|  |
| --- |
| **Data Science Project Training Report**  **on**  **Wine Quality Prediction using Machine Learning Techniques**  **BACHELOR OF TECHNOLOGY**  **Session 2021-22 in**  **Computer Science**  **By -**   1. **Shivanshu Baranwal (2000320120162)** 2. **Rupesh Verma (200320120143)** 3. **Yash Kumar Sharma (2000320120196)**   **MS SAPNA JAIN**  **ASSISTANT PROFESSOR**  **DEPARTMENT OF COMPUTER SCIENCE**  **ABES ENGINEERING COLLEGE, GHAZIABAD**  **AFFILIATED TO**  **DR. A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY, U.P., LUCKNOW (Formerly UPTU)** |

**Student’s Declaration**

I / We hereby declare that the work being presented in this report entitled **“WINE QUALITY PREDICTION USING MACHINE LEARNING”** is an authentic record of my / our own work carried out under the supervision of Ms. **SAPNA JAIN, Assistant Professor, Computer**

|  |  |  |
| --- | --- | --- |
|  | **Science** Date: |  |
| **Signature of student student** | **Signature of student** | **Signature of** |
| **Shivanshu Baranwal** | **Rupesh Verma** | **Yash Kumar Sharma** |
| **(2000320120162 )** | **(2000320120143)** | **(2000320120196)** |
| **Department: CS** | **Department: CS** | **Department: CS** |

This is to certify that the above statement made by the candidate(s) is correct to the best of my knowledge.

|  |  |
| --- | --- |
| **Signature of HOD** | **Signature of Teacher** |
| **Prof.(Dr.) Pankaj Kumar Sharma** | **Sapna Jain** |
| **Computer Science** | **Assistant Professor**  **Computer Science** |

Date:..........................

# TABLE OF CONTENTS-

|  |  |  |
| --- | --- | --- |
| **S.NO** | **CONTENTS** | **PAGE NO.** |
|  | **Students declaration** |  |
| **1.** | **INTRODUCTION** |  |
| 1.1 | Problem statement |  |
| 1.2 | Objective |  |
| 1.3 | Methodology |  |
| 1.4 | Literature Survey |  |
| **2.** | **DATA EXPLORATION** |  |
| 2.1 | Dataset |  |
| 2.2 | Exploratory Data Analysis |  |
| **3.** | **PREPROCESSING** |  |
| 3.1 | Train test splitting of data |  |
| **4.** | **MODELING** |  |
| 4.1 | Using Standard Scaler |  |
| 4.2 | Using Logistic Regression |  |
| 4.3 | Using SVM |  |
| 4.4 | Using SVC |  |
| **5.** | **IMPLEMENTATION ( CODE WITH SCREENSHOTS)** |  |
| **6.** | **CONCLUSION** |  |
| **7.** | **REFERENCES** |  |

## 1. INTRODUCTION

The quality of the wine is a very important part for the consumers as well as the manufacturing industries. Industries are increasing their sales using product quality certification. Nowadays, all over the world wine is a regularly used beverage and the industries are using the certification of product quality to increases their value in the market. Previously, testing of product quality will be done at the end of the production, this is time taking process and it requires a lot of resources such as the need for various human experts for the assessment of product quality which makes this process very expensive. Every human has their own opinion about the test, so identifying the quality of the wine based on humans experts it is a challenging task.

There are several features to predict the wine quality but the entire features will not be relevant for better prediction.

The research aims to what wine features are important to get the promising result by implementing the machine learning classification algorithms such as Support Vector Machine (SVM), Naïve Bayes (NB), and Artificial Neural Network (ANN), using the wine quality dataset.

The wine quality dataset is publically available on the UCI machine learning repository (Cortez et al., 2009). The dataset has two files red wine and white wine variants of the Portuguese “Vinho Verde” wine. It contains a large collection of datasets that have been used for the machine learning community. The red wine dataset contains 1599 instances and the white wine dataset contains 4898 instances. Both files contain 11 input features and 1 output feature. Input features are based on the physicochemical tests and output variable based on sensory data is scaled in 11 quality classes from 0 to 10 (0-very bad to 10-very good).

### 1.1 PROBLEM STATEMENT

There are many moments in life that make our hearts sink. One of the worst is surely coming home [from a wine shop](https://learn.winecoolerdirect.com/wine-buying-tips/) with a disappointing bottle. One of the problems with wine is that it is sometimes inherently mysterious: until you’ve popped the cork, the contents of the bottle can be either thrillingly exciting, or painfully underwhelming. Often, you simply don’t know until you’ve poured out that first glass, and taken a tentative sip.

Choosing wine can be hard. It really doesn’t help that the world of wine has a tendency to be irritatingly pretentious at times, and bottle labels are more often than not [filled with elitist nonsense](https://learn.winecoolerdirect.com/minerality/), impenetrable bits of jargon, and archaic French adjectives that block our understanding and put us in a state of confusion when browsing. Making the most of your trip to the wine store takes time, practice, and you’re going to make mistakes at some point.

### 1.2 OBJECTIVE

To achieve our goal of developing machine learning model to predict the quality of wine, we need perform following tasks in the same order as stated.

-Data Collection and Analysis

-Pre-processing the data

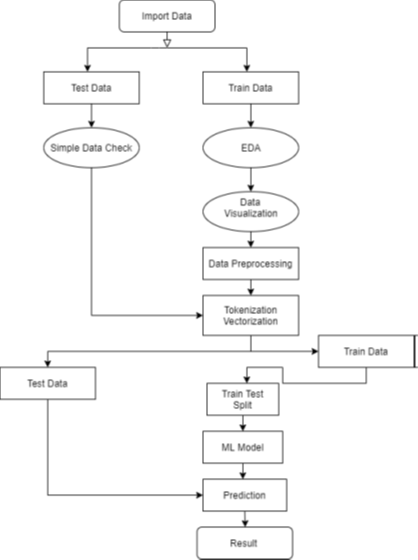
-Graphical Representation of data according to quantities in Dataset.

-Using different algorithms

-Taking the best algorithm and predicting method.

-Predicting the quality of wine.

### 1.3 METHODOLOGY



### 1.4 LITERATURE SURVEY

There are many moments in life that make our hearts sink. One of the worst is surely coming home from a wine shop with a disappointing bottle. Choosing wine can be hard. It really doesn’t help that the world of wine has a tendency to be irritatingly pretentious at times, and bottle labels are more often than not filled with elitist nonsense, impenetrable bits of jargon, and archaic French adjectives that block our understanding and put us in a state of confusion when browsing. Making the most of your trip to the wine store takes time, practice, and you’re going to make mistakes at some point.

## The good news is, understanding some basic information about wine can help you learn more about which wines are likely to be a good fit for you — and which ones you’ll probably want to avoid. By learning about your tastes and the general characteristics of wine, you can become a wine expert in no time and feel confident in selecting a delectable wine to pair with your meal!

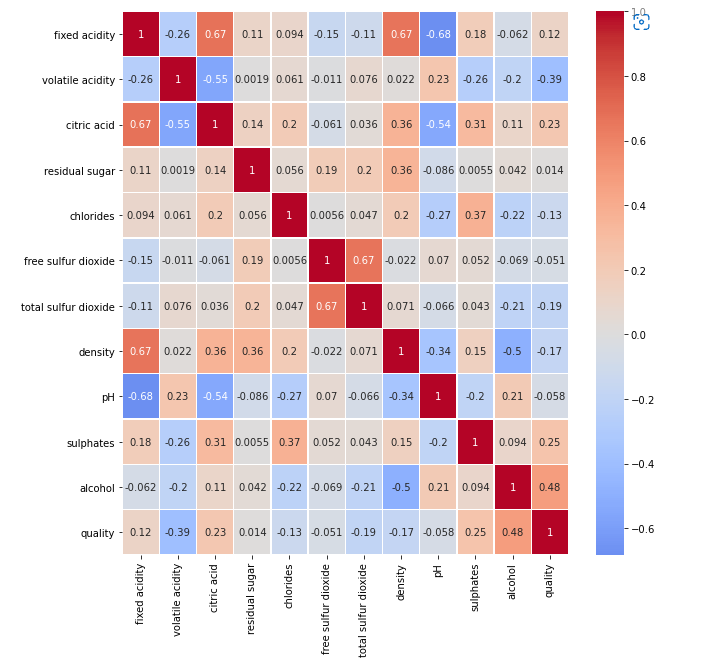
Choosing a good wine is completely subjective. How each person defines a good wine is unique to them and their taste buds. Whether you prefer delicate, bold, sweet, tart or even spicy flavors, it is possible to find a wine you adore. These essential characteristics that define each variety of wine can be helpful to keep in mind as you navigate picking a bottle.

* **Sweetness:** Wine labels often use the terms “sweet,” “semi-sweet” or “dry.” A dry wine will not be sweet at all.
* **Acidity:** Wines with high acidity will be more tart, whereas low-acidity wines will taste rounder or richer.
* **Tannin:** Tannins are phenolic compounds in the skins of grapes. When tannins are naturally present in the winemaking process or added through aging, the wine will have a more bitter taste. Because tannins also tend to dry out your mouth, people often confuse the tannin level with the “dryness” of a wine, which actually refers to how sweet or not sweet a wine is. The red winemaking process incorporates more tannins, giving some red wines a distinctively dry and bitter finish.
* **Body:** Wines get characterized as having a light body, full body or somewhere in between. The “body” of the wine refers to how heavy or light it feels in your mouth. Generally, red wines have a fuller body than whites, as do wines made from grapes grown in warmer regions, rather than cooler ones.
* **Alcohol:** The higher the percentage of alcohol in your glass of wine, the more it will warm your throat and the back of your mouth. Measured in percentage of alcohol by volume (ABV), most wines contain 11 to 13 percent alcohol, but can range from 5.5 percent all the way up to 20 percent.

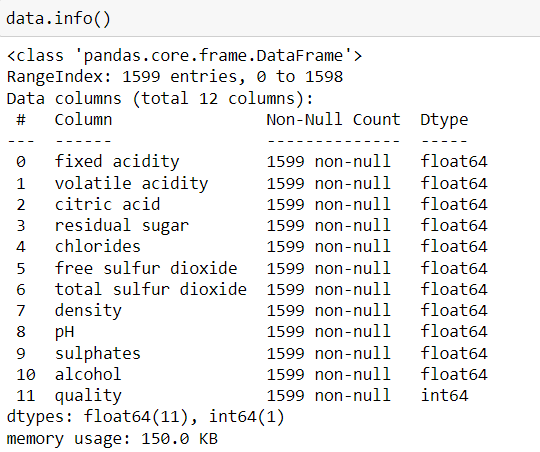
## 2. DATA EXPLORATION

My analysis will use Red Wine Quality Data Set, available on the UCI machine learning repository (<https://archive.ics.uci.edu/ml/datasets/wine+quality>). I obtained the red wine samples from the north of Portugal to model red wine quality based on physicochemical tests. The dataset contains a total of 12 variables, which were recorded for 1,599 observations. This data will allow us to create different regression models to determine how different independent variables help predict our dependent variable, quality. Knowing how each variable will impact the red wine quality will help producers, distributors, and businesses in the red wine industry better assess their production, distribution, and pricing strategy.

### 2.1 DATASET



### 2.2 EXPLORATORY DATA ANALYSIS



## 3. PREPROCESSING

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

Data preprocessing is done to make data ready for actual processing i.e. implementation of various machine learning models. Data preprocessing includes various steps such as dropping unnecessary columns or dealing with null values of categorical as well as numerical parameters or loading or exploring the data or analysis of different parameters of the data to find the correct parameters for the prediction.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

**3.1 TRAIN TEST SPLITTING OF DATA**

Splitting dataset into Training and Testing Datasets.

* + Training as 70%
  + Testing as 30%

## 4. MODELING

This section deals with training the classifier. Different classifiers were investigated to predict the class of the text. We explored specifically four different machine-learning algorithms –

1. Standard Scaler
2. Logistic Regression
3. Support Vector Machine (SVM)
4. Support Vector Classification (SVC)

**4.1 USING STADARD SCALER**

Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance). It standardizes features by subtracting the mean value from the feature and then dividing the result by feature standard deviation.The standard scaling is calculated as:

**z = (x - u) / s**

Where,

* z is scaled data.
* x is to be scaled data.
* u is the mean of the training samples
* s is the standard deviation of the training samples.

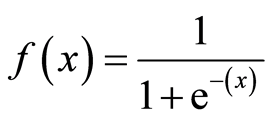
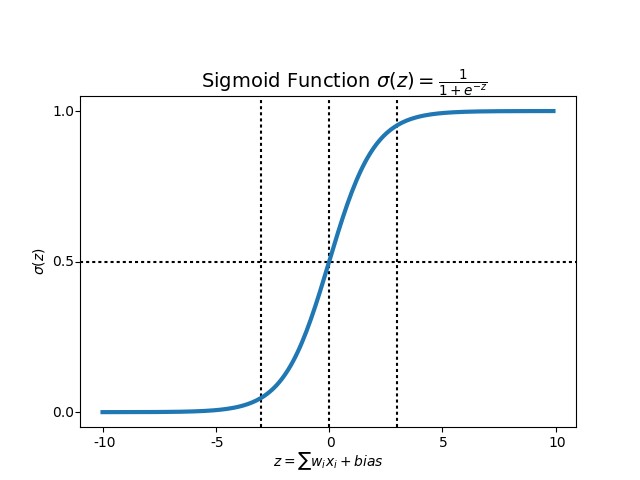
### 4.2 USING LOGISTIC REGRESSION

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.

We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the ‘**Sigmoid function**’ or also known as the ‘logistic function’ instead of a linear function. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.



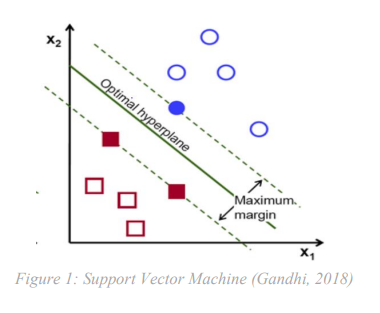
In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.



### 4.3 Using Support Vector Machine (SVM)

The support vector machine (SVM) is the most popular and most widely used machine learning algorithm. It is a supervised learning model that can perform classification and regression tasks. However, it is primarily used for classification problems in machine learning (Gandhi, 2018).

The SVM algorithm aims to create the best line or decision boundary that can separate n-dimensional space into classes. So we can put the new data points easily in the correct groups. This best decision boundary is called a hyperplane.

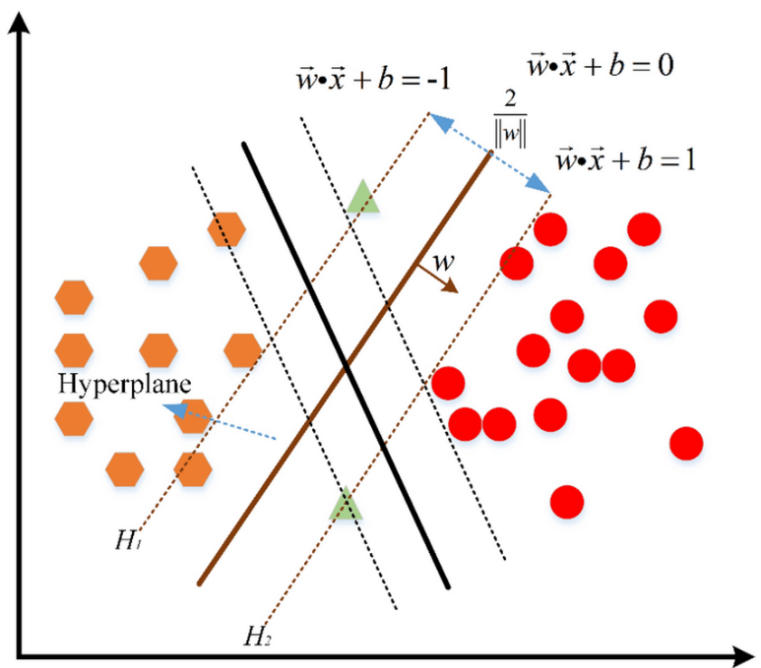


The SVM model is used for both non-linear and linear data. It uses a nonlinear mapping to convert the main preparing information into a higher measurement. The model searches for the linear optimum splitting hyperplane in this new measurement. A hyperplane can split the data into two classes with an appropriate nonlinear mapping to suitably high measurements and for the finding, this hyperplane SVM uses the support vectors and edges (J. Han et al., 2012). The SVM model is a representation of the models as a point in space, the different classes are isolated by the gap to mapped with the aim that instances are wide as would be careful. The model can perform out a nonlinear form of classification (Kumar et al., 2020).

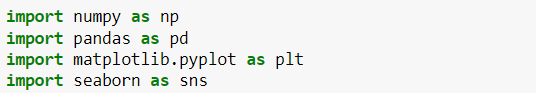
**4.4 Using Support Vector Classification** (**SVC)**

“Support Vector Classifier” (SVC) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

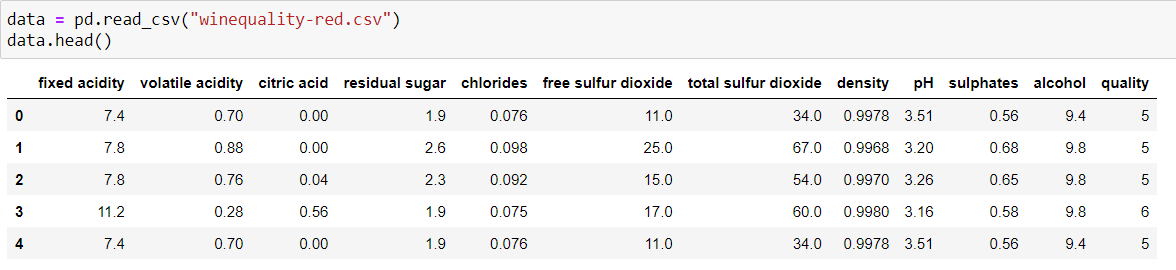
It is C-support vector classification whose implementation is based on libsvm. The module used by scikit-learn is sklearn.svm.SVC. This class handles the multiclass support according to one-vs-one scheme.

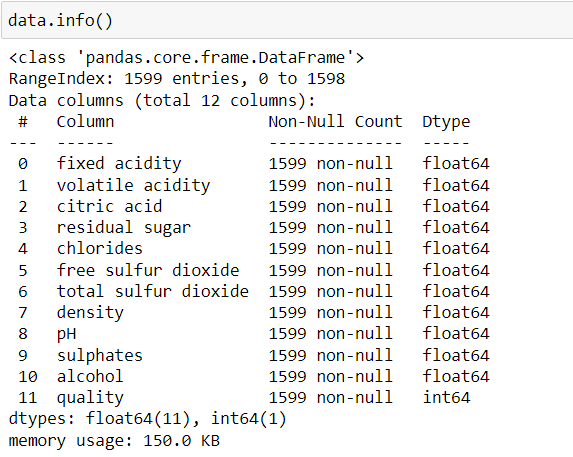


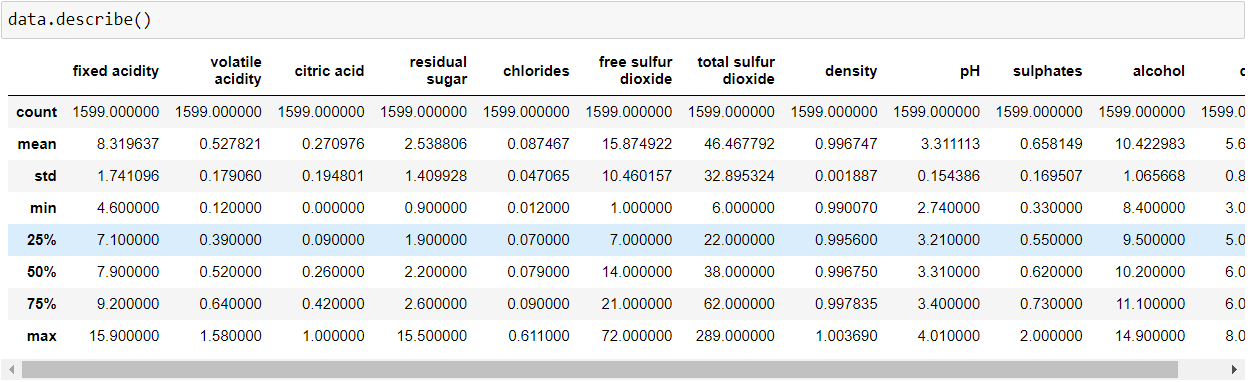
## 5. IMPLEMENTATION( CODE AND SCREENSHOTS)

**1.Loading Libraries and datasets** 

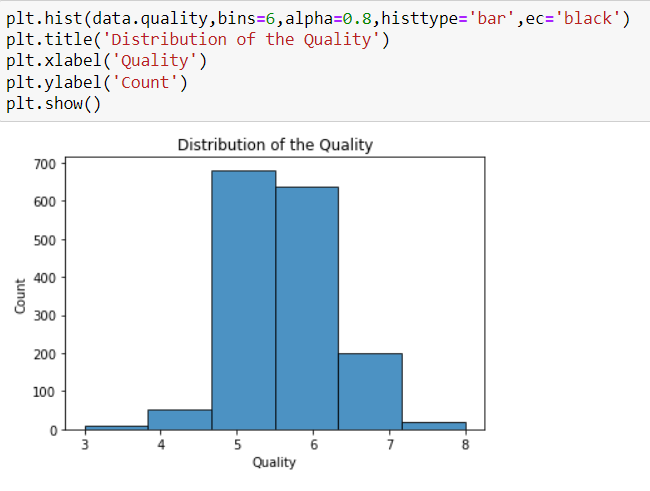
### 2. Exploratory Data Analysis



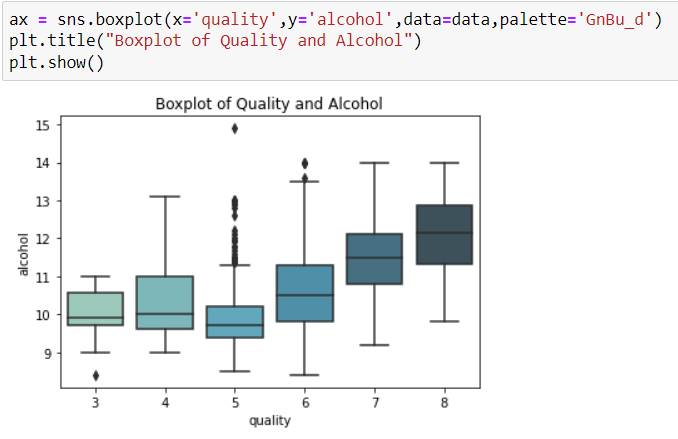


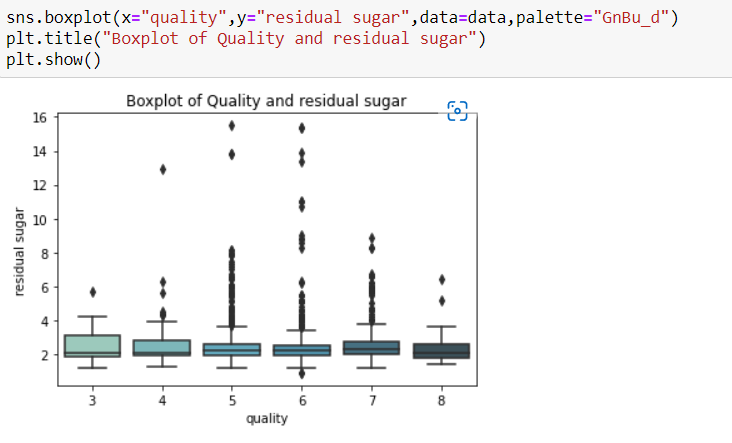


**3. Visualization**







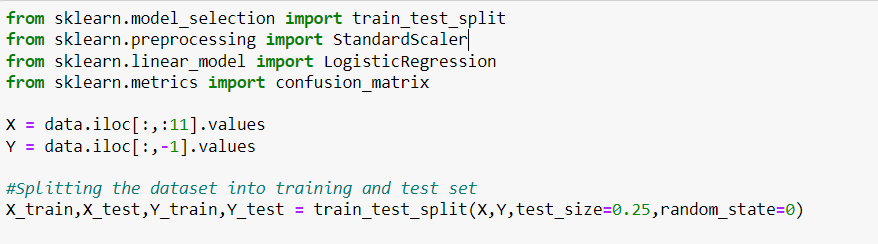


#### Train-Test split of data

● Splitting dataset into Training and Testing Datasets.

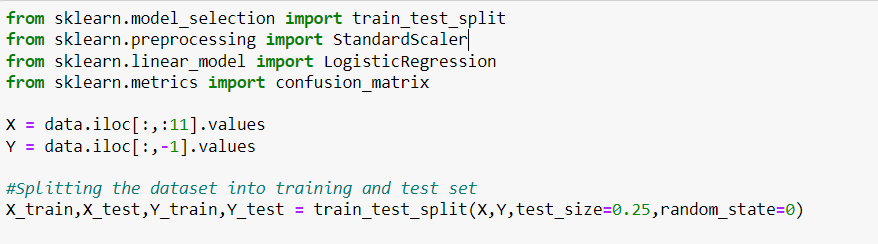
− Training as 70%

− Testing as 30%

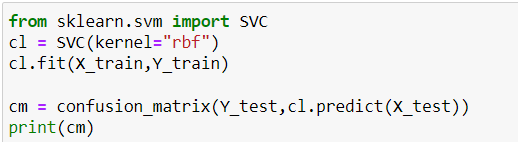


### 4.Modeling

#### a. Standard Scaler and Logistic Regression



#### b. SVC



## 6. CONCLUSION

This report uses the two types of wine dataset red and white, of Portuguese “Vinho Verde” wine to predict the quality of the wine based on the physicochemical properties.

First, 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 is balancing 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.

Therefore, in the classification algorithms by selecting the appropriate features and balancing the data can improve the performance of the model

## 7. 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.** [**https://doi.org/10.1007/0-387-25465-X\_40**](https://doi.org/10.1007/0-387-25465-X_40)
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.** [**https://doi.org/10.1016/j.dss.2009.05.016**](https://doi.org/10.1016/j.dss.2009.05.016)
3. **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**](https://doi.org/10.4236/ojs.2021.112015)
4. **Dastmard, B., 2013. A statistical analysis of the connection between test results and field claims for ECUs in vehicles.**
5. **Drummond, C., Holte, R.C., 2003. C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling. pp. 1–8.**
6. **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**](https://doi.org/10.18201/ijisae.265954)
7. **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**](https://doi.org/10.18201/ijisae.265954)
8. **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.** [**https://doi.org/10.1007/3-540-44816-0\_4**](https://doi.org/10.1007/3-540-44816-0_4)
9. **Fauzi, M., Arifin, A.Z., Gosaria, S., Prabowo, I.S., 2017. Indonesian News Classification Using Naïve Bayes and Two-Phase Feature Selection Model. Indones. J. Electr. Eng. Comput. Sci. 8, 610–615.** [**https://doi.org/10.11591/ijeecs.v8.i3.pp610-615**](https://doi.org/10.11591/ijeecs.v8.i3.pp610-615)
10. **Gandhi, R., 2018. Support Vector Machine — Introduction to Machine Learning Algorithms [WWW Document]. Medium. URL https://towardsdatascience.com/support-vector-machineintroduction-to-machine-learning-algorithms-934a444fca47 (accessed 6.6.21).**
11. **Guo, X., Yin, Y., Dong, C., Yang, G., Zhou, G., 2008. On the Class Imbalance Problem, in: 2008 Fourth International Conference on Natural Computation. Presented at the 2008 Fourth International Conference on Natural Computation, pp. 192–201. https://doi.org/10.1109/ICNC.2008.871**