A Project Report in partial fulfillment of the degree

# Bachelor of Technology

in

# Computer Science & Engineering/ Electrical & Electronics Engineering

## By

2103A51291 SAHITHI SIRIKONDA

2103A51269 SUHANI KUSUMA

2103A51127 FARIHA

**Under the Guidance of**

# D. Ramesh

**Submitted to**



# DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

## S R UNIVERSITY, ANANTHASAGAR, WARANGAL





**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**CERTIFICATE**

This is to certify that the Project Report entitled “WINE QUAILTY PREDICTION” is a record of bonafide work carried out by the student(s) Sahithi , Suhani , Fariha , bearing Roll No(s) 2103A51291 ,2103A51269, 2103A51127 during the academic year 2022-2023 in partial fulfillment of the award of the degree of ***Bachelor of Technology*** in **Computer Science Engineering/ Electrical & Electronics Engineering** by the SR University, Warangal.

.

# Supervisor Head of the Department

**External Examiner**

# ABSTRACT

Nowadays people try to lead a luxurious life. They tend to use the things either for show off or for their daily basis. These days the consumption of red wine is very common to all. So it became important to analyze the quality of red wine before its consumption to preserve human health. Hence this research is a step towards the quality prediction of the red wine using its various attributes. Dataset is taken from the sources and the techniques such as Random Forest, Support Vector Machine and Naïve Bayes are applied. Various measures are calculated and the results are compared among training set and testing set and accordingly the best out of the three techniques depending on the training set results is predicted. Better results can be observed if the best features out from other techniques are extracted and merged with one another to improve the accuracy and efficiency.

# Table of Contents

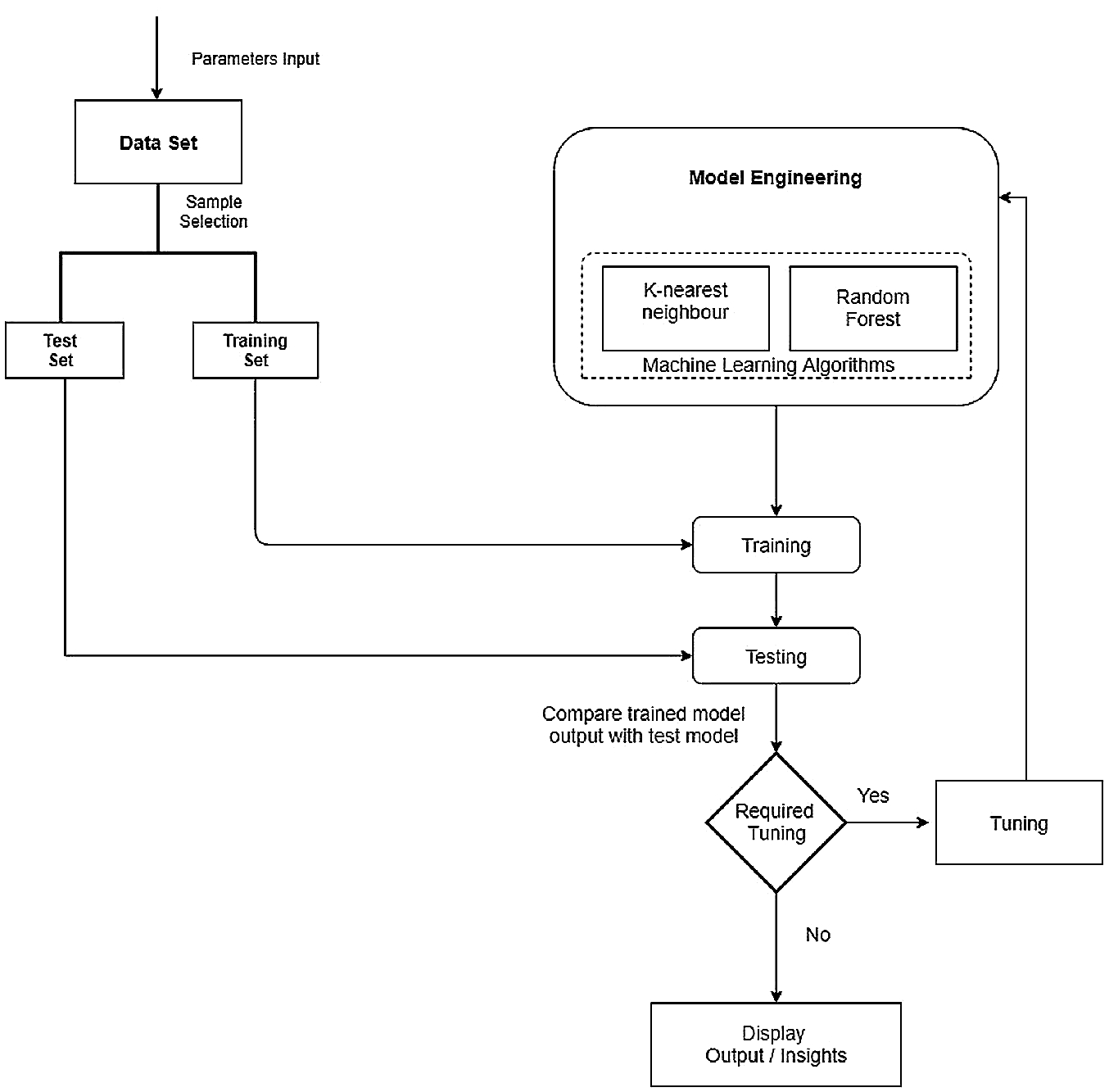
|  |  |  |
| --- | --- | --- |
| **S.NO** | **Content** | **Page No** |
| 1 | Introduction | 1 |
| 2 | Literature Review | 3 |
| 3 | Design | 5 |
| 4 | Dataset | 8 |
| 5 | Data Pre-processing | 19 |
| 6 | Methodology | 20 |
| 7 | Results | 26 |
| 8 | Conclusion | 27 |
| 9 | References | 28 |

1. **INTRODUCTION:**

In the wine industry, Product Quality Certification is used to promote their products and become a concern for every individual who consumes their products. With such a high demand for wine, it is impossible to ensure its quality by experts, since this would increase costs. By selecting the important parameters, users can build a machine learning model which predicts the quality of the wine using machine learning techniques.

As well as consumers, the wine industry places a high importance on its quality. Certifications are helping increase sales in many industries. An extensive team of human experts is required to review a product, which is an expensive and time-consuming process. To replace human experts, machine learning models are applied. Using machine learning techniques, we can build a model that can predict the quality of wine by selecting the parameters. The quality of wine is determined by factors such as Fixed acidity, Volatile acidity.

It is subjective to evaluate wine quality by human experts, but not by a machine learning model that has been trained to detect wine quality. Wine experts follow their personal preferences, but ML models provide accurate predictions in a more objective way. Machine learning systems are led by humans, but the best results are determined by the right input data (labelled, of course).



* 1. **OBJECTIVES:**

In this mini project, there are two objectives. This can be listed as below:

1. The major goal of this project is to develop classification models that can determine whether a wine is of good quality or not using machine learning techniques.
2. Identifying the characteristics of a good wine will be the objective.

# LITERATURE REVIEW

Based on its composition, it has been used to predict the quality of red wine. To make the prediction, they used random forests, support vector, machine, and naive Bayes techniques. To measure performance, they used this formula: precision, recall, f1-score, accuracy, specificity, and misclassification error. They achieved the best result from the support vector machine as compare to the random forest and using naive Bayes techniques, the accuracy of the support vector machine technique is 67.25%

The algorithm has used various machine learning algorithms such as linear regression, neural network, and support vector machine techniques to extract important characteristics of red wine and white wine quality. There are two methods they used. In the first case, the dependence between the independent variable and the target variable and in the second case, predicting the value of the target variable and concluding that all features are not necessary in order to predict the wine quality, as opposed to only selecting the necessary attributes.

A variety of machine learning models, including rigid regression, support vector machines, gradient boosting regressors, and multilayer artificial neural networks, have been applied to predict the wine quality. In their analysis, they found that the gradient boosting regressor performed the best compared to other models with the MSE, R, and MAPE values of 0.3741.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SI NO DATE OF AUTHORS NAME METHADOLOGY DATASET ACCURACY  PUBLICATION | | | | | | |
| 1 | 29 July, 2021. | ArielZeev | classification-comments-instagram | various NLP techniques | LIAR ISOT | 60.8 98 |
| 2 | April, 2018. | Daniar Nur Amin1 , Ednawati Rainarli2 | Classification of instagram content based on comments using SVM | SVM | GIT HUB -  Labelled as authentic. | 98.4 |
| 3 | 25 July 2019 | Beta Priyoko; Ainul Yaqin | Implementation of Naïve based algorithm for spam comments | Naïve based algorithm | FA-KES ISOT | 60 99 |
| 4 | July, 2019 | Nur Azizul Haqimi | Detection of Spam Comments on Instagram Using Complementary Naïve Bayes | Naive Bayes | online media. | 92 |
| 5 | April, 2019 | Nur Rokhman2, Sigit Priyanta | Spam Comments on Instagram | SVM | From Kaggle | 87 |
| 6 | June, 2017. | Andrew Hutchinson | Toxic Comments | lSTM | From Kaggle | 93.2 |
| 7 | July, 2019. | Nabanita | MAX-Toxic-Comment-Classifier | NLP alogorithms | Kaggle | 95 |
| 8 | October, 2021. | Neha Narwal | Detecting and classifying toxic messages | Multilayer perceptron | Jigsaw | 91.8 |

# DESIGN:

* 1. **Requirement Specifications (S/W & H/W)**

## Hardware Requirements

Processor : Min. Core i5 processor

RAM : 8GB (Min.) and more

Hard Disk Space : 50GB+

CPU or GPU (GPU is recommended)

## Software Requirements

Operating System : Windows 7 or later versions of windows

Notebooks : Jupiter Notebook, Google Collab

Business Tier : Python

Model : Random forest classifier

**FLOW CHART**

CREATE A DATASET

PREPROCESS THE DATA

BUILD A DATASET CLASS

BUILD A MODEL

TRAIN A MODEL

EVALUATE THE MODEL FOR TEST DATA

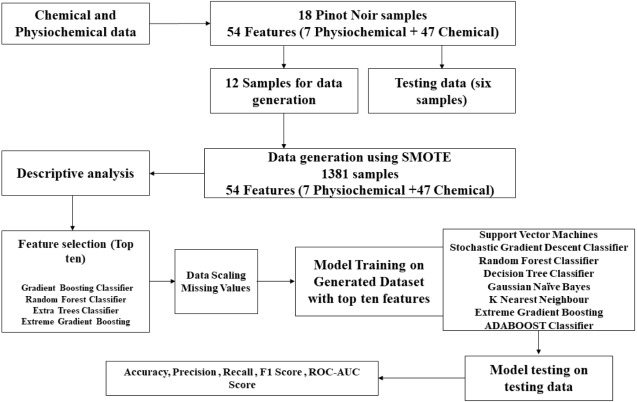
CREATE A RANDOM FOREST CLASSIFIER

TAKE ALL THE VALUES FROM THE DATASET

FIND THE VALUES WHICH PREDICTS THE GOOD QUALITY

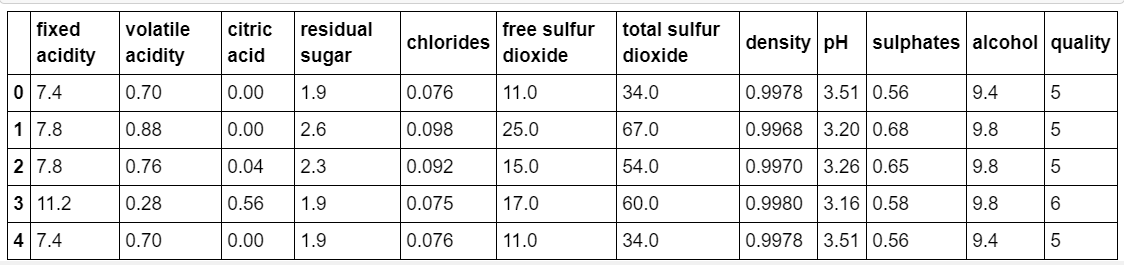
TRAIN A MODEL

MEASURE THE QUALITY OF THE WINE



# DATASET:

The data set for building the classification model was acquired from the competition site and it included the training set as well as the test set. The steps elaborated in the workflow below will describe the entire process from Data Pre-Processing to Model Testing. There are 1500 samples. The input is comment\_text and the output Label is a model which predicts a probability of each type of toxicity for each comment.



# 5.DATA PREPROCESSING:

After collecting and preparing wine data, a simple classification approach is determined using machine learning methods. When given some input data, a machine learning classifier attempts to predict whether it is low-quality wine, mediocre wine, or high-quality wine. The Random Forest Classifier gives us the best accuracy score among the multiple Regression and Classification models, so we will use it.

## ALGORITHM:

## In this project, the text classification is done based on the dataset which is in the form of CSV, it has different texts which are labelled under Fixed acidity,Volatile acidity, Citric acid, Residual sugar, Chlorides, Free sulfur dioxide, Total sulfur dioxide , Density, pH, Sulfates.

## Random forest: credit: github user kjw0612

Fig.5.1 Random forest classifier

**5.2. PREDICTION USING RANDOM FOREST CLASSIFIER:**

Among the input variables that determine the quality of a wine are Fixed acidity, Volatile acidity, pH, and Sulfates. Gathering and preparing wine data is the first step, followed by determining the most accurate and effective classification approach with ML methods. Analyzing certain input data, the machine learning classifier predicts outcomes whether it is low-quality wine, mediocre wine, or high-quality wine. To check the accuracy score, we apply multiple Regression and Classification models. The Random Forest Classifier has the best accuracy among all of the models, so we will use it. The random forest builds multiple decision trees and merges them to produce a more accurate and stable prediction.

While the trees grow, the random forest adds more randomness to the model. While splitting a node, it looks for the best feature among a random subset instead of looking for the most important feature.

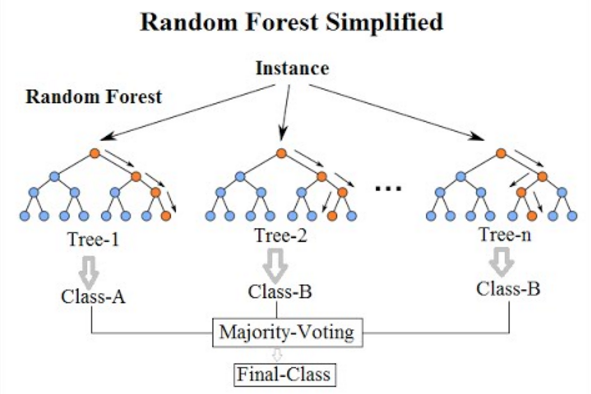
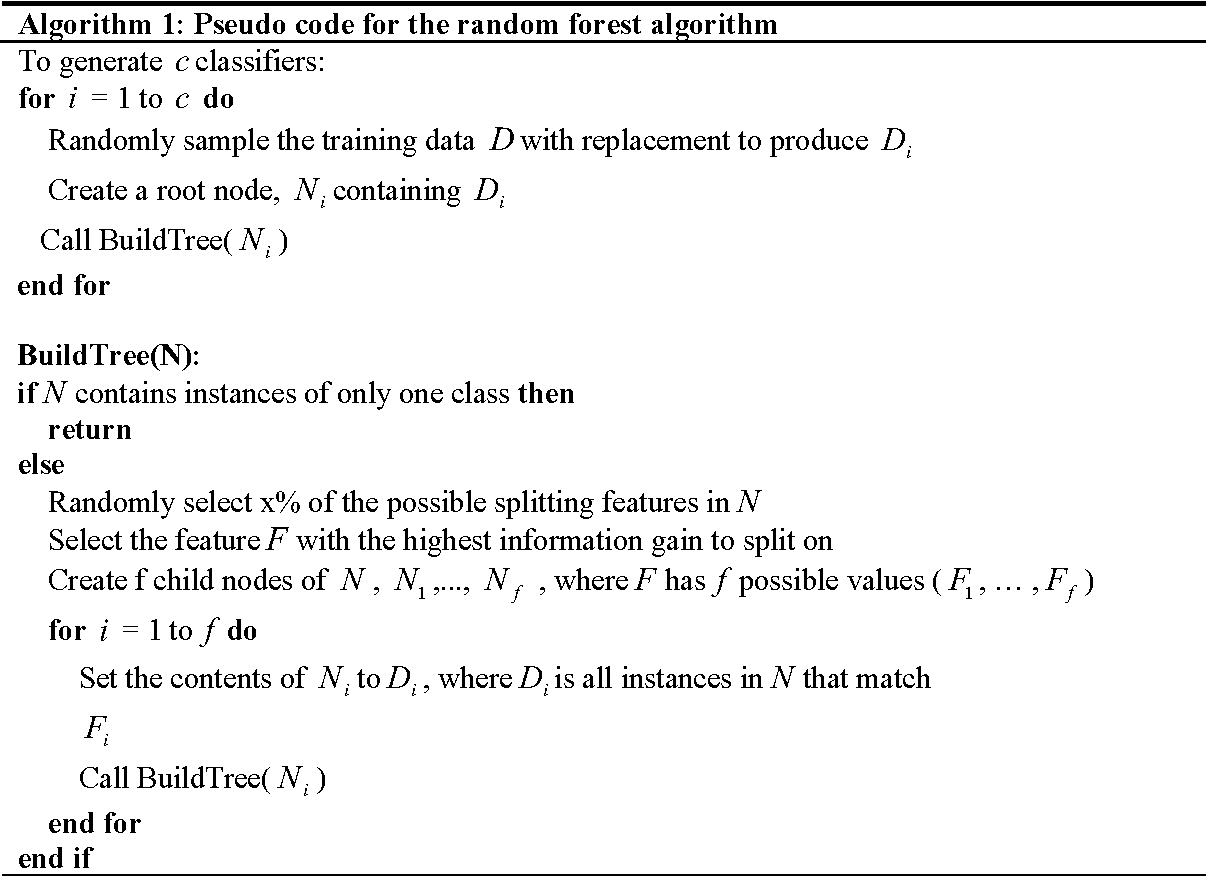


Fig.3.1.1 Random forest Tree

**5.3 ADVANTAGES OF RANDOM FOREST MODEL:**

* Machine learning models can replace human tasks and predict the quality with accuracy.
* Using these machine learning models is more efficient and less time consuming than using the traditional method
  1. **DISADVANTAGES OF RANDOM FOREST MODEL:**
* No capability to handle missing data
* Requires higher space complexity and time complexity



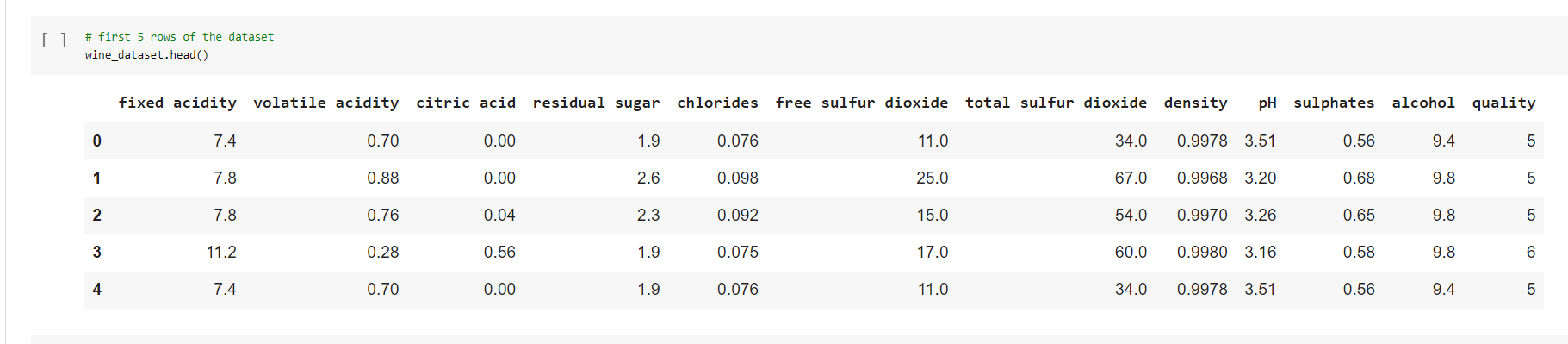
Algorithm For Random forest classifier

## First, we need to import necessary libraries for the project.

## Screenshot (68)

Fig 4.1.1 Importing the dependencies

**DATA DESCRIPTION:**

****

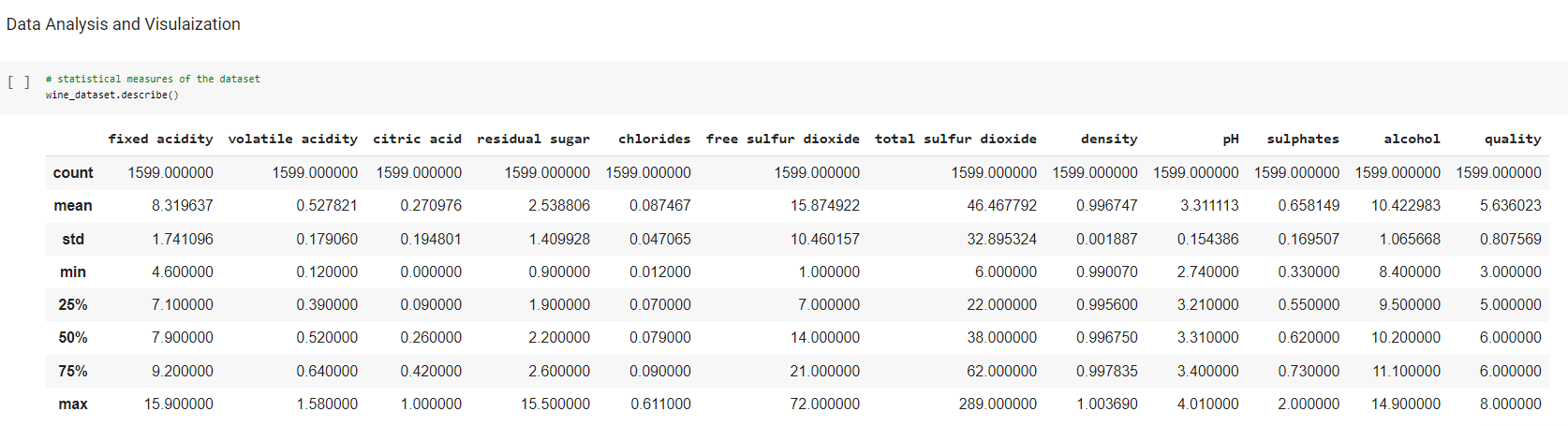
****

Fig 4.2 Data Analysis

**DATA ANALYSIS AND VISUALISATION:**

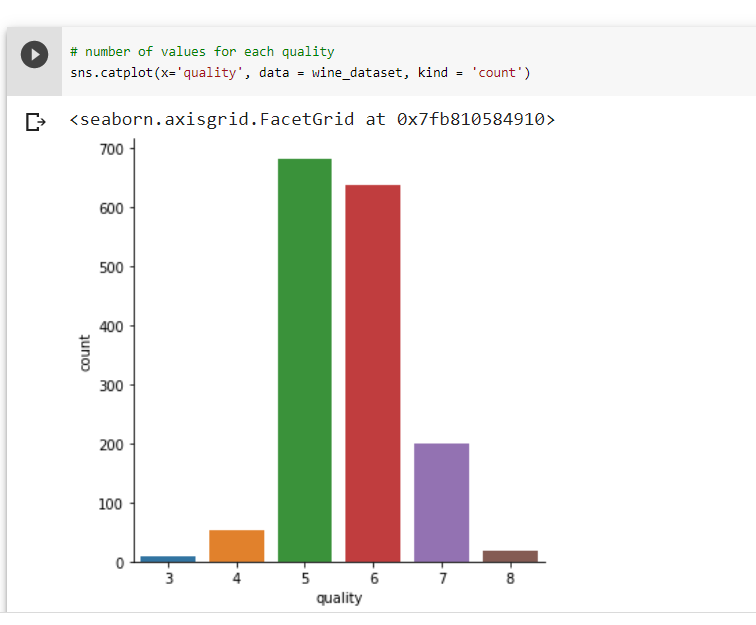
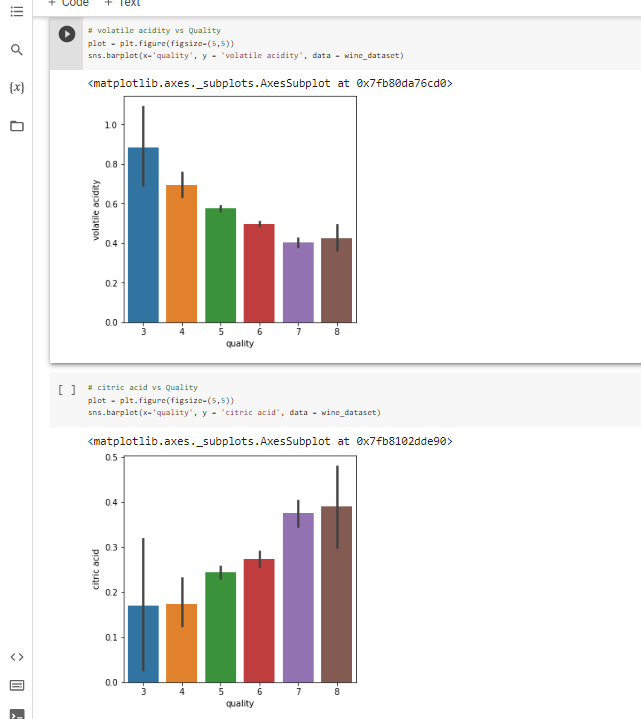
****

Fig 4.3 Values count

****

**CORRELATION:**

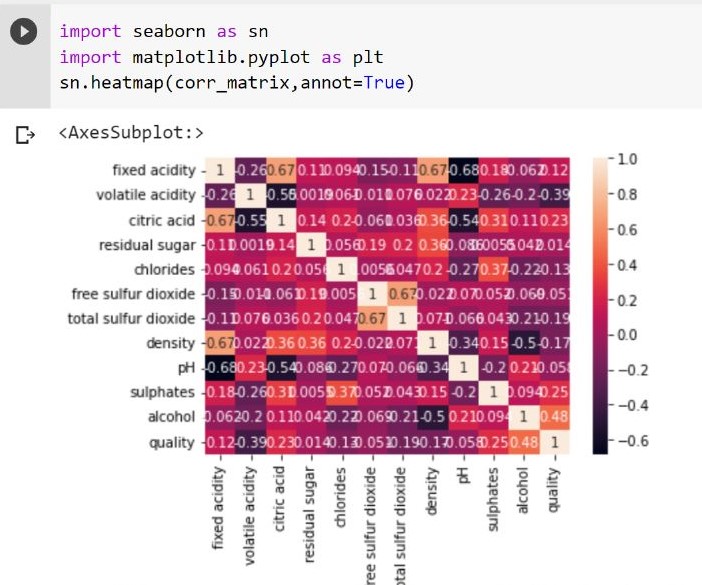
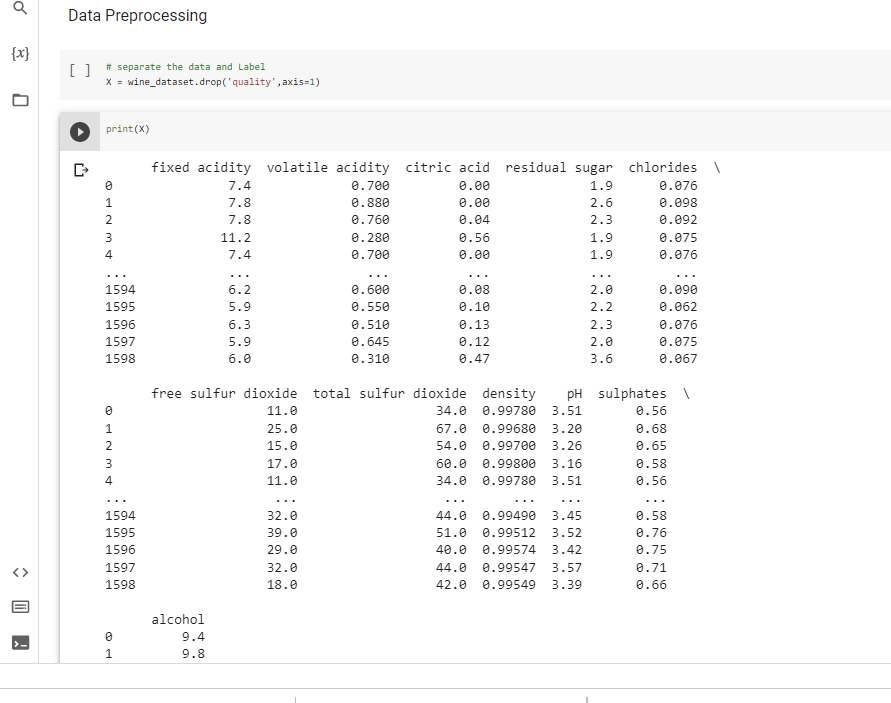
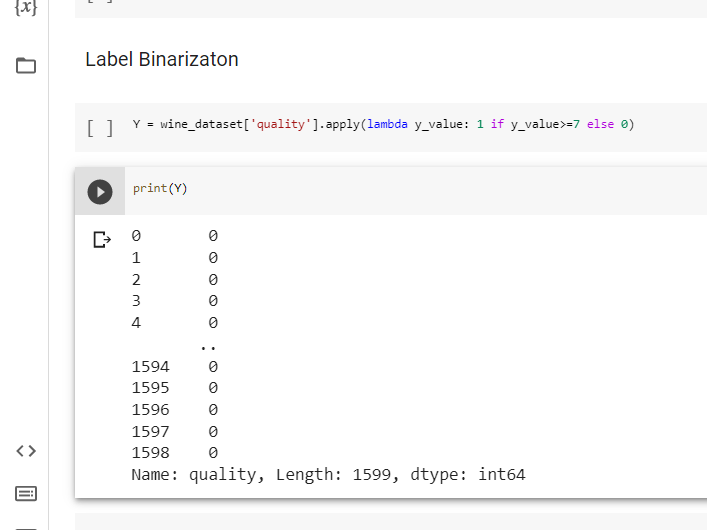


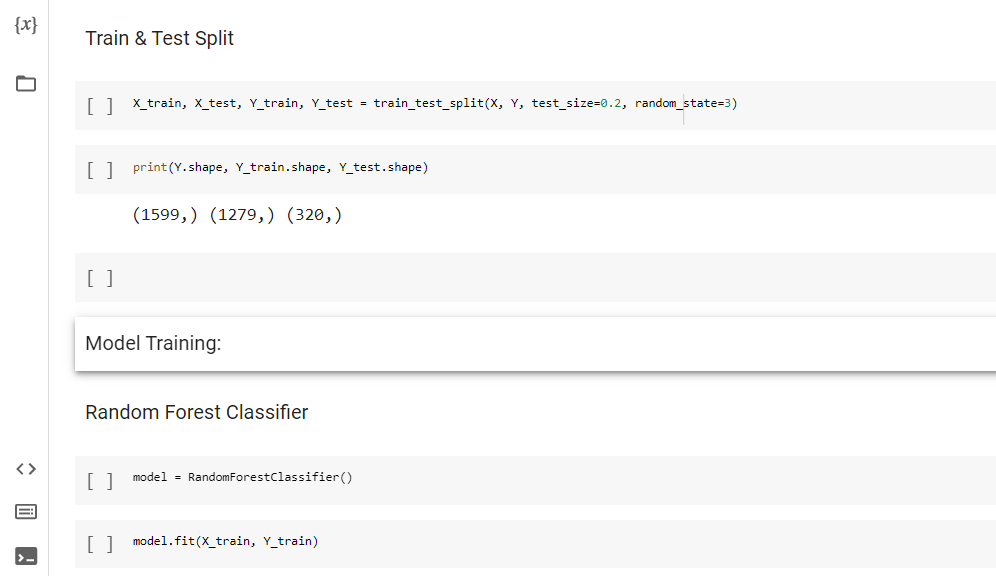
Fig 4.4 Correlation

**DATA PREPROCESSING**

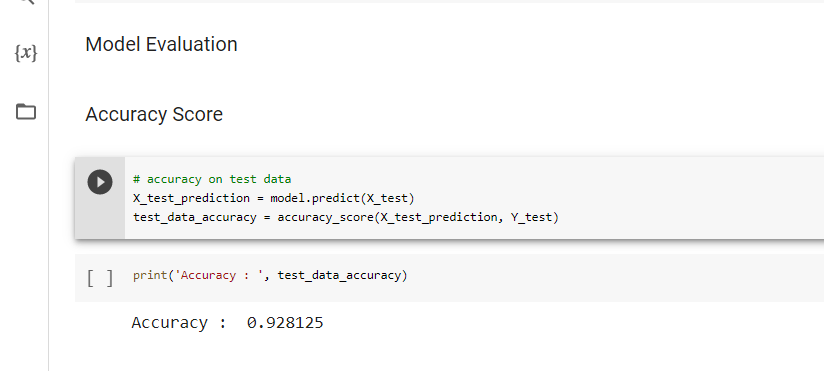
****

****

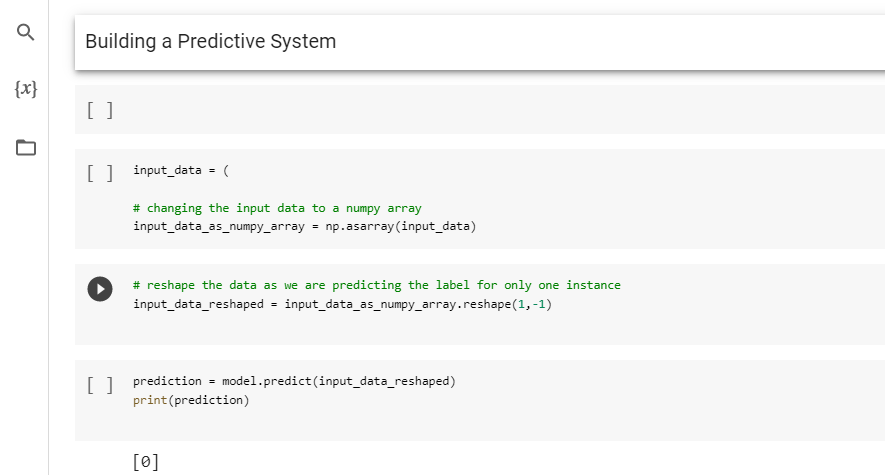
**TRAIN AND EVALUATE THE MODEL:**

****

Train the model

****

Evaluate the model

****

Building the predictive system

**6.METHODOLOGY:**

* First, we need to take the input should be a sequence of tokens and the output will be a vector of size the number of classes in our classification task.
* Next, we need to create a dataset which should be in CSV format and labeled under one of 5 categories: entertainment, sport, tech, business, or politics.
* Write an algorithm to preprocess the data, create a dataset class for our dataset, model building, an algorithm to train the model and evaluate model on test data.
* Software and hardware will be selected based on requirements.
* Implementation of the algorithm will be done with the dataset.
* Finally, a report will be prepared and uniqueness of report (plagiarism) will be checked.

**THE ALGORITHMS USED HERE ARE**

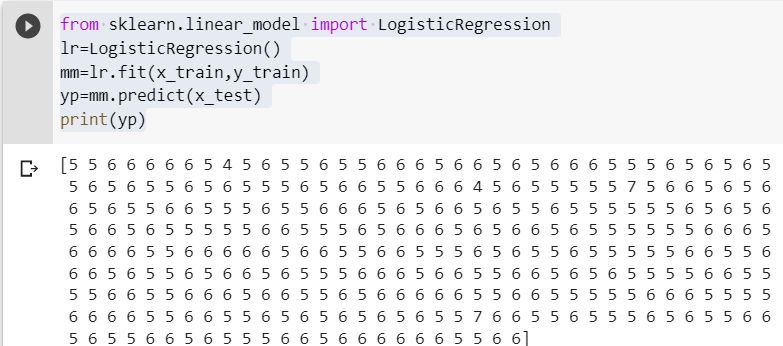
**1.LOGISTIC REGRESSION:** It is supervised learning, it is a classification algorithm, used when the value of the target variable is categorical in nature. Logistic regression is used when data in question has binary output. It uses a complex cost function called sigmoid function.

**ADVANTAGES**

* More simple to use than remaining algorithms.
* Predict possibilities of binary outcomes.

**DISADVANTGES**

* Limited to binary classification.
* Requires sufficient data, unlike other algorithms.



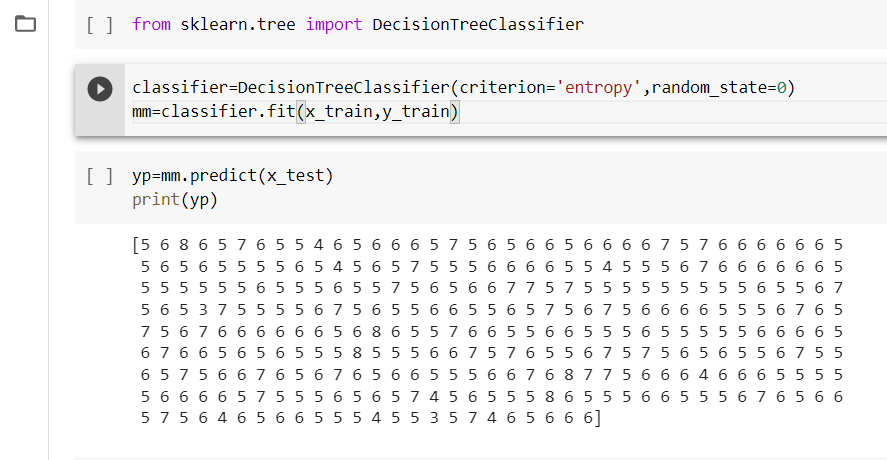
**2.DECISION TREE**: Decision tree is a supervised learning algorithm which is used in both classification and regression. This algorithm is used to divide the dataset into a tree structure and constructed into nodes and sub nodes. Nodes are created based on some features. This algorithm is widely used in machine learning because it is easy to use. The goal is to predict the target variable by using some rules. The deeper the tree, more complex the decision rules. It has 2 attributes ENTROPY and GINI INDEX.

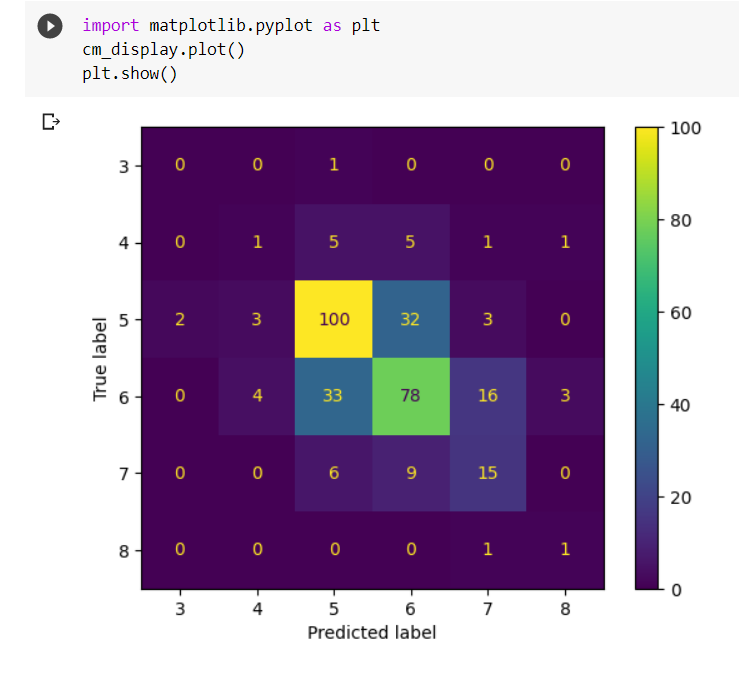
**ADVANTAGES**

* Can handle both numerical and categorical data
* Easy to handle
* Possible to account reliability of model

**DISADVANTAGES**

* If any mistake has taken place it results to different tree.
* Less effective in predicting the outcome of a continuous variable.



****

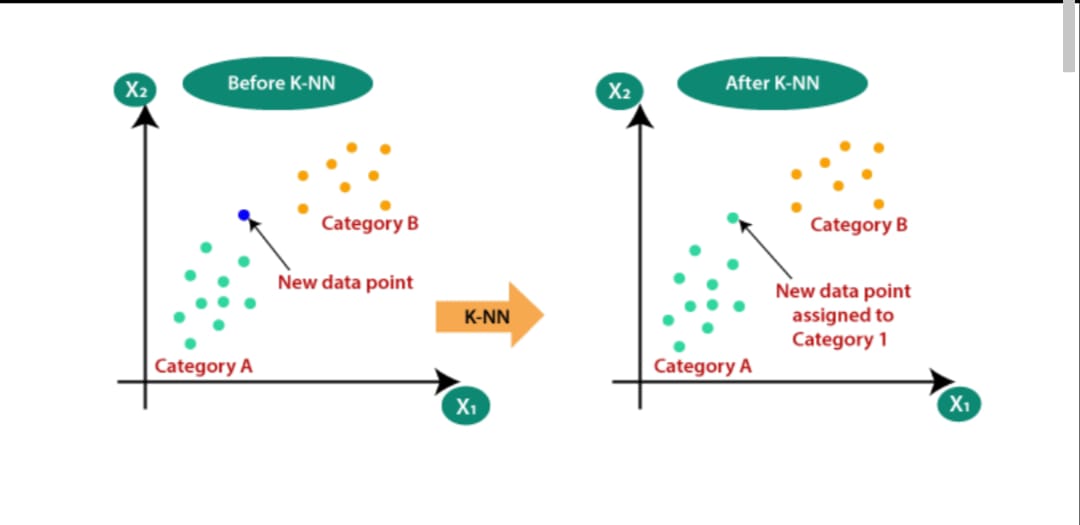
**3.K-NEAREST NEIGHBOUR:** It is a supervised learning used for both classification and regression. It works based on the principle that the data points tend to have similar outcomes. For example, if the nearest data point has value has true then the value become true.

**ADVANTAGES:**

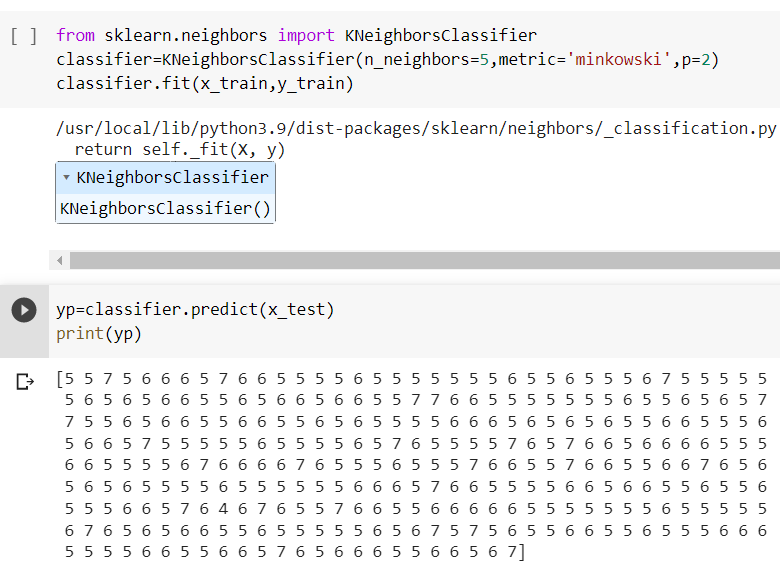
* Easy to understand and implement
* Doesn’t require training phase

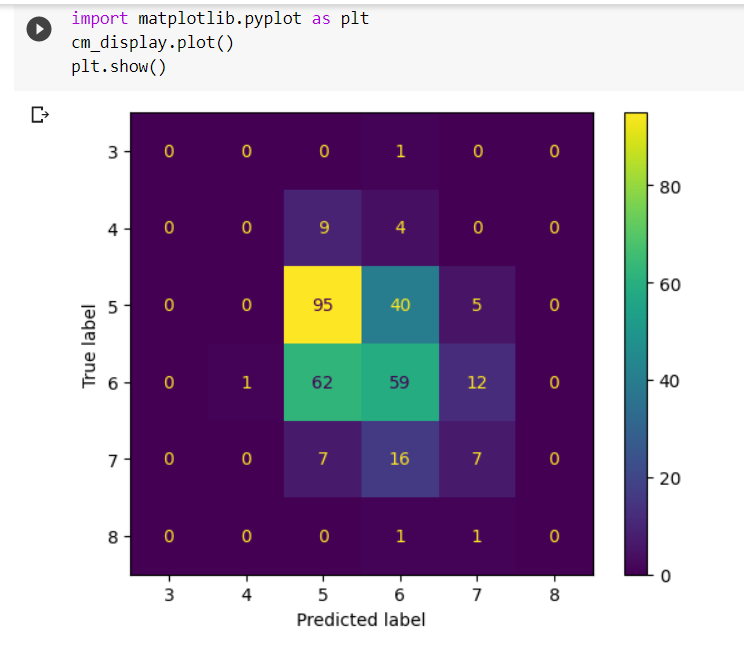
**DISADVANTAGES:**

* Expensive
* As it moves to the data which takes the nearest value it becomes biased data.



CODE.





**4.BAYES THEOREM:** Bayesian machine learning is a supervised learning which is used for both classification and regression that incorporates Bayesian probability principles to make predictions and infer model parameters. It is used to indentify conditional probabilities.

P(H/X) ={P(X/H) \*P(H)}/P(X)

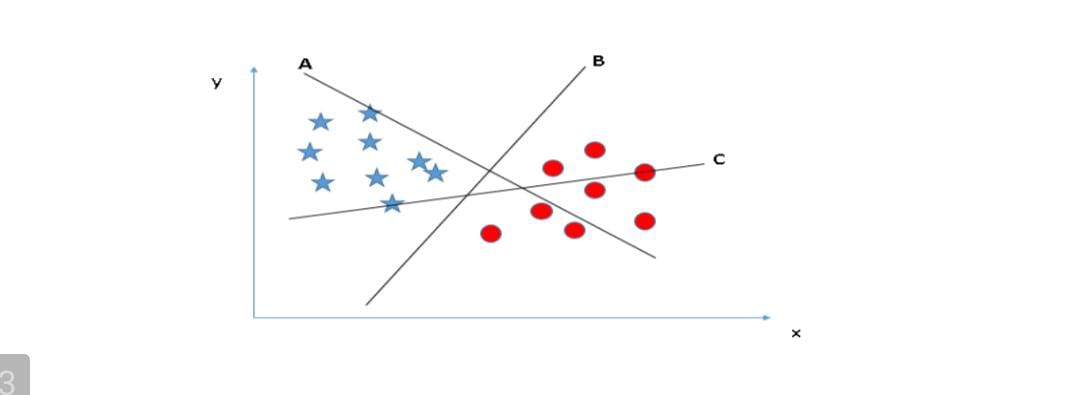
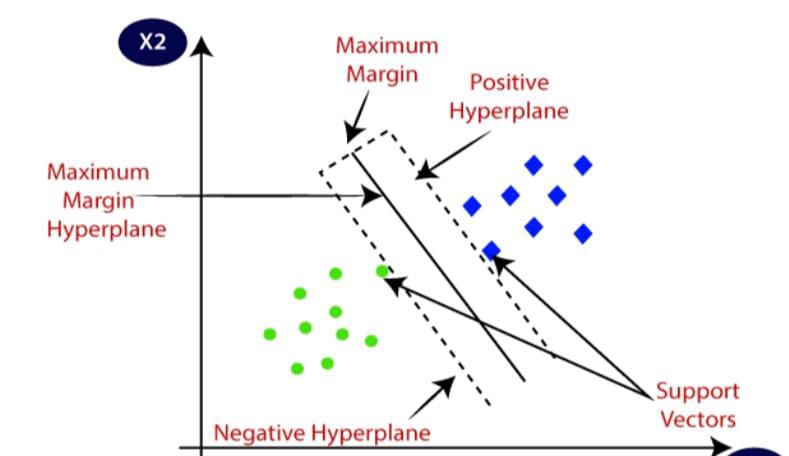
**ADVANTAGES:**

* Bayes theorem is more flexible
* Perform model averaging

**DISADVANTAGES**

* Expensive
* Faces problem in very large data sets

**5.SUPPORT VECTOR MACHINE:** SVM is a supervised machine learning which is used for both classification and regression. The main concept of SVM is to find a hyperplane that separates the datapoints into two classes. The distance between the hyperplane and the nearest data point from either set is known as MARGIN. The goal is to divide the hyperplane with the greatest possible margin. There will never be any data point inside the margin.

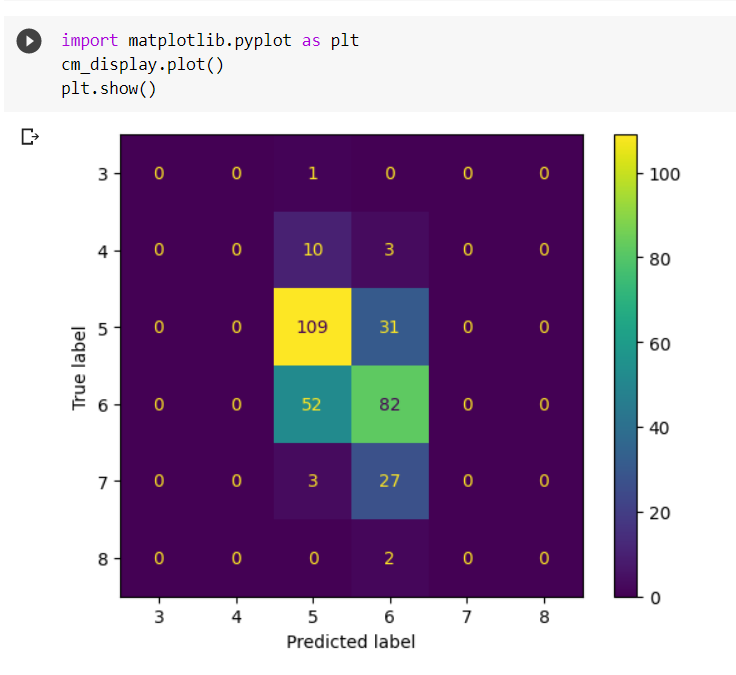
**ADVANTAGES:**

* Effective in high-dimensional spaces
* SVM aims to maximize the margin between classes, which can result in better generalization performance and improved model robustness.

**DISADVANTAGES:**

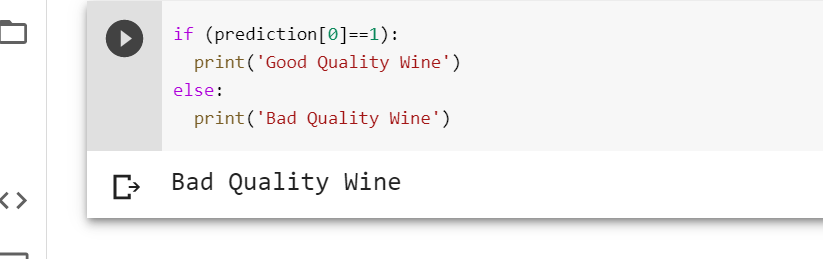
* Expensive for large datasets
* Requires sufficient memory to store SVMs



**ss**

# 7. RESULTS:

Here are the results we obtained from training our model i.e. training loss is around 0.05 and validation loss is 0.0574 whereas training accuracy is 97.1% and validation accuracy is 97.3%

****

# 8. CONCLUSION:

The method has become significantly more efficient and time-consuming since the invention of machine learning algorithms. To assess the quality of wine based on the variables that it depends on, a variety of machine learning approaches and models were applied. These models improve the efficiency of the manufacturing industry's quality assessment process, which has a huge impact on our economy.

**FUTURE SCOPE:**

## It is evident that the method or the data must be updated in the future to increase the classifier's performance. We advocate employing feature engineering, possible wine quality associations, or the boosting algorithm on the more accurate way. Furthermore, by using different performance measurement and machine learning methods for better outcomes comparison. This research will aid the manufacturing businesses in predicting the quality of various sorts of wines depending on certain characteristics, as well as assisting them in producing a decent product.

# 9. REFERENCES:

1. 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.
2. 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.
3. 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
4. Estabrooks, A., Japkowicz, N., 2001. A Mixture-of-Experts Framework for Learning from Imbalanced Data Sets, in: Hoffmann, F., Hand, D.J., Adams, N., Fisher, D., Guimaraes, G. (Eds.), Advances in Intelligent Data Analysis, Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 34–43