### **Wine Quality Prediction Using Machine Learning**

## **1 Introduction**

### **Overview**

Wine is the most commonly used beverage globally, and its values are considered important in society. Wine is an alcoholic drink that is made up of fermented grapes. Quality of wine is important for its consumers, mainly for producers in the present competitive market to raise the revenue. Wine quality refers to the factors that go into producing a wine, as well as the indicators or characteristics that tell you if the wine is of high quality. Historically, wine quality used to be determined by testing at the end of the production.

If you have come across wine then you will notice that wine has also their type, they are red and white wine. According to experts, wine is differentiated according to its smell, flavor, and color, but we are not wine experts to say that wine is good or bad. Every person has their own opinion about the tastes, so identifying a quality based on a person’s taste is challenging. Judging the quality of wine manually is a really tough task, even the professional wine tasters have the accuracy of 71%.

### **purpose**

In this project, we present a wine quality prediction technique that utilizes historical data to train simple machine learning models which are more accurate and can help us know the quality of wine. The models can be run on much less resource intensive environments. From this the best model is selected and saved in pkl format. We will be doing flask integration and IBM deployment.

# 2 LITERATURE SURVEY

### 

### **2.1 Exiting problem**

To predict the quality of the wine accurately,whether the taste of the wine is good or bad.

**2.2 Proposed solution**

* Data Collection.
* Collect the dataset or Create the dataset
* Data Preprocessing.
* Import the Libraries.
* Importing the dataset.
* Checking for Null Values.
* Data Visualization.
* Taking care of Missing Data.
* Label encoding.
* One Hot Encoding.
* Feature Scaling.
* Splitting Data into Train and Test.
* Model Building
* Training and testing the model o Evaluation of Model(decision tree classification)
* Application Building
* Create an HTML file
* Build a Python Code

### **THEORETICAL ANALYSIS**

### Block diagram

### 

### 

### 

### 

## 

### **Hardware and software requirements of the project**

* **Hardware**

Processor : Processor Intel CORE i3 and above

Internet Connection : Existing telephone lines, Data card, Fiber net

RAM : 4 GB

* **Software designing**

Operating System : Windows, Mac, Linux

Language : R Programming – R-4.1.1

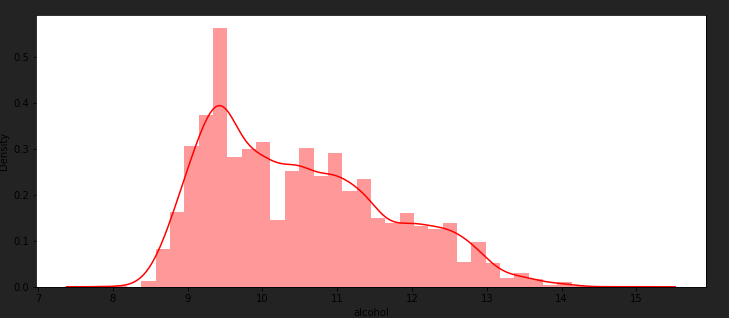
GUI : R Studio

**⦁ EXPERIMENTAL INVESTIGATIONS**

**⦁ Analysis or the investigation made while working on the solution.**

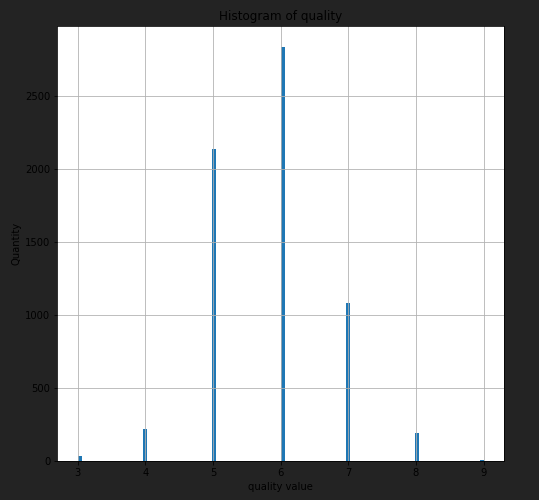
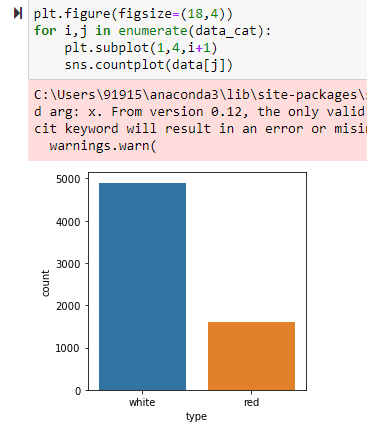
A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally.The curve through the bars increases and decreases of the observations of white and red wine.

* In our dataset we have categorical features. With the countplot function, we are going to count the unique category in that feature. With for loop and subplot we have plotted this below graph.
* From the plot we came to know, the count of white wine observations is much more than the red wine.



**Distribution plot:**

The distribution plot is suitable for comparing range and distribution for groups of numerical data. Data is plotted as value points along an axis.

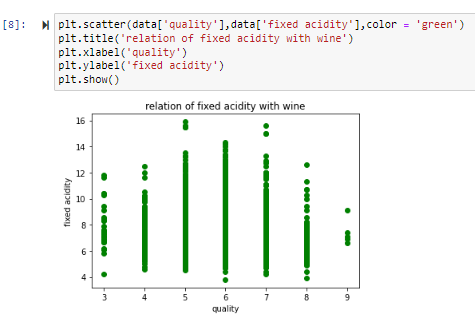


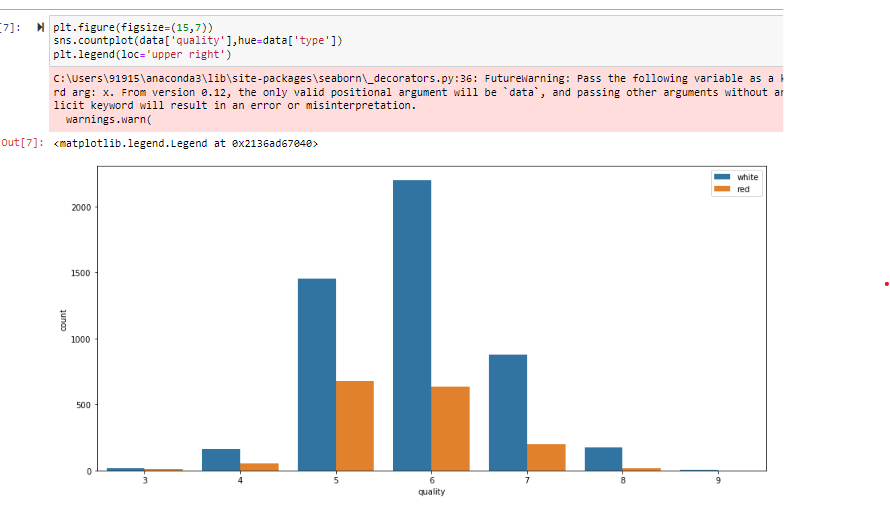
* hist() function in the pyplot module of matplotlib library is used to plot a histogram.
* As we can see, the most common vote is '6', when the lowest vote is '3', and the highest vote is '6'. In general, we may see that most of the parameters (except the "type" parameter, which is binary parameter) are normally distributed.

### 

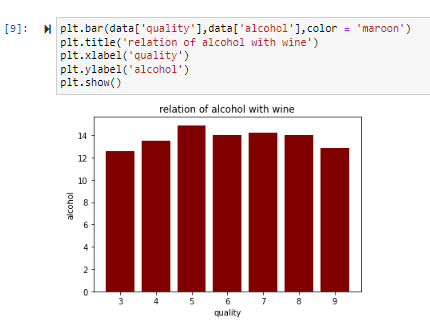
### **Bivariate Analysis**

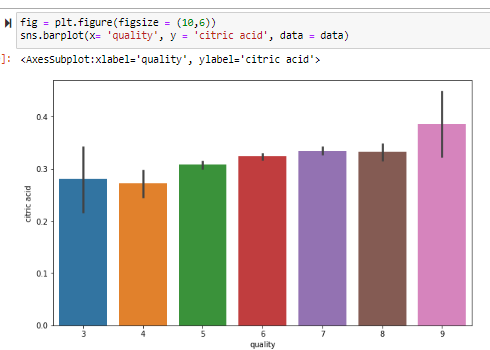
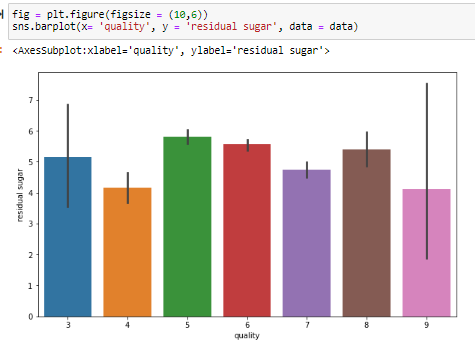
To find the relation between two features we use bivariate analysis.

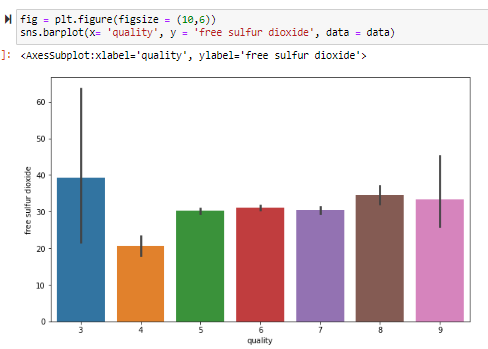
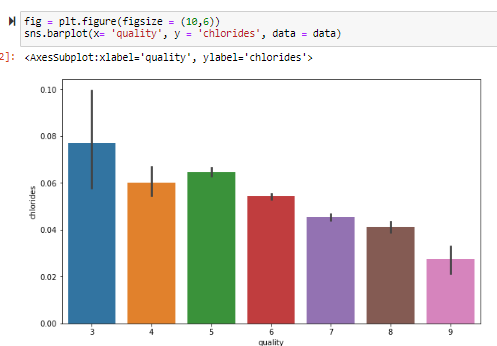
* Countplot is used here. As a 1st parameter we are passing x value and as a 2nd parameter we are passing hue value.
* From this plot we can see the relationship between type and the quality of the data

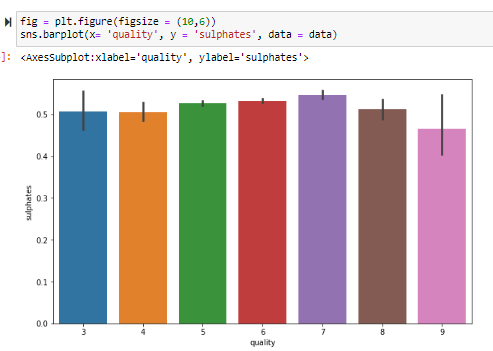


* A scatter plot is a means to represent data in a graphical format. A simple scatter plot makes use of the Coordinate axes to plot the points, based on their values. Scatter plots uses dots to represent individual pieces of data.
* The following scatter plot represents the relationship between quality and fixed acidity as a scatter plot.
* A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.
* The following visualization represents the variation of fixed acidity in the different qualities of wine.

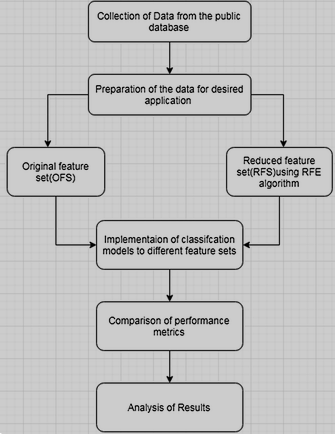






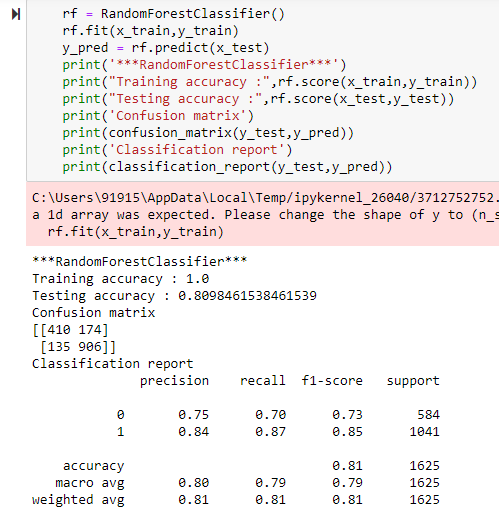


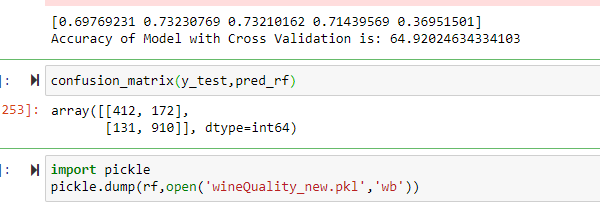
* The following visualization represents the relationship between Residual sugar and our target variable, Quality. We can see the variation in Residual sugar values over the quality.
* The following visualization represents the relationship between Chlorides and our target variable, Quality. We can see the variation in Chlorides values over the quality.
* We can see that the Composition of chloride also goes down as we go higher in the quality of the wine, so we can say that Chlorides and Quality are inversely related.
* The following visualization represents the relationship between free sulfur dioxide and our target variable, Quality. We can see the variation in free sulfur dioxide values over the quality.
* The following visualization represents the relationship between sulphates and our target variable, Quality. We can see that there’s not much variation in sulphates values over the quality.
* As we can see that like the above two items do not have very strong relation to the dependent variable, we have to showcase a correlation plot to check which of the items are more related to the dependent variable and which items are less related to the dependent variables.

**⦁ FLOWCHART**

**6 RESULT**

Final findings (Output) of the project along with screenshots.





## 

## ***7. Application***

* *By the use of several Machine learning models, we will predict the quality of the wine. Here we will only deal with the white type wine quality, we use classification techniques to check further the quality of the wine i.e. is it good or bad.*
* *We can see that good quality wines have higher levels of alcohol on average, have a lower volatile acidity on average, higher levels of sulphates on average, and higher levels of residual sugar on average.*

## ***8. Conclusion***

## *We used oversampling to balance the dataset in the data preprocessing stage to optimize the performance of the model. Then we look for features that can provide better prediction results. For this, we used Pearson coefficient correlation matrices and ranked the features according to the high correlation among the features. After applying the sampling datasets which are balancing the dataset the performance of the model is improved. In general, removing irrelevant features of the datasets improved the performance of the classification model. To conclude that the minority classes of a dataset will not get a good representation on a classifier and representation for each class can be solved by oversampling and undersampling to balance the representation classes over datasets.*

*In the classification algorithms by selecting the appropriate features and balancing the data can improve the performance of the model.*

***9 Future Scope***

*In the future, to improve the accuracy of the classifier, it is clear that the algorithm or the data must be adjusted. We recommend feature engineering, using potential relationships between wine quality, or applying the boosting algorithm on the more accurate method. In addition, by applying the other performance measurement and other machine learning algorithms for the better comparison of results.*

*This study will help the manufacturing industries to predict the quality of the different types of wines based on certain features, and also it will be helpful for them to make a good product.*

***10 References***

Chawla, N.V., 2005. Data Mining for Imbalanced Datasets: An Overview, in: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer US, Boston, MA, pp. 853–867. https://doi.org/10.1007/0-387-25465-X\_40

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J., 2009. Modeling wine preferences by data mining from physicochemical properties. Decis. Support Syst. 47, 547–553. https://doi.org/10.1016/j.dss.2009.05.016

Dahal, K., Dahal, J., Banjade, H., Gaire, S., 2021. Prediction of Wine Quality Using Machine Learning Algorithms. Open J. Stat. 11, 278–289. https://doi.org/10.4236/ojs.2021.112015

Dastmard, B., 2013. A statistical analysis of the connection between test results and field claims for ECUs in vehicles.

Drummond, C., Holte, R.C., 2003. C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling. pp. 1–8.

Er, Y., Atasoy, A., 2016. The Classification of White Wine and Red Wine According to Their Physicochemical Qualities. Int. J. Intell. Syst. Appl. Eng. 4, 23–26. https://doi.org/10.18201/ijisae.265954

Er, Y., Atasoy, Ayten, 2016. The Classification of White Wine and Red Wine According to Their Physicochemical Qualities. Int. J. Intell. Syst. Appl. Eng. 23–26. https://doi.org/10.18201/ijisae.265954

Rish, I., 2001. An Empirical Study of the Naïve Bayes Classifier. IJCAI 2001 Work Empir Methods Artif Intell 3.

says, D. Shafei, 2020. Artificial Neural Network - Applications, Algorithms and Examples [WWW Document]. TechVidvan. URL

https://techvidvan.com/tutorials/artificial-neural-network/ (accessed 6.6.21).

scikit-learn, developer, 2021. scikit-learn: machine learning in Python — scikit-learn 0.24.2 documentation [WWW Document]. URL https://scikit-learn.org/stable/ (accessed 5.31.21)

Wolf, L., Shashua, A., 2005. Feature Selection for Unsupervised and Supervised Inference: The Emergence of Sparsity in a WeightBased Approach 33.

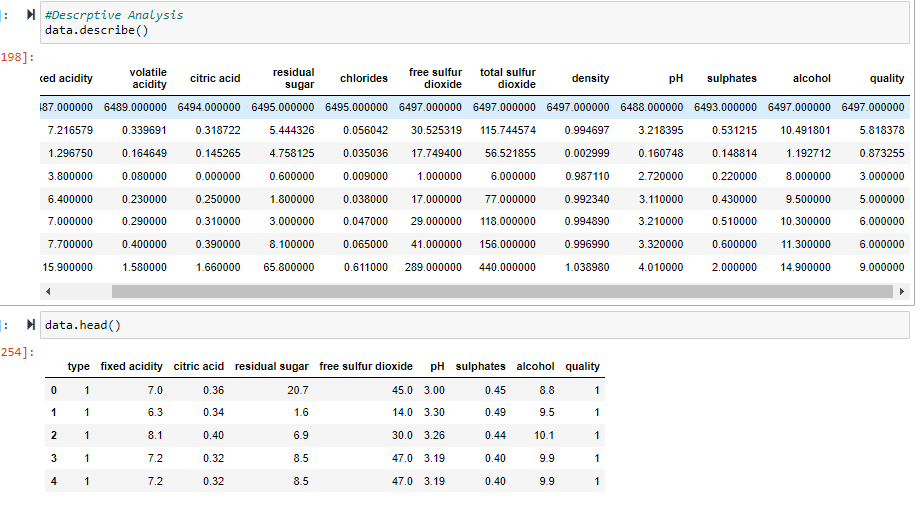
Yu, T., Zhu, H., 2020. Hyper-Parameter Optimization: A Review of Algorithms and Applications. ArXiv200305689 Cs Stat. Zhang, P., 2000. Neural Networks for Classification:

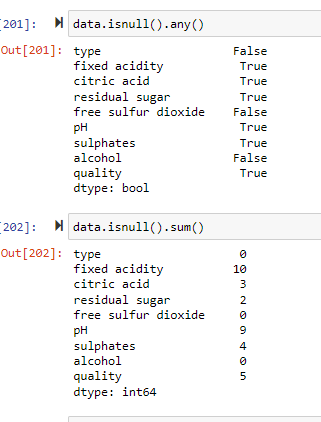
A Survey. Syst. Man Cybern. Part C Appl. Rev. IEEE Trans. On 30, 451–462. https://doi.org/10.1109/5326.897072

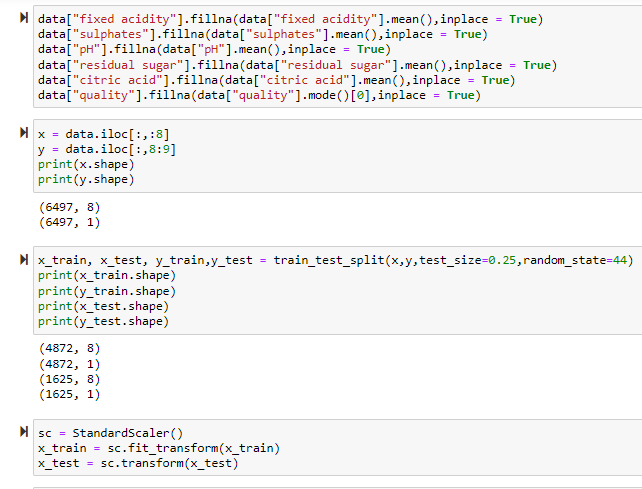
***11 Appendix***

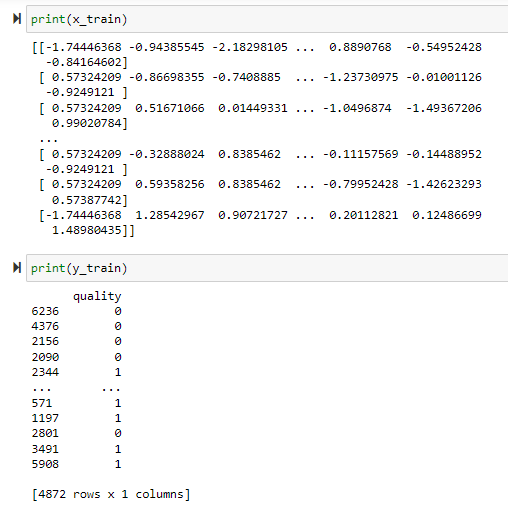
**11.1 Dataset and Importing Libraries**



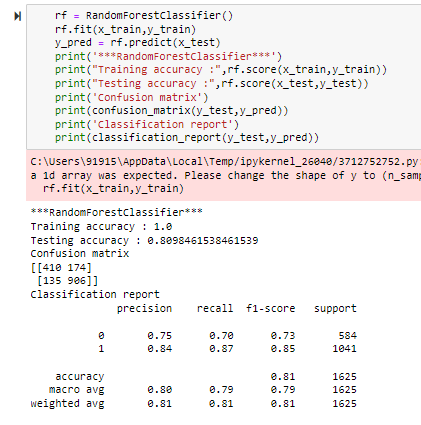
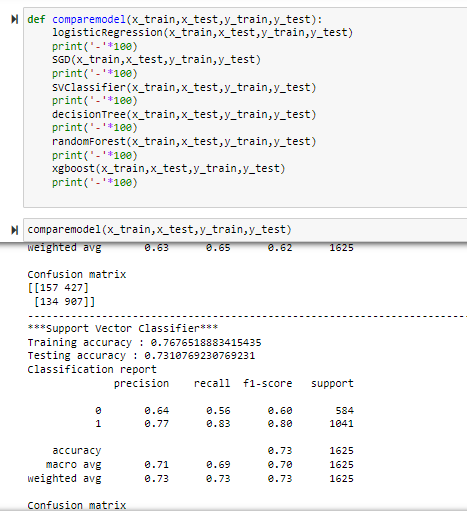
***11.2 PreProcessing***



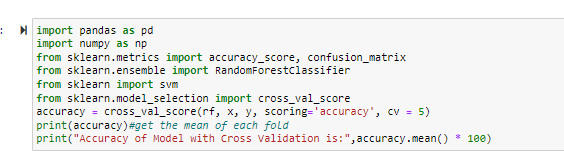
***11.3 Training Model***

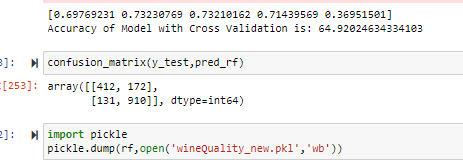






,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,

***11.4 Evaluation Metrics***



***12.5 Flask Implementation***

*Wine quality prediction*