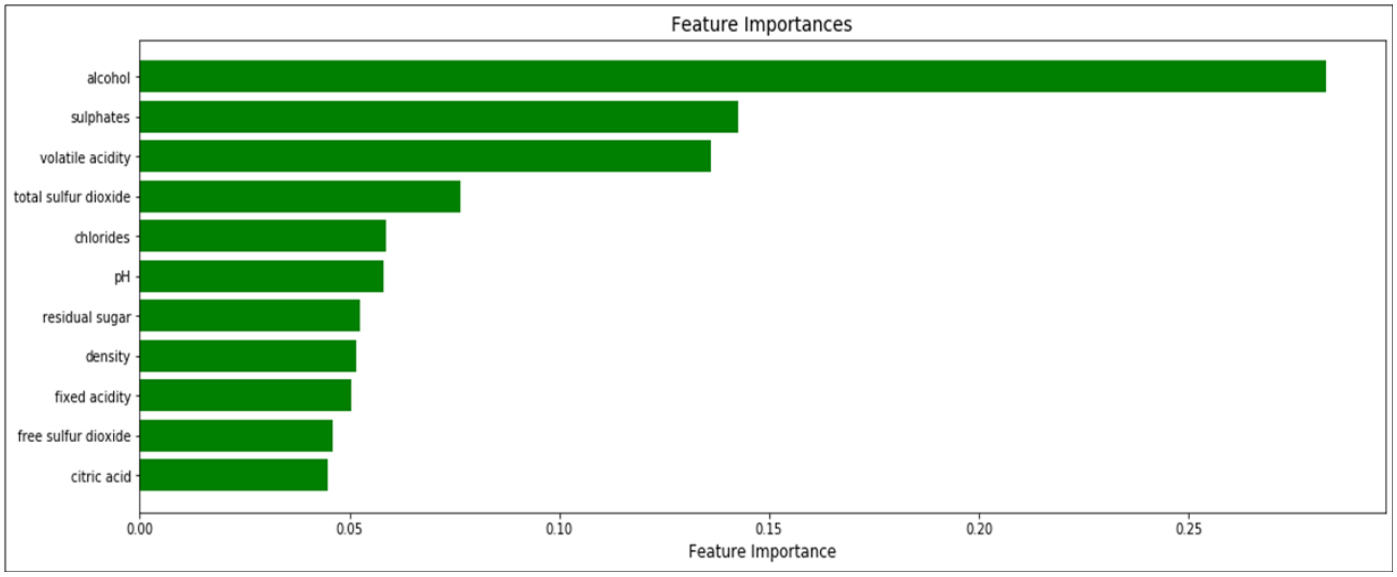

Fig 4 Feature Importances of 11 Physicochemical Variables

➢ *Correlation Analysis*

A slight multi collinearity is observed in a few features. But, the level of collinearity is not very high. So we will not drop any feature.

ISSN No:-2456-2165

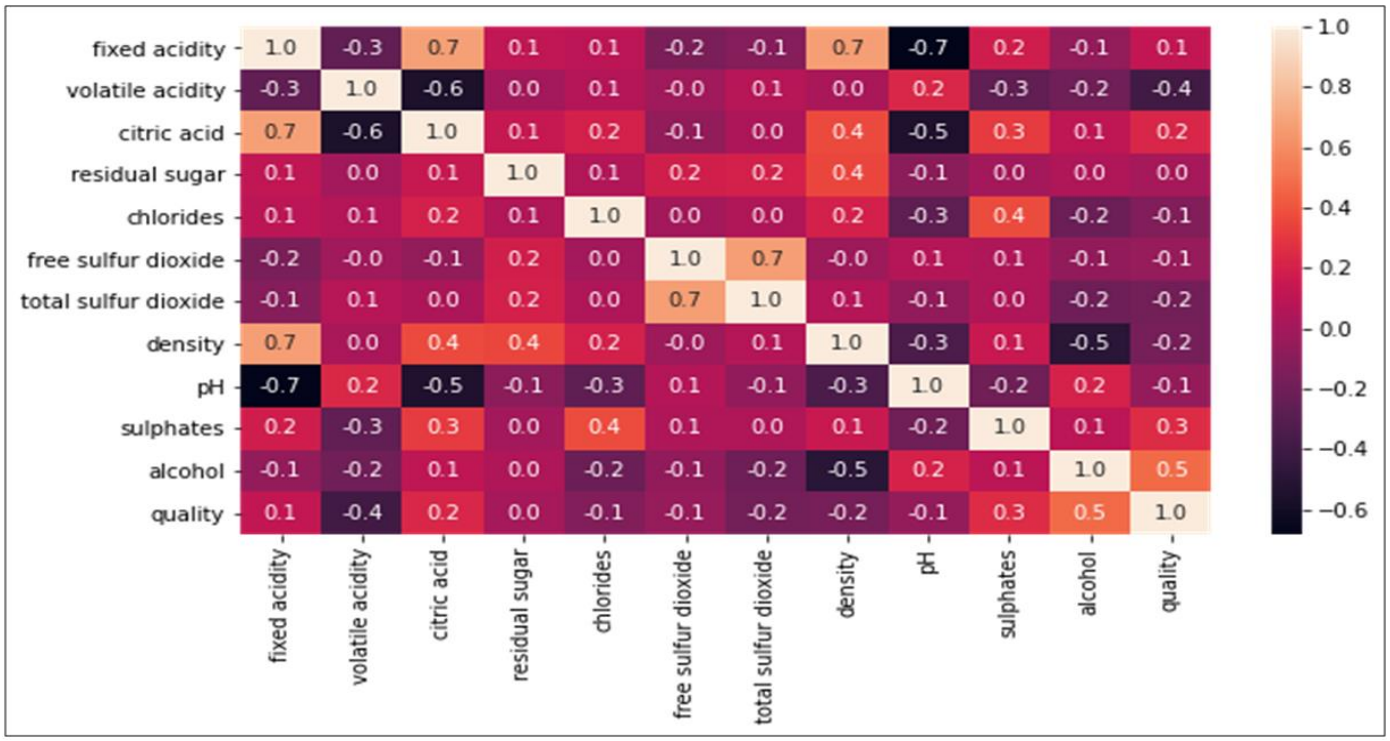

Fig 5 Correlation Matrix

➢ *Over-Sampling the Training Data*: -

Imbalanced dataset poses a challenge for predictive modeling as most of the machine learning algorithms used for classification are designed around the assumption of an equal number of examples for each class. So, training a model on imbalanced dataset results in poor predictive performance, specifically for the minority class. This is a problem because typically, the minority class is more important and therefore the problem is more sensitive to classification errors for the minority class than the majority class. For this purpose We had over-sampled the minority class (1's) in the target variable and make the number of 0's(majority class) and 1's(minority class) equal.

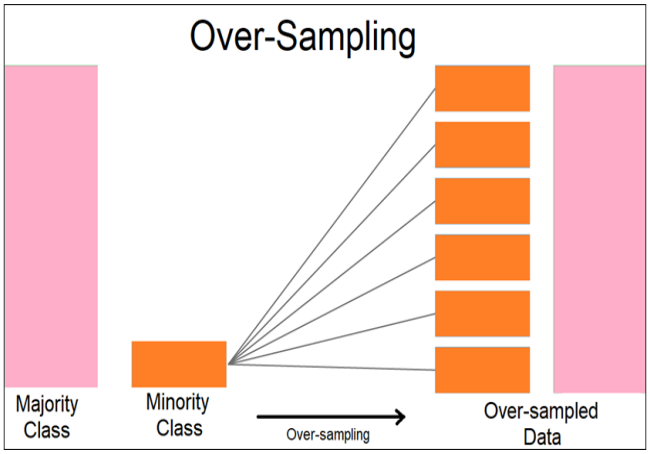

Fig 6 Over-Sampling of Minority Class

We will be using SMOTE for over-sampling purpose. It works by selecting minority samples that are close in the feature space, drawing a line between these samples in the feature space and drawing a new sample at a point along that line.

Specifically, a random sample from the minority class is first chosen. Then k of the nearest neighbors for that sample are found (typically k=5). A randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in feature space.

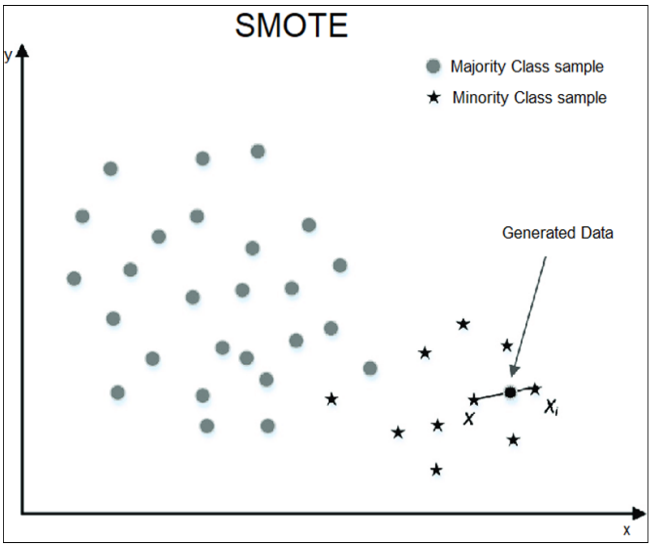

Fig 7 Synthetic Minority Over Sampling Technique

E. Model Selection:

Machine learning is divided into three main types: supervised, unsupervised, and reinforcement. We implemented supervised learning technology for the dataset used. We have two types of supervised learning: - Regression and Classification. We used several classification and regression algorithms such as Logistic Regression, Support Vector Classifier, KNN, Decision trees, Random Forest and Gradient Boosting Classifier.

ISSN No:-2456-2165

➢ *Random Forest*

Random Forest is an ensemble and supervised machine learning algorithm that is widely used for both classification and regression tasks in machine learning. It operates by constructing a multitude of decision trees during training and outputs the mode (classification) or mean prediction (regression) of the individual trees. The three main components of the random forest algorithm are node size, the number of trees, and the number of extracted elements, and these components must be determined before training. From there, the random forest partitioner can be used to solve classification or regression problems. The random forest algorithm is based on decision tree prediction. It estimates the size or yield rate of different trees and is made up of cut trees. Increasing the number of trees increases the accuracy of the results.

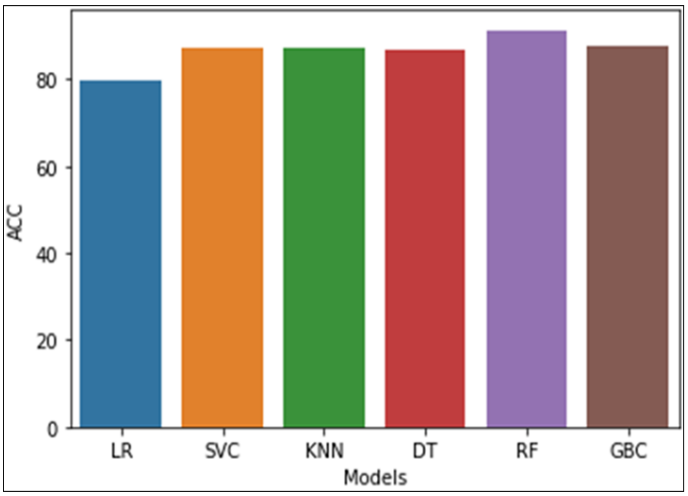

Fig 8 Accuracy Plot for Various Models

F. Model Evaluation

➢ *Model Building*: -

We will be using Artificial Neural Network (ANN) Model for this project. For Neural Network model building we used TensorFlow Library.

TensorFlow is an open-source library developed by Google primarily for deep learning applications. It also supports traditional machine learning. TensorFlow was originally developed for large numerical computations without keeping deep learning in mind. However, it proved to be very useful for deep learning development as well.

➢ *Model Prediction*: -

We saw that the model has learned to predict 0 and 1. It means it has learned the patterns present in the training data well. We also compared these values with a threshold value (for example, threshold = 0.5) and classify the output as 0 or 1 for better clarity of model predictions.

IV. RESULTS

The model performance in machine learning classification and regression models is usually measured through the following performance evaluation concepts which are accuracy, precision, recall, F1 score. These are defined as below:

➢ *Accuracy:* - Accuracy measures the overall correctness of the model's predictions.

$$Accuracy = TP+TN / TP+TN+FP+FN$$

➢ *Precision:* - Precision measures the accuracy of the positive predictions.

$$Precision = TP / TP+FP$$

➢ *Recall:* - Recall measures the ability of the model to capture all the positive instances.

$$Recall = TP / TP+FN$$

➢ *F1-score:* - F1-score is the harmonic mean of precision and recall, providing a balance between the two.

$$F1\_score = 2\times Precision\times Recall / Precision+Recall$$

Table 2 Comparison of the Performances of Six Machine Learning Algorithms

	LR	SVC	KNN	DT	RF	GBC
Accuracy	0.79.	0.87	0.87	0.86	0.91	0.87
Precision	0.76	0.82	0.80	0.83	0.87	0.83
Recall	0.82	0.88	0.86	0.89	0.93	0.89
F1 score	0.79	0.79	0.96	0.86	0.91	0.87

V. CONCLUSION

From this research, we can conclude that the random forest classifier is best in terms of accuracy and precision than the other six classification and regression models.

➢ *Key Take-Aways from EDA:*

- Based on the univariate analysis of feature, it was observed that a few Features are highly skewed which implies that extreme outliers are present.
- Based on the Bi-variate analysis, it was observed that a Few features were positively or negatively correlated with the target variable.
- The target variable 'quality' is highly imbalanced which can affect machine learning model's performance.

✓ So, for a sample of 100 red wines, model will correctly predict the wine quality for 69 wines.
✓ There were 6 classes in the target variable, out of which 3 classes had a very few records (support signifies the number of records in testing data corresponding to each class).

- ✓ The performance of model for classes 2,3,4 which had good number of records(support) in the testing data was good (>=65%) in terms of precision, recall and accuracy.
- ✓ Even after over-sampling model performance for classes 0,1 was not good.
- ✓ For class 5, even though it was minority class, model performance was decent.

REFERENCES

[1]. https://medium.com/@m.ariefrachmaann/wine-quality-prediction-with-machine-learning-model-10c29c7e3360.
[2]. https://www.ijset.in/wp-content/uploads/IJSET_V8_issue4_231.pdf
[3]. https://www.ijraset.com/best-journal/regression-modeling-approaches-for-red-wine-quality-prediction-individual-and-ensemble
[4]. https://ruomo.lib.uom.gr/bitstream/7000/1413/10/2021_IJSAMI-final_pre_print_2.pdf
[5]. https://www.irjet.net/archives/V8/i12/IRJET-V8I12183.pdf
[6]. https://unitech-selectedpapers.tugab.bg/images/papers/2019/s5/s5_p120.pdf
[7]. https://www.ijrte.org/wp-content/uploads/papers/v10i1/A58540510121.pdf.
[8]. https://thescipub.com/pdf/jcssp.2021.1099.1103.pdf
[9]. https://ijarsct.co.in/Paper2526.pdf
[10]. https://scikit-learn.org
[11]. https://medium.com/@m.ariefrachmaann/wine-quality-prediction-with-machine-learning-model-10c29c7e3360
[12]. https://www.geeksforgeeks.org
[13]. https://www.kaggle.com
[14]. https://www.ibm.com/cloud/learn/random-forest
[15]. https://ijcat.com/archieve/volume8/issue9/ijcatr08091010.pdfGaurang.