## Artificial Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

Wine Quality Analysis

A red and white sign

Description automatically generated with low confidence

**Supervised By: Submitted By:**

Faculty Name: Bhavya Sood, 2210991443(G-18)

Mr. Animesh Ojha Bhuvesh Mittal, 2210991450(G-18)

Daanushi Sharma, 2210991469(G-18)

**Department of Computer Science and Engineering**

## Chitkara University Institute of Engineering & Technology,

## Chitkara University, Punjab

**ABSTRACT**

With the greater development of technology and automation human history is predominantly updated. The technology movement shifted from large mainframes to PCs to cloud when computing the available data for a larger period. This has happened only due to the advent of many tools and practices that elevated the next generation in computing. A large number of techniques has been developed so far to automate such computing. Research dragged towards training the computers to behave similar to human intelligence. Here the diversity of machine learning came into play for knowledge discovery. Machine Learning (ML) is applied in many areas such as medical, marketing, telecommunications, stock, health care and so on. This paper presents reviews about machine learning algorithm foundations, its types and flavors together with R code and Python scripts possibly for each machine learning technique.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Contents** | **Page No.** |
|  | Introduction | 6 |
|  | Research Methodology | 8 |
|  | Dataset | 13 |
|  | Machine Learning Project | 16 |
|  | Findings and Discussion | 40 |
|  | Implementation | 41 |
|  | Conclusion | 42 |

**TABLE OF CONTENTS**

1. **Introduction** 
   1. Background of the Study
   2. Need and Significance of the Study
   3. Statement of the Problem
   4. Objective of the Study
   5. Scope of the Study
   6. Limitations of the Study
2. **Research Methodology**
   * 1. Objective
     2. Hypothesis
   1. Software Requirements
   2. Tools Description
3. **Dataset**

3.1 Source of Data

3.2 Origin

3.3 Usage

3.4 Number of Cases

3.5 Order

3.6 Variables

3.7 Sample Design

3.8 Data Analysis

3.8.1 Data Analysis Techniques

**4. Machine Learning Project**

4.1 Wine Quality Analysis with Classification

4.2 Visualization and Analysis

4.3 Using Algorithms

4.4 Evaluation Metrics

**5. Findings and Discussion**

**6. Implementation**

**7. Conclusion**

**Chapter 1**

**INTRODUCTION**

**1.1 BACKGROUND OF THE STUDY**

This dataset is related to red variants of the Portuguese "Vinho Verde" wine. This dataset describes the amount of various chemicals present in wine and their effect on it's quality. The study aims to provide insights for winemakers to enhance production processes and wine quality through data-driven approaches.

**1.2 NEED AND SIGNIFICANCE OF THE STUDY**

Predicting wine quality using machine learning algorithms is essential for the wine industry to meet consumer expectations and maintain competitiveness in the market. Traditional methods of quality assessment are subjective, time-consuming, and costly. By leveraging data analytics, winemakers can streamline the quality evaluation process and optimize production practices. This study addresses the need for accurate and efficient wine quality prediction models by utilizing a dataset containing physicochemical properties and sensory quality scores of red "Vinho Verde" wine. The significance lies in its potential to provide actionable insights for winemakers, enabling them to make informed decisions regarding production techniques, ingredient selection, and quality control measures.

**1.3 STATEMENT OF PROBLEM**

Wine quality assessment is crucial for both producers and consumers, yet accurately predicting wine quality remains a challenge due to the complex interplay of physicochemical properties. In the context of the Portuguese "Vinho Verde" red wine, this project aims to develop predictive models using machine learning techniques. By analyzing the dataset's physicochemical attributes and sensory quality scores, the objective is to create models that can reliably classify wine quality. The ultimate goal is to provide stakeholders in the wine industry with actionable insights to enhance production processes and ensure consumer satisfaction.

**1.4 OBJECTIVE OF THE STUDY**

➢The goal of this project is to create a regression model that is able to accurately predict the quality of a wine based on the given features.

➢ To compare and contrast multiple logistic regression algorithms and their training set/test set performance on the given dataset.

➢ To correctly build the classification models to predict the wine quality.

**1.5 SCOPE OF THE STUDY**

This study aims to evaluate the efficacy of various machine learning algorithms, including Data Mining, Visualization, and classification models, in predicting wine quality. The primary objective is to deepen understanding and proficiency in regression and classification methods within machine learning, specifically applied to wine quality assessment. The dataset will be processed to enhance model performance, including feature selection to identify relevant attributes for predicting wine quality accurately. Given that each wine possesses unique characteristics contributing to its quality, feature selection methods will help eliminate irrelevant variables and improve model interpretability.

**1.6 LIMITATIONS OF THE STUDY**

There is no guarantee that the data will be available in time nor contains the exact requested list of features. Thus, there might be a risk that the access will be denied or delayed. If so, the study will be accomplished based only on the public dataset. Moreover, this study will not cover all regression algorithms; instead, it is focused on the chosen algorithm, starting from the basic regression techniques to the advanced ones. Likewise, the artificial neural network that has many techniques and a wide area and several training methods that do not fit in this study.

**Chapter 2**

**2.1 RESEARCH METHODOLOGY**

**2.1.1 OBJECTIVES**

Research is a careful and detailed study of a particular problem or concern, using scientific methods. An in-depth analysis of information creates space for generating new questions, concepts and understandings. The main objective of research is to explore the unknown and unlock new possibilities. It’s an essential component of success. The objectives of research should be closely related to the problem statement, giving way to specific and achievable goals. Here are the two types of research objectives for you to explore:

❖ GENERAL OBJECTIVE

Also known as secondary objectives, general objectives provide a detailed view of the aim of a study. In other words, you get a general overview of what you want to achieve by the end of your study. For example, if you want to study an organization’s contribution to environmental sustainability, your general objective could be: a study of sustainable practices and the use of renewable energy by the organization.

❖ SPECIFIC OBJECTIVES

Specific objectives define the primary aim of the study. Typically, general objectives provide the foundation for identifying specific objectives. In other words, when general objectives are broken down into smaller and logically connected objectives, they’re known as specific objectives. They help define the who, what, why, when and how aspects of your project. Once you identify the main objective of research, it’s easier to develop and pursue a plan of action.

Let’s take the example of ‘a study of an organization’s contribution to environmental sustainability’ again. The specific objectives will look like this:

• To determine through history how the organization has changed its practices and adopted new solutions

• To assess how the new practices, technology and strategies will contribute to the overall effectiveness

**HYPOTHESIS**

With respect to the Wine Quality dataset study, we have 5 different studies of research hypotheses. These hypotheses serve as starting points for exploring the relationships between different wine attributes and quality ratings.

1) Hypothesis 1: Relationship between acidity and quality:

* Null Hypothesis (H0): There is no significant relationship between acidity levels (fixed acidity, volatile acidity, citric acid) and wine quality.
* Alternative Hypothesis (H1): Acidity levels have a significant impact on wine quality, with certain acidity levels associated with higher or lower quality ratings.

2) Hypothesis 2: Influence of alcohol content on quality:

* Null Hypothesis (H0): Alcohol content does not affect wine quality significantly.
* Alternative Hypothesis (H1): Wines with higher alcohol content are associated with higher quality ratings due to enhanced flavor complexity and perceived quality.

3) Hypothesis 3: Role of sulfur dioxide in preserving quality:

* Null Hypothesis (H0): The amount of sulfur dioxide (free and total) does not affect wine quality.
* Alternative Hypothesis (H1): Sufficient levels of sulfur dioxide contribute to preserving wine quality by preventing oxidation and microbial spoilage, leading to higher quality ratings.

4) Hypothesis 4: Influence of residual sugar on perceived sweetness:

* Null Hypothesis (H0): Residual sugar content does not influence perceived sweetness and overall wine quality.
* Alternative Hypothesis (H1): Wines with higher residual sugar levels are perceived as sweeter and may receive higher quality ratings, especially in certain wine styles such as dessert wines.

5) Hypothesis 5: Importance of pH in flavor balance:

* Null Hypothesis (H0): pH levels do not impact the balance of flavors and overall sensory perception of wine.
* Alternative Hypothesis (H1): Optimal pH levels contribute to flavor balance, acidity perception, and overall harmony in wine, leading to higher quality ratings.

**2.2 SOFTWARE REQUIREMENTS**

This project uses the following tools for its implementation:

1) Python: the programming language used to implement this project

2) Jupyter notebook: used to provide an interactive environment for python, for the implementation of this project

3) Scikit-Learn: used to implement the several machine learning models used in this project to view their accuracies.

4) Matplotlib: used to plot graphs for us to understand the trend in accuracy/performance of the various machine learning algorithms implemented.

5) Seaborn: use for visualization of the statistical behavior of the data set.

6) Pandas: - used to analyze and clean our dataset,

7) Numpy: for mathematical Calculations.

**2.3 TOOLS DESCRIPTION**

The various tools used in this project are described below.

1. Python

Python is an interpreted, high level programming language created by Guido ban Rossum in 1991 it emphasizes on code readability by using whitespace to terminate statements and blocks. It is easy to use and its syntax makes it a language of preference for data analysis.

1. Jupyter Notebook

It is a web application that provides an integrated development environment for python using this one can share documents that contain equations, visualizations such as graphs, test as a live for this reason it is a highly used tool for the purposes of data analysis.

1. Scikit-Learn

It is a python package used for data modeling. It provides a number of supervised and unsupervised machine learning models. Scikit-learn makes it extremely simple to train models with simple function calls on the input data being all that is needed.

1. Matplotlib

Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension.

1. Pandas

Pandas is an open-source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named NumPy, which provides support for multi-dimensional arrays.

1. Numpy

NumPy can be used to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.

**Chapter 3**

**DATASET (WINE QUALITY ANALYSIS)**

**3.1 SOURCE OF DATA**

Scikit-learn comes with a few small standard datasets. Wine quality dataset is retrieved from the sklearn-dataset load\_wine library.

**3.2 ORIGIN**

The origin of the Wine Quality dataset is Natural.

**3.3 USAGE**

This dataset may be used for Assessment.

**3.4 NUMBER OF CASES**

The dataset contains a total of 4898 cases.

**3.5 ORDER**

The order of the cases is mysterious.

**3.6 VARIABLES**

There are 12 attributes in each case of the dataset. They are:

1. Fixed Acidity: The total concentration of acid present in the wine.

2. Volatile Acidity: The level of volatile acids, which can contribute to unpleasant flavors if present in high amounts.

3. Citric Acid: The presence of citric acid, which can contribute to freshness and fruitiness.

4. Residual Sugar: The amount of sugar remaining in the wine after fermentation, affecting sweetness.

5. Chlorides: The concentration of chloride ions, which can influence taste perception.

6. Free Sulfur Dioxide: The amount of sulfur dioxide present in a free form, which acts as an antioxidant and antimicrobial agent.

7. Total Sulfur Dioxide: The total amount of sulfur dioxide, including both free and bound forms, crucial for wine preservation.

8. Density: The mass per unit volume of the wine, influenced by sugar and alcohol content.

9. pH: A measure of acidity or alkalinity, affecting the wine's stability and flavor.

10. Sulphates: The presence of sulphates, which can act as antioxidants and antimicrobial agents.

11. Alcohol: The percentage of alcohol by volume, influencing the wine's body and perceived sweetness.

12. Quality: Overall assessment of the wine's quality, often based on sensory evaluation and chemical analysis.

**3.7 SAMPLE DESIGN**



**3.8 DATA ANALYSIS**

**3.8.1 DATA ANALYSIS TECHNIQUES**

I. Importing all the Required modules / libraries.

II. Data Loading

III. Data Cleaning

IV. Data Mining

V. Data Visualization

VI. Determining correlation Matrix using Heat Map.

VII. Splitting the Data

VIII. Data Model Fit

IX. Classification Methods.

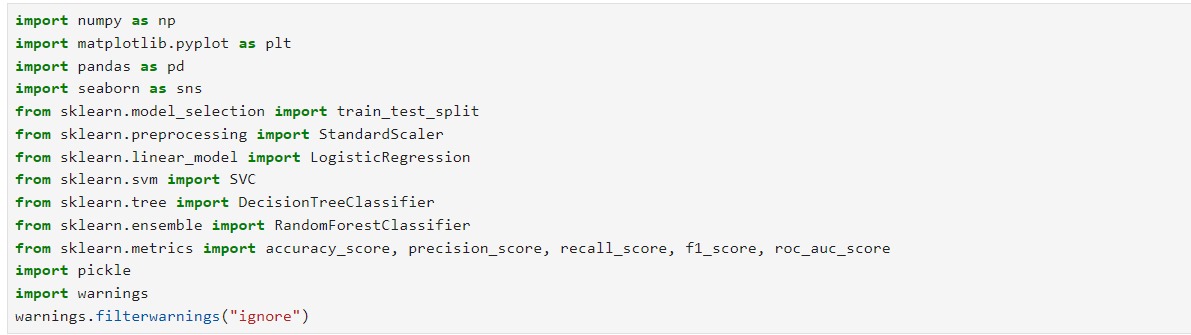
**Chapter 4**

**MACHINE LEARNING PROJECT:**

**4.1 WINE QUALITY ANALYSIS WITH CLASSIFICATION**

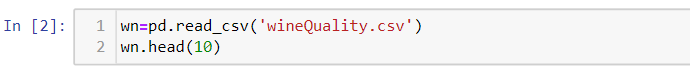
In this project, our aim is to construct and assess the efficacy of a model trained and validated on data sourced from various wines. Upon achieving a satisfactory model, we intend to employ it to forecast the perceived quality of wines based on their chemical attributes. Such a model holds immense value for wine sellers, sommeliers, and enthusiasts, empowering them with actionable insights for daily decision-making and enhancing their understanding of wine quality.

**IMPORT LIBRARIES NECESSARY FOR THIS PROJECT:**

****

**Fig 4.1.1**

**IMPORTING THE WINE QUALITY ANALYSIS DATASET:**

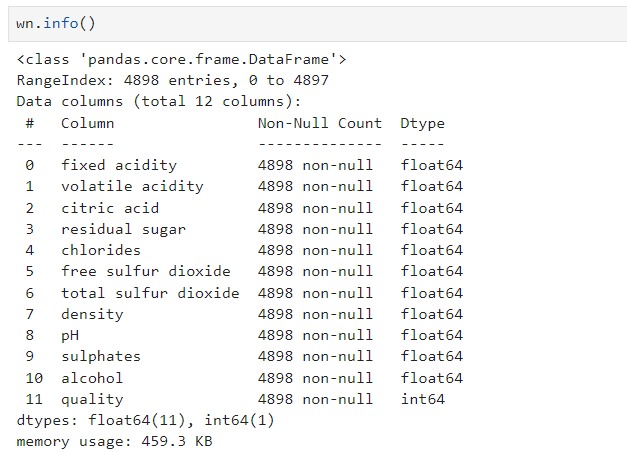
****

**Fig 4.1.2**

**DESCRIPTION OF WINE ANALYSIS DATASET: **

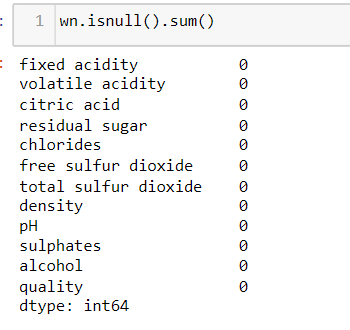
**Fig 4.1.3**

**INFORMATION OF DATASET:**

****

**Fig 4.1.4**

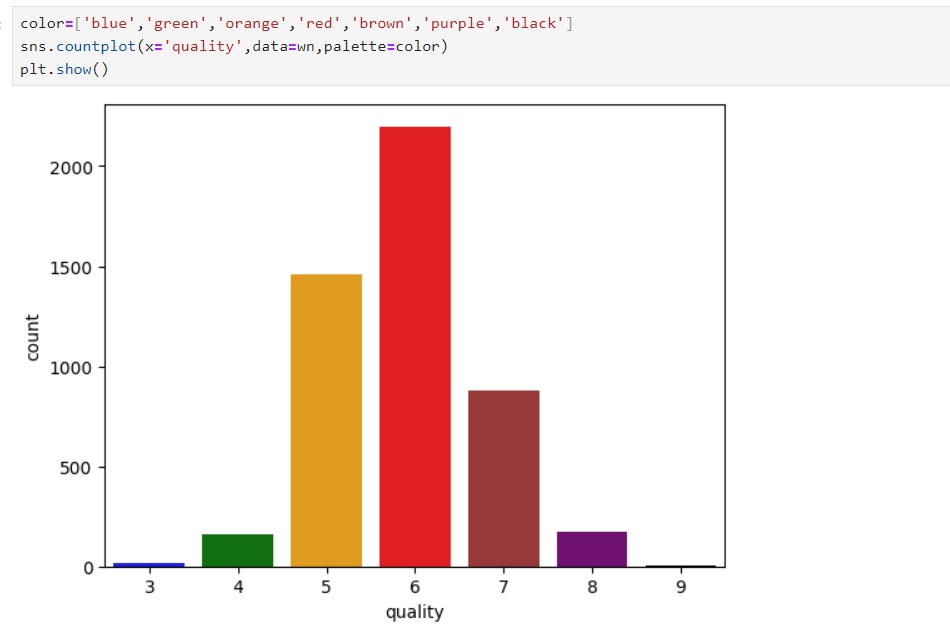
**CHECK FOR MISSING VALUES:**

****

**Fig 4.1.5**

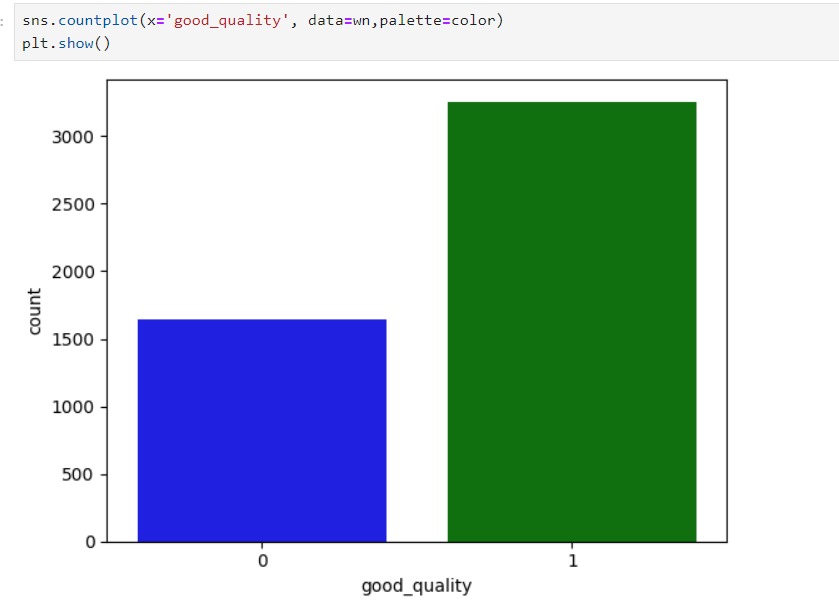
**4.2 VISUALIZATION & ANALYSIS**

**QUALITY COUNT**

****

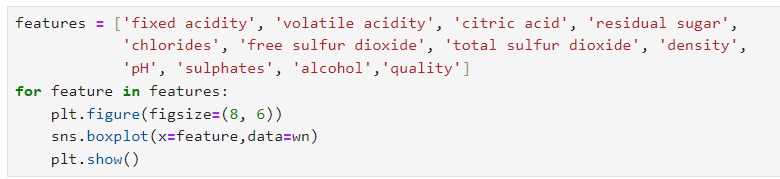
**Fig 4.2.1**

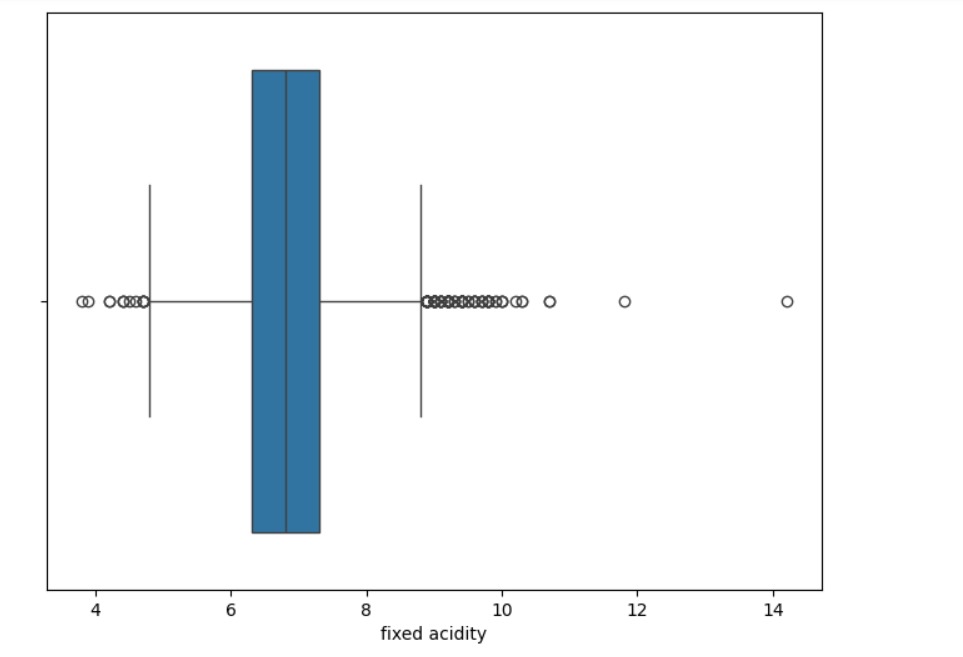
**GOOD\_QUALITY CREATED FROM QUALITY AS 0 AND 1.**

****

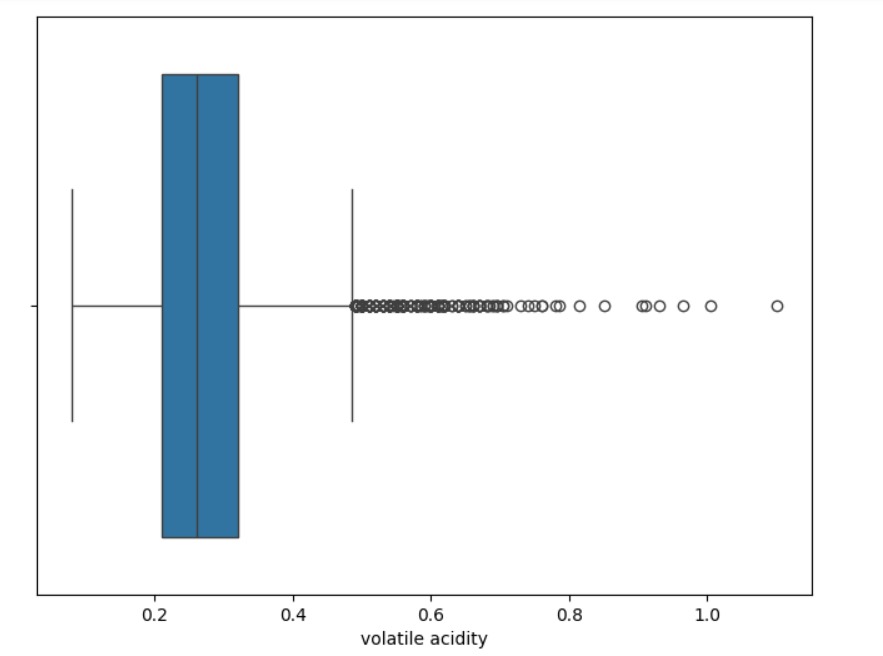
**Fig 4.2.2**

**OUTLIERS DETECTION**

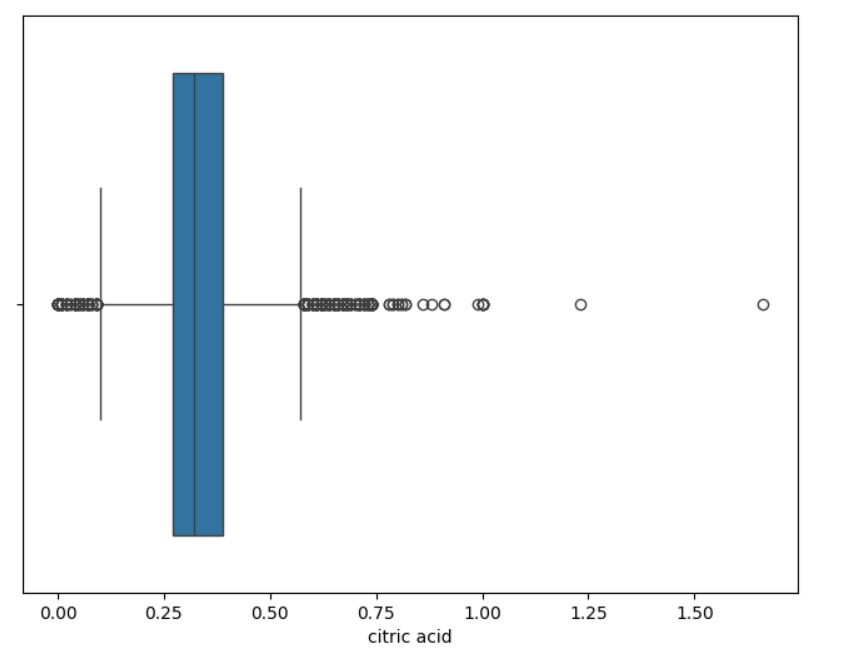
****

****

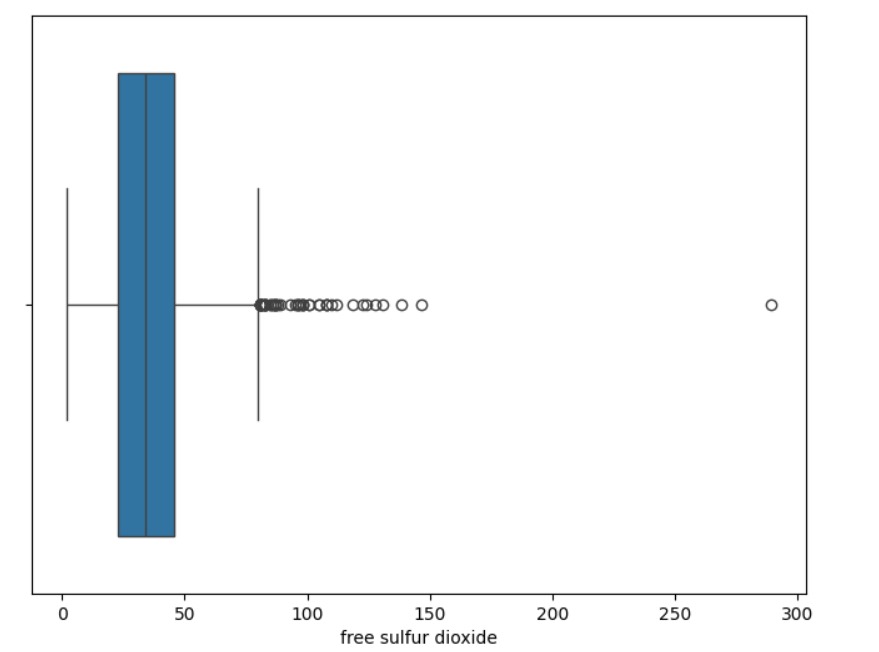
**Fig 4.2.3**

****

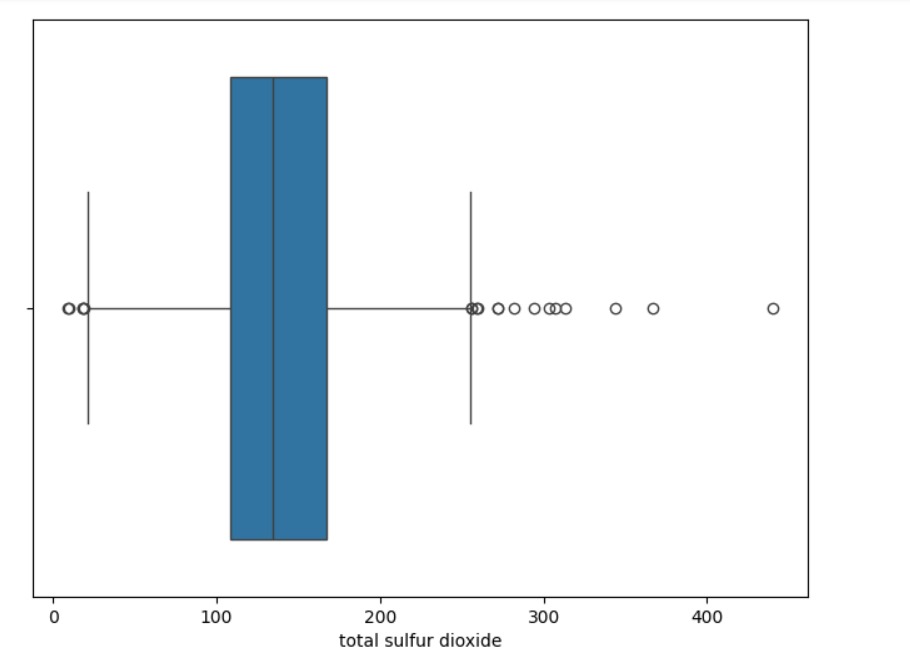
**Fig 4.2.4**

****

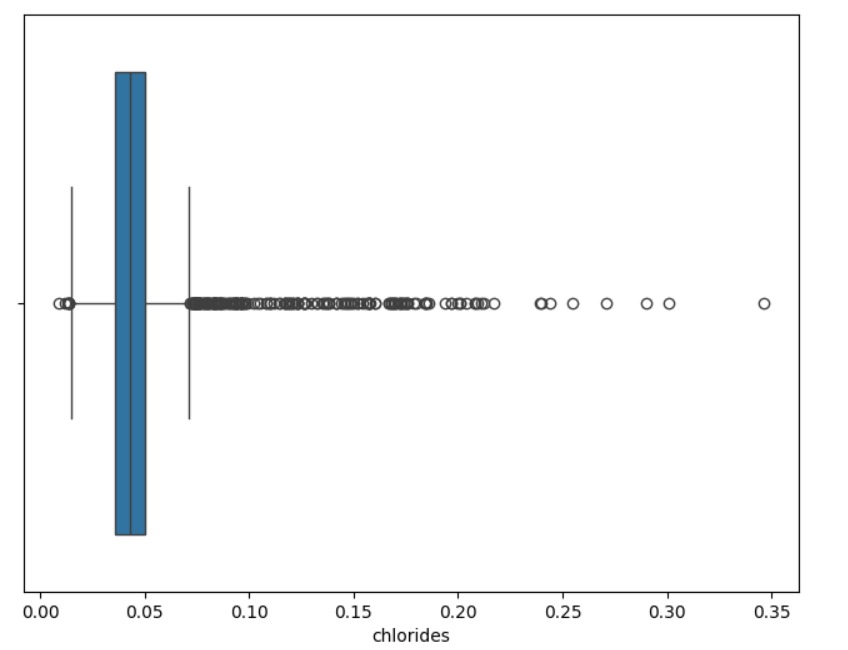
**Fig 4.2.5**

****

**Fig 4.2.6**

****

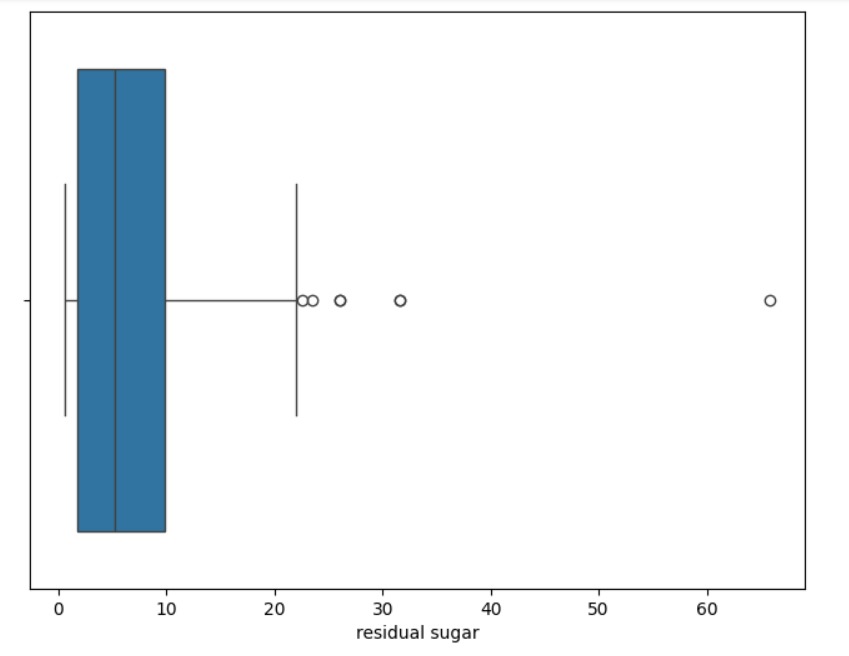
**Fig 4.2.7**

****

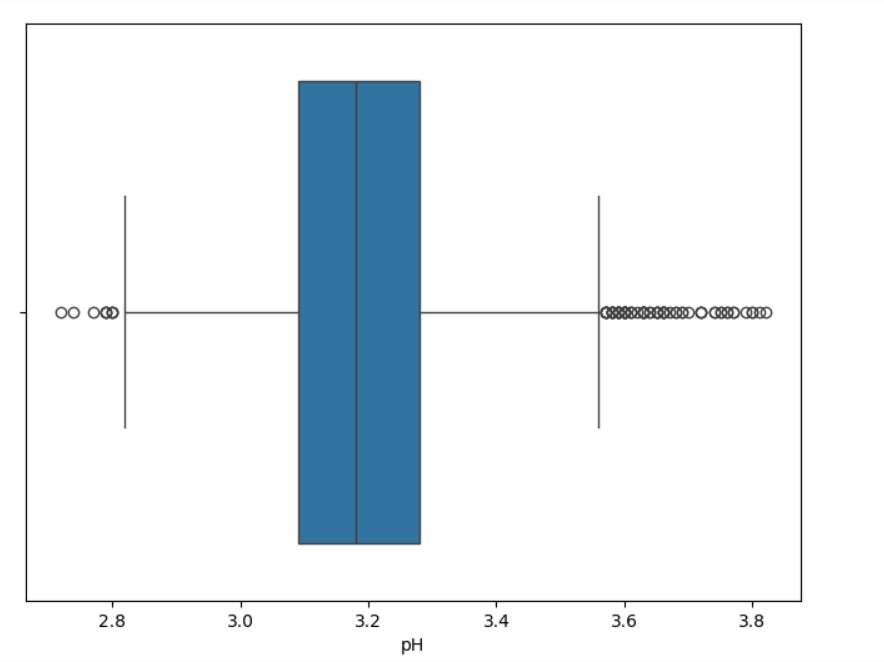
**Fig 4.2.8**

****

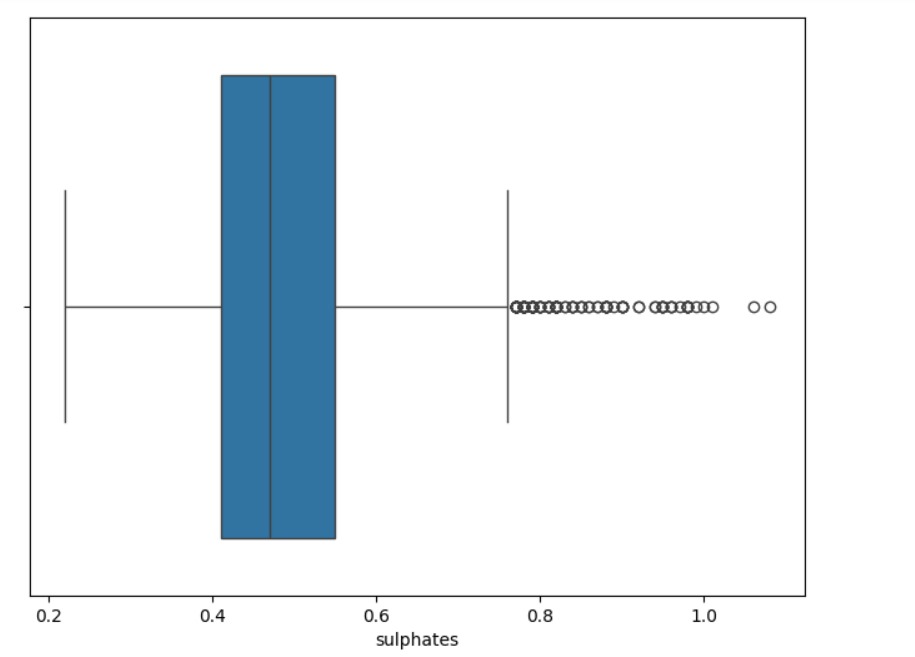
**Fig 4.2.9**

****

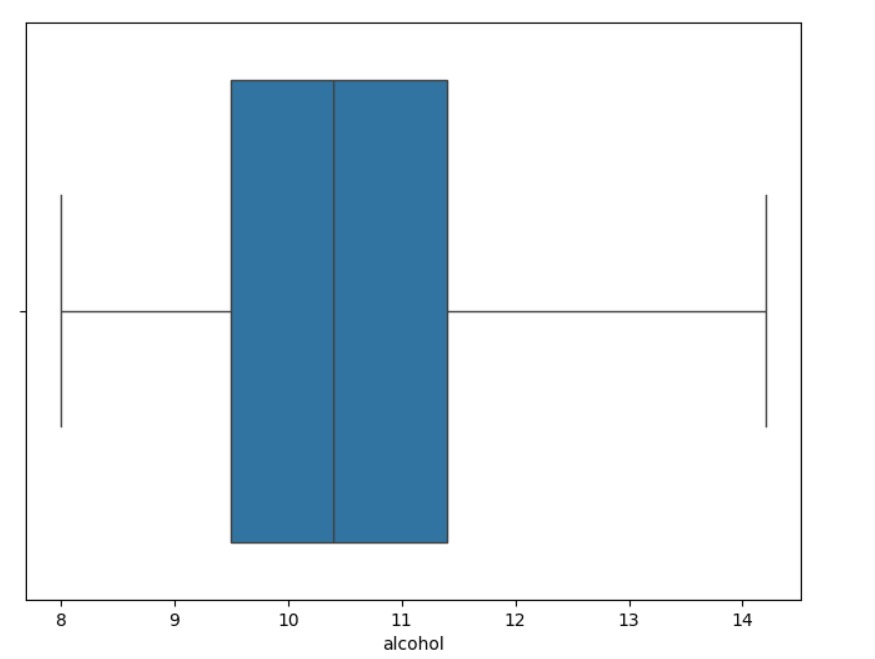
**Fig 4.2.10**

****

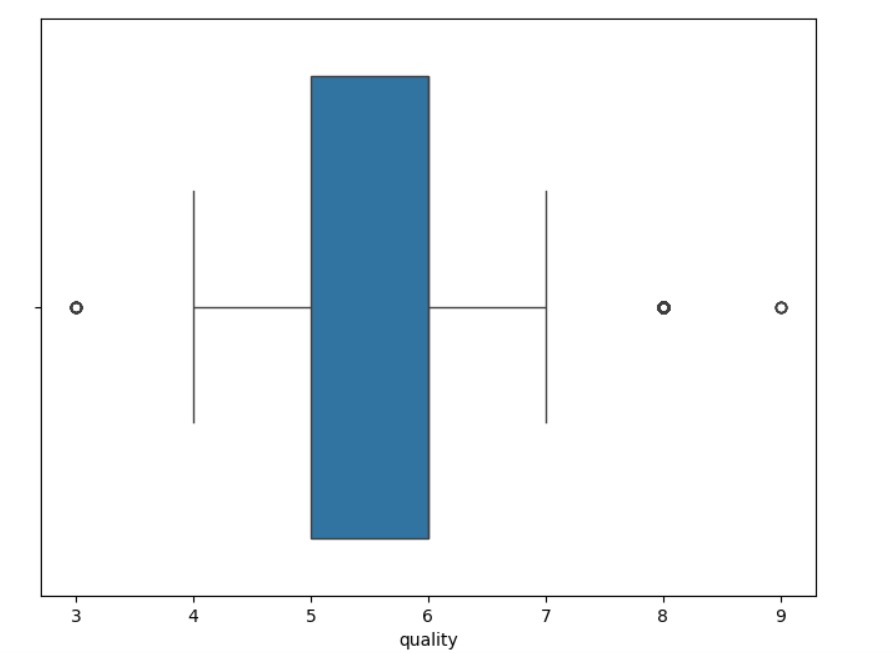
**Fig 4.2.11**

****

**Fig 4.2.12**

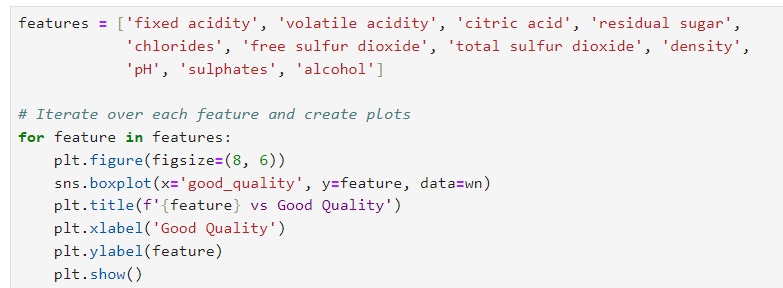
****

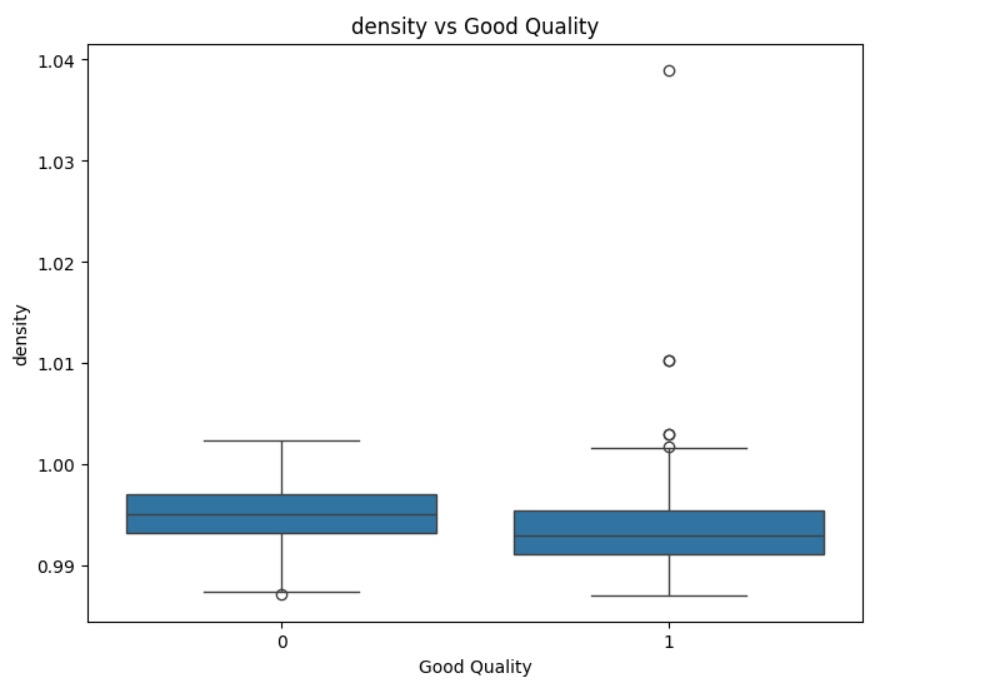
**Fig 4.2.13**

****

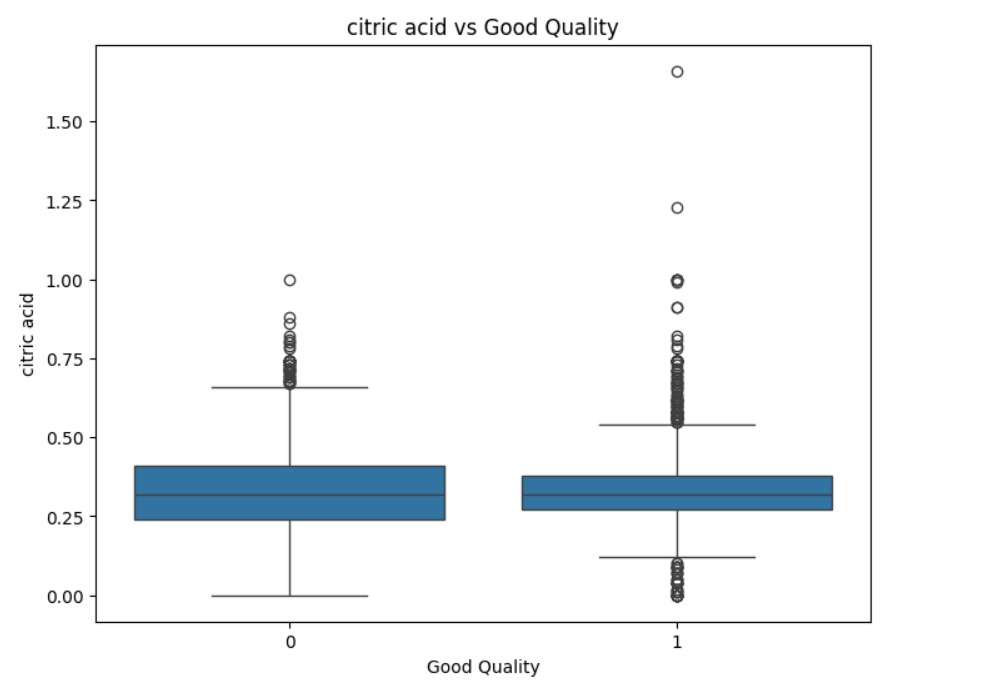
**Fig 4.2.14**

**GOOD\_QUALITY COMPARED WITH EVERY FEATURE IN BOXPLOT:**

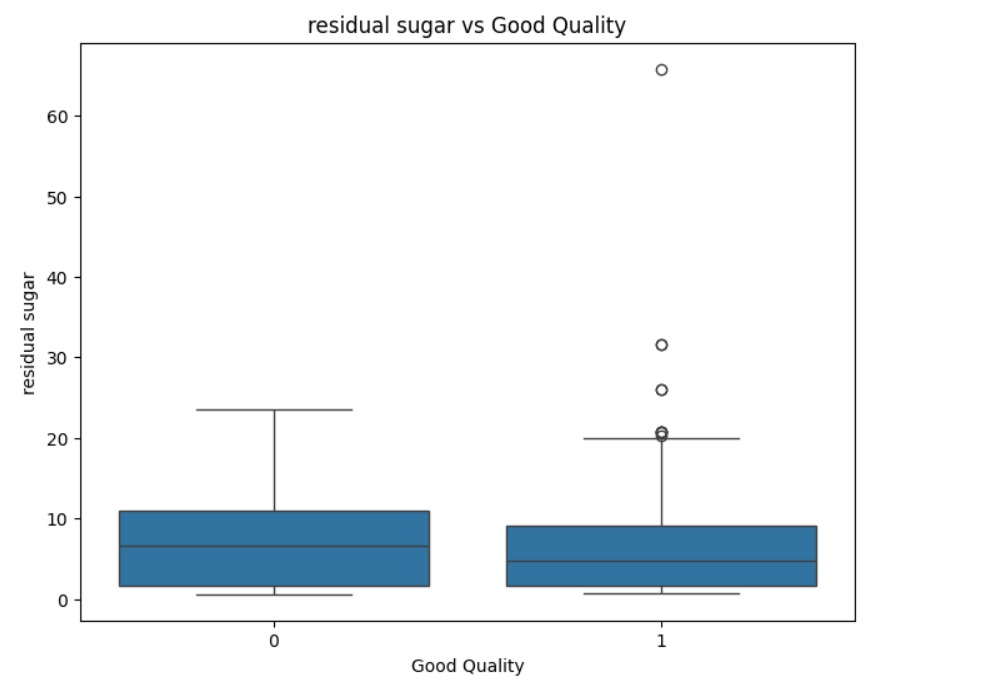
****

****

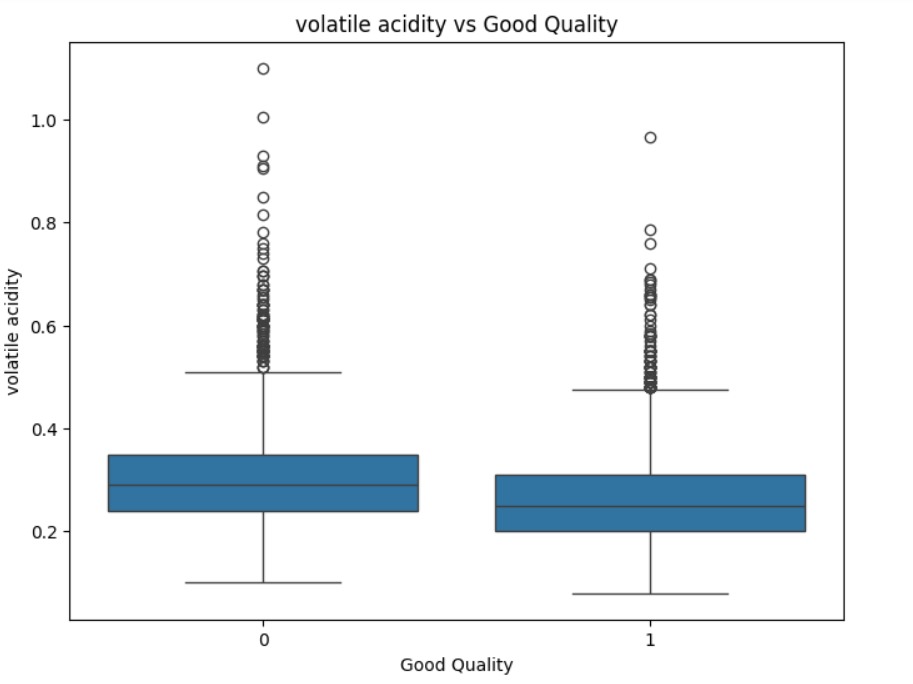
**Fig 4.2.15**

****

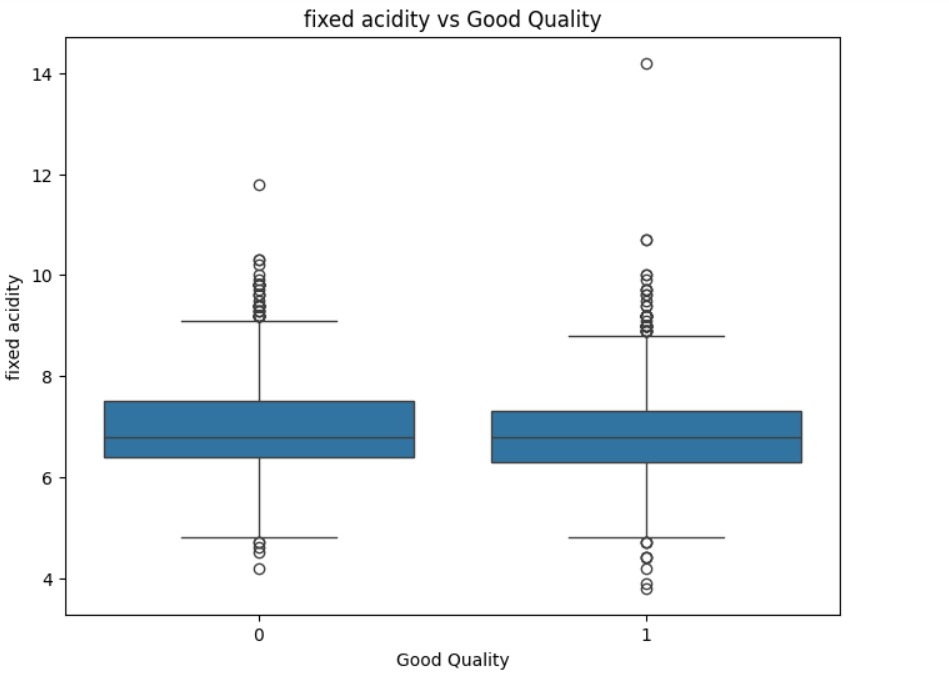
**Fig 4.2.16**

****

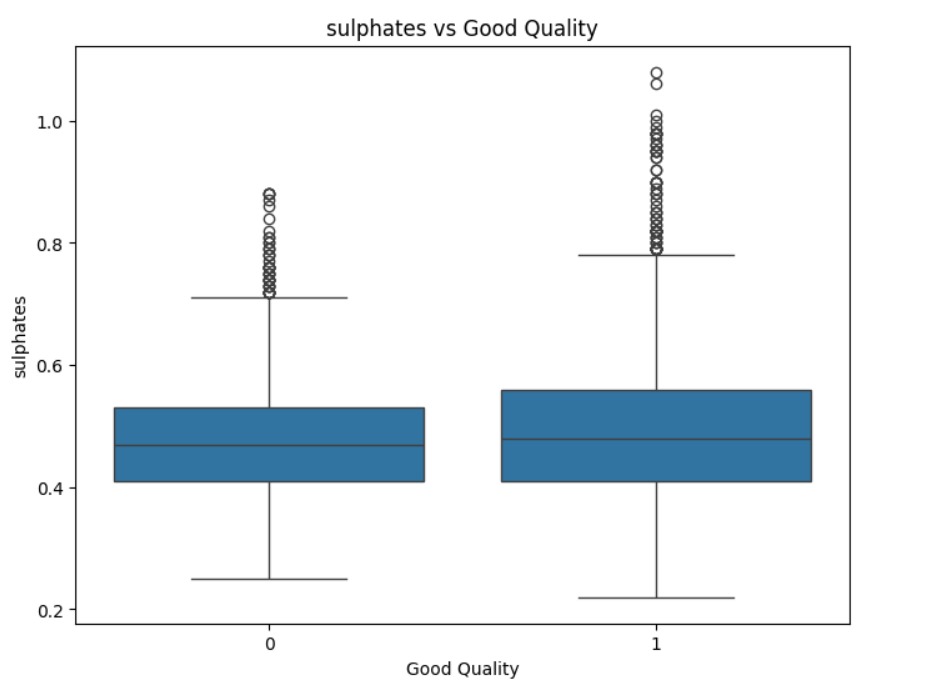
**Fig 4.2.17**

****

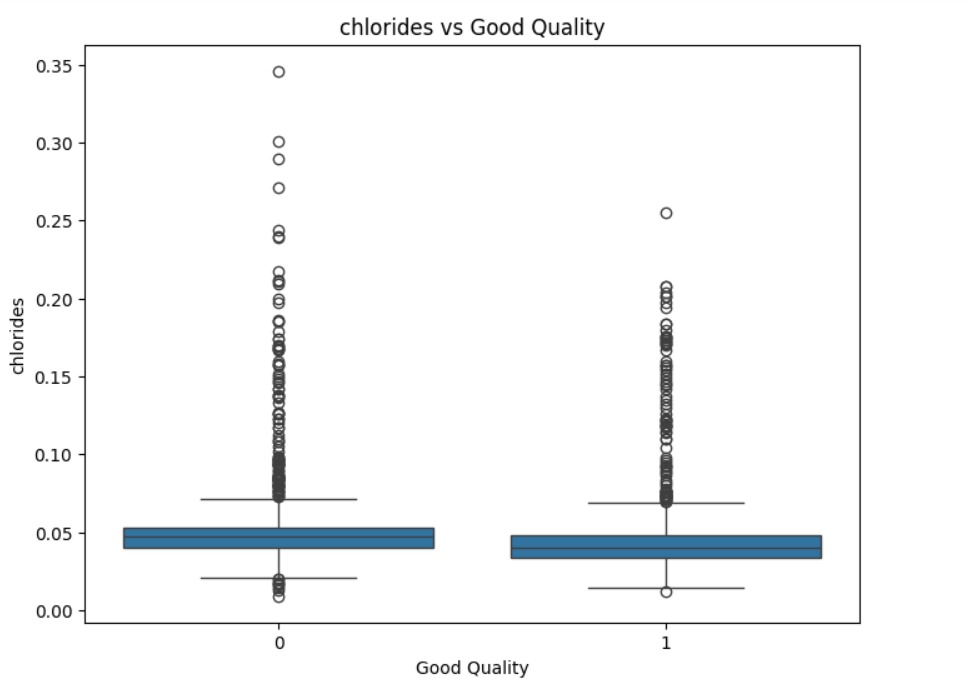
**Fig 4.2.18**

****

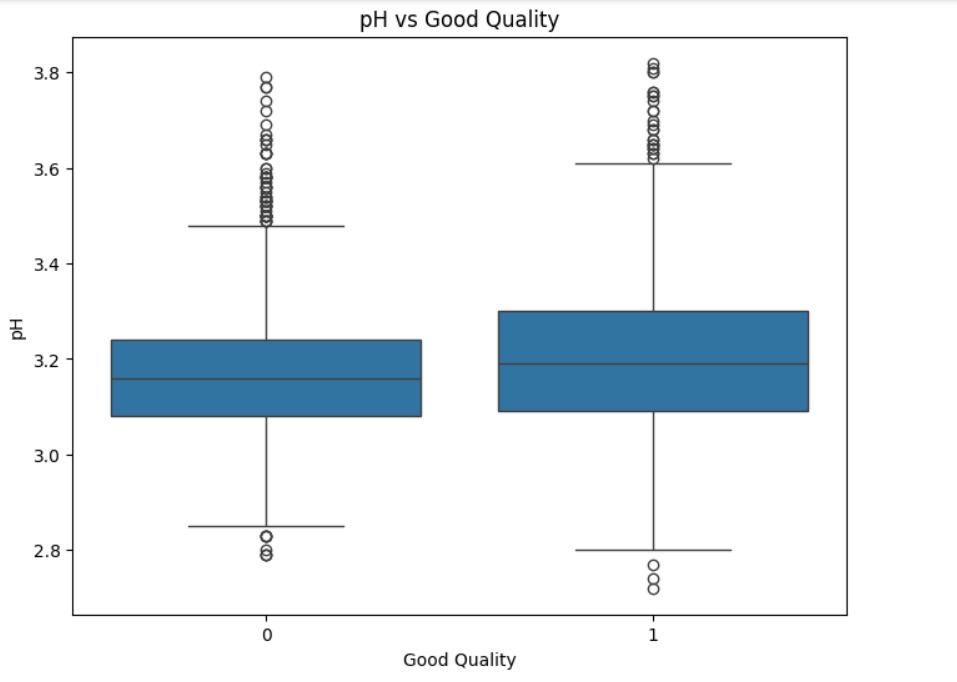
**Fig 4.2.19**

****

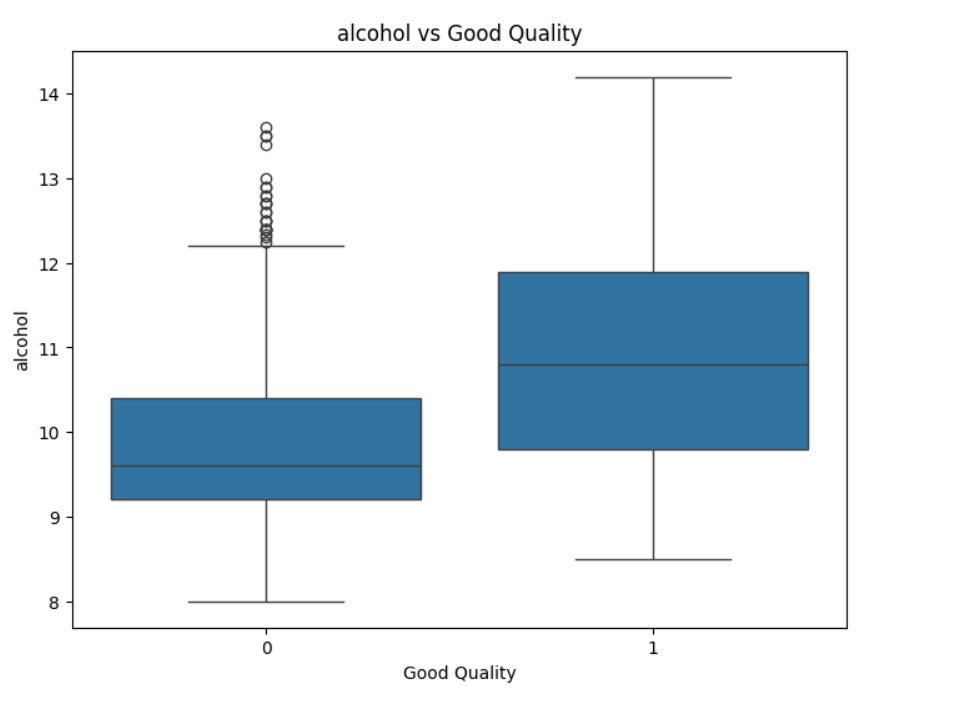
**Fig 4.2.20**

****

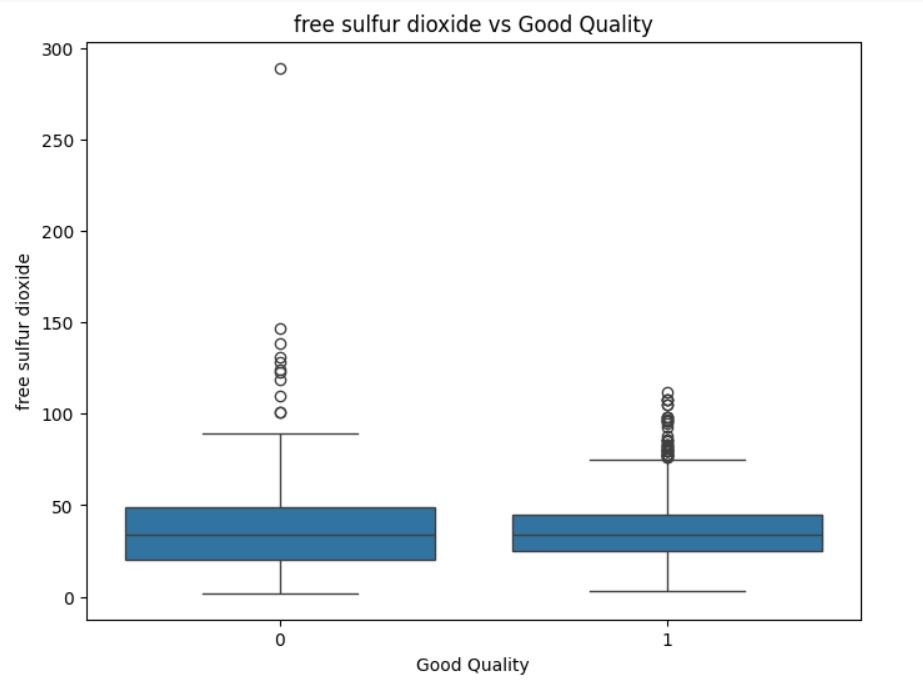
**Fig 4.2.21**

****

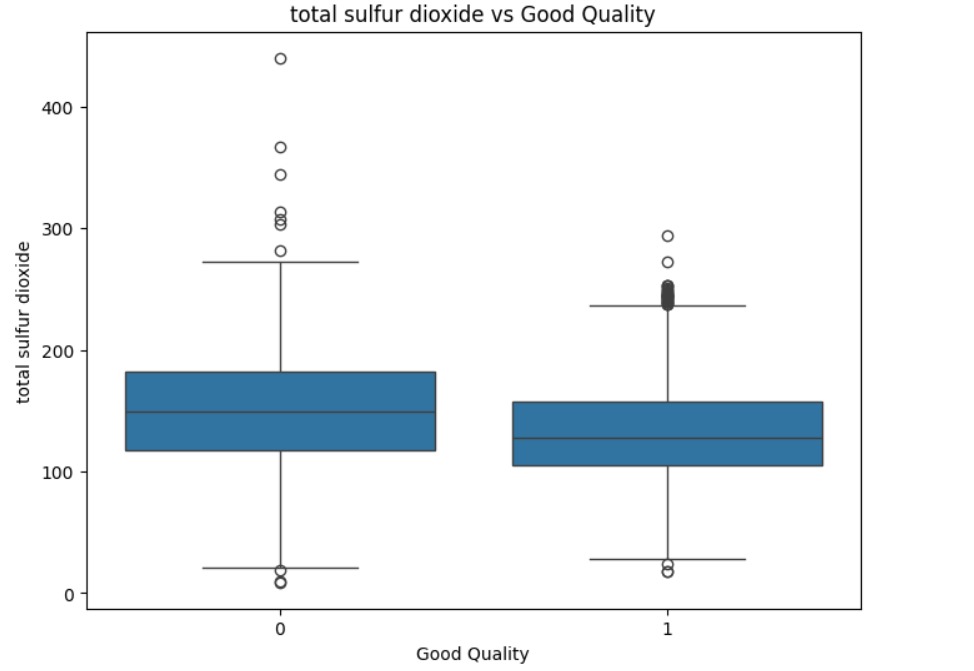
**Fig 4.2.22**

****

**Fig 4.2.23**

****

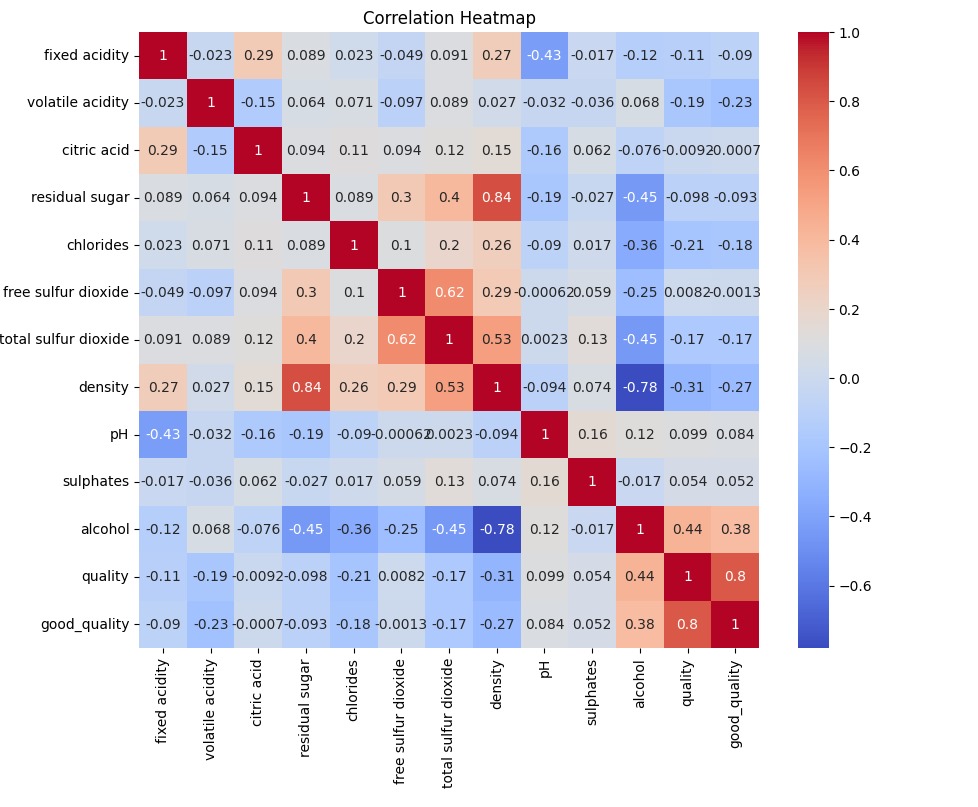
**Fig 4.2.24**

****

**Fig 4.2.23**

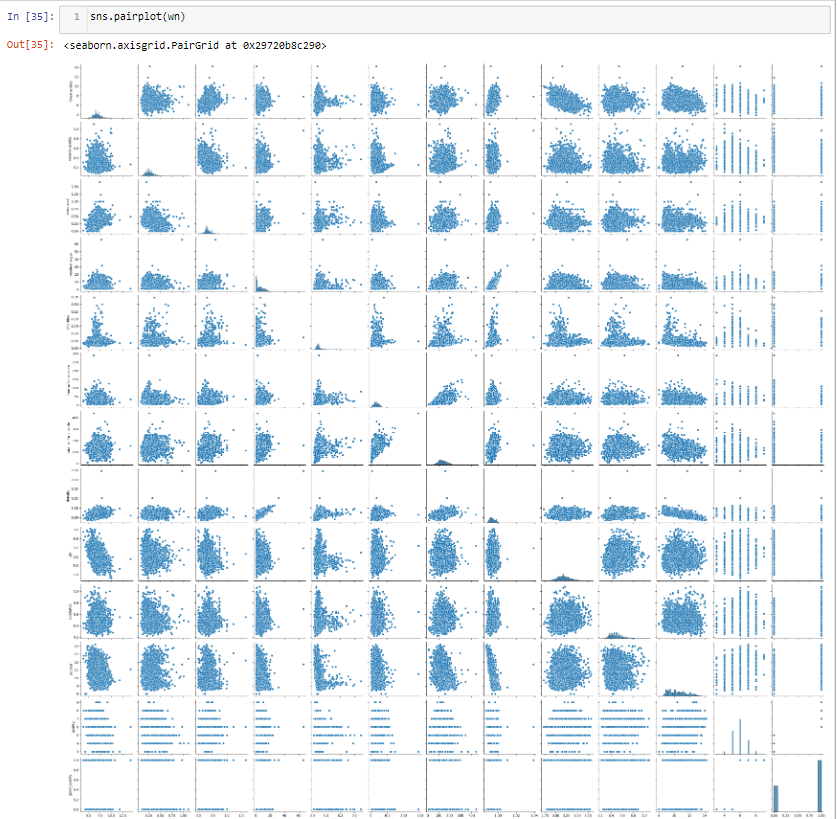
**CORRELATION MATRIX:**

****

****

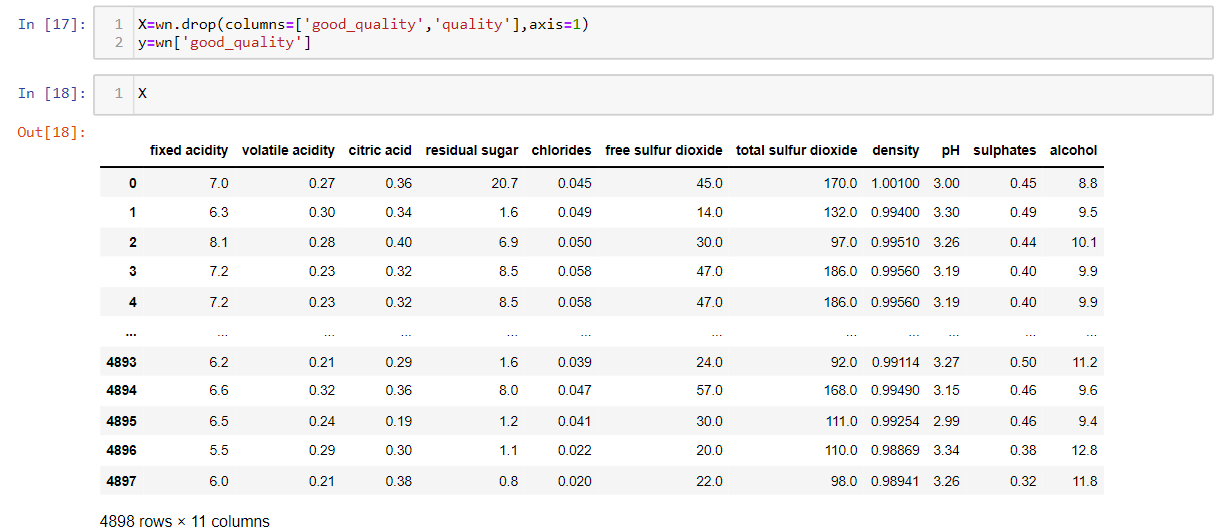
**Fig 4.2.24**

**PAIRPLOT**

****

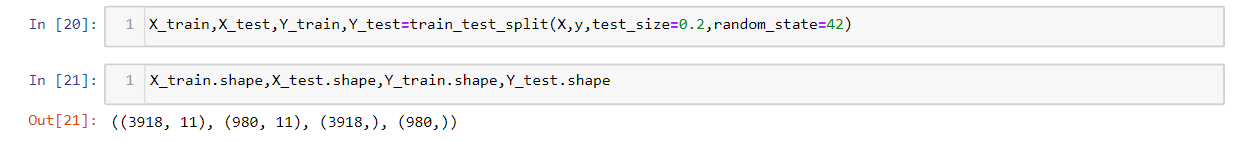
**Fig 4.2.25**

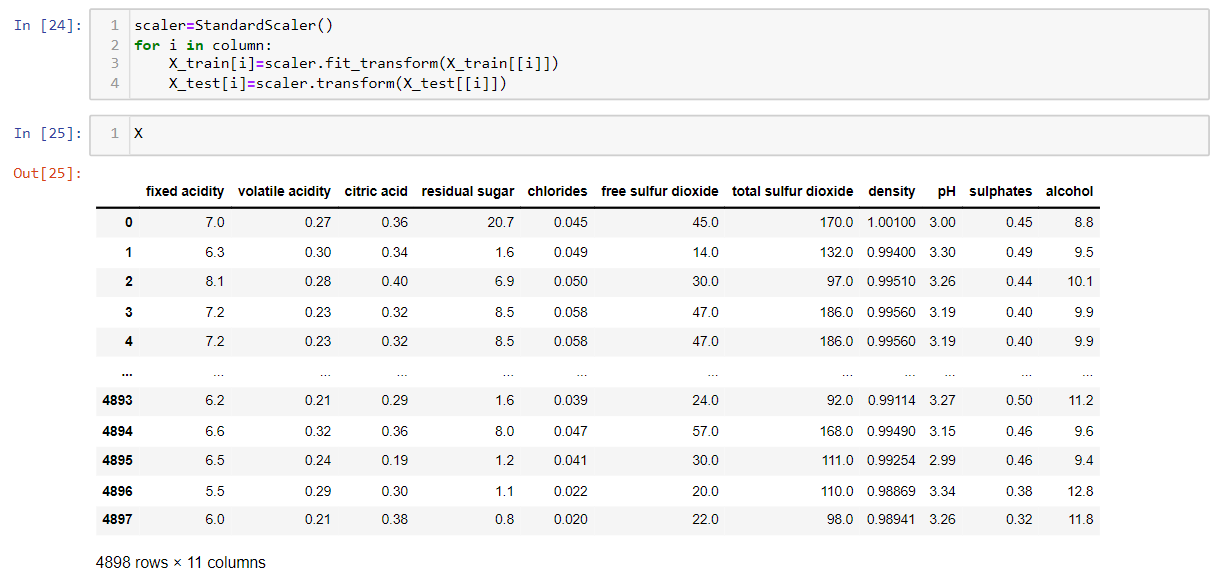
**4.3 SPLITTING DATASET FOR USE IN MODEL:**

****

****

**RANDOMLY SPLITING DATASET INTO TRAINING DATA & TESTING DATA:**

****

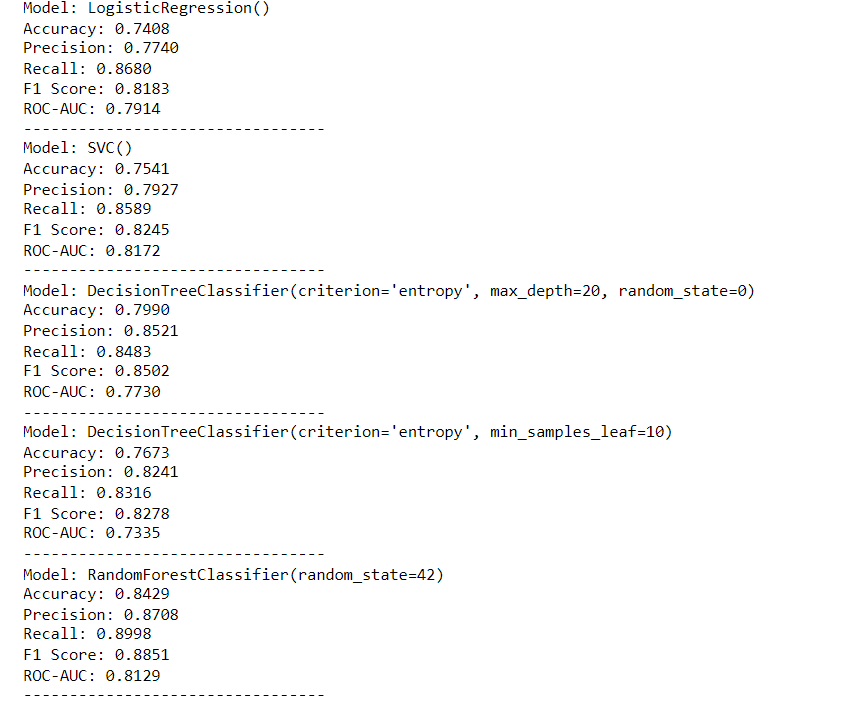
****

**CHECKING TRAINING & TESTING SCORE FOR DIFFERENT MODELS:**

****

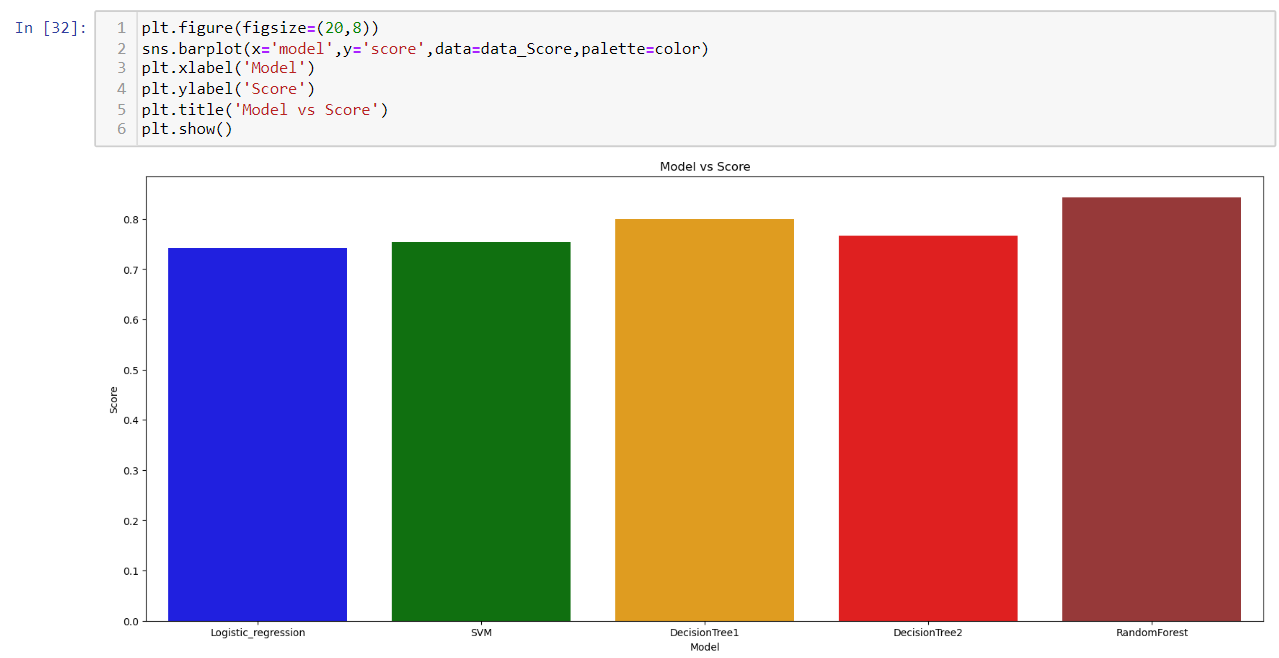
**4.4 EVALUATION METRICS**

****

****

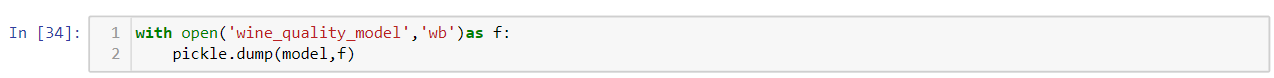
**ACCURACY COMPARISON BETWEEN DIFFERENT MODELS**

****

****

**Fig 4.4.1**

**STORING PYTHON OBJECT INTO FILE:**

****

**Chapter 5**

**FINDINGS & DISCUSSION**

- Through classification analysis, we discerned notable patterns in wine quality based on chemical attributes. For instance, wines with higher levels of sulphates and citric acid tended to be classified as higher quality, while those with elevated levels of volatile acidity were often categorized as lower quality.

- Although our classification model doesn't achieve complete accuracy (with a classification accuracy of 79%), it still demonstrates a significant relationship between wine quality and the analyzed chemical factors.

- Employing a supervised machine learning approach, our classification algorithm assigns wines to discrete quality categories rather than predicting continuous values. Despite its limitations, classification analysis serves as a valuable tool for discerning quality distinctions among wines based on their chemical composition.

**Chapter 6**

**IMPLEMENTATION**

The implementation of this project involved the following steps: -

• Import the necessary packages

• Import the required dataset and clean the dataset

• Visualize the dataset

• Split the dataset into a training and test set

• Apply the algorithm

• Generate graphs comparing train and test accuracies

**Chapter 7**

**CONCLUSION**

- The classification model exhibits limitations, evident from a modest accuracy rate.

- The model's performance, indicated by metrics such as precision, recall, and F1-score, suggests room for improvement.

- Despite its shortcomings, the model demonstrates potential for making reasonable predictions regarding wine quality classification.

- Identification of outliers within the dataset underscores the need for further refinement and feature engineering to enhance model robustness.

- This study underscores the application of classification techniques in analyzing wine quality based on chemical attributes, highlighting areas for future optimization and model enhancement.