**Team 4:**

**DATS 6103: Summary Report**

**Professor Ning Rui**

**December 14, 2022**

**Predicting Wine Quality**

**Introduction**

Within the ever-evolving scene of information science and prescient analytics, machine learning has risen as an effective apparatus for unraveling perplexing designs and estimating results over assorted spaces. One such captivating application lies within the domain of viticulture, where the quality and characteristics of wines are scrutinized through a focal point of computational ability. This inquires about endeavors to dig into the subtleties of wine quality expectation, with a specific center on a dataset comprising 32,485 occasions, fastidiously curated from the world of oenology. The chosen dataset unwinds the complex exchange of 12 particular highlights or fixings, extending from citric corrosive to liquor substance, all of which serve as the bedrock for foreseeing the quality and sort of wines.

The core of our investigation lies within the application of machine learning calculations to distill important bits of knowledge from these physicochemical properties, inferred from laboratory-based tests. Our request points not as it were to foresee the quality of wines but moreover to disentangle the perplexing connections between these input factors and the resultant wine quality. A wise understanding of the physicochemical characteristics gets to be foremost in chiseling and precise and successful machine learning demonstrates wine quality expectation. Through this ponder, we look to address relevant questions that direct our expository travel, examining the profundities of the dataset to disentangle the privileged insights typified inside the chemical composition of wines. This investigation endeavors to contribute to the developing body of information at the crossing point of machine learning and oenology, advertising a nuanced viewpoint on foreseeing wine quality through the focal point of computational insights.

**SMART Questions**

1. Do certain types of wine (red or white) tend to have higher quality scores on average?
2. Can we quantitatively measure the correlations between all attributes and wine quality ratings?
3. What is the range of wine quality scores, and how can we improve this range through analysis and recommendations?
4. Can machine learning models accurately predict wine quality based on its chemical composition, and if yes, which algorithms perform the best?
5. How will understanding and improving wine quality benefit winemakers, distributors, and wine consumers?

**Literature Review**

In the dynamic realm of data-driven predictions, machine learning stands out as a powerful tool for identifying patterns and forecasting outcomes across diverse domains. One intriguing application is in the wine industry, where computers are harnessed to assess wine quality and characteristics.Previous research on wine quality prediction emphasizes the importance of employing advanced techniques, particularly machine learning, to discern patterns within wine-related data. This approach facilitates the creation of models that enhance our understanding of the factors influencing wine quality. Concurrently, technology is being leveraged in vineyards to improve grape cultivation and winemaking. Machine learning applications aid in tasks ranging from estimating grape yields to optimizing the fermentation process, contributing to precision farming and sustainable winemaking practices.

The physical and chemical properties of wine play a crucial role in determining its quality. Existing studies emphasize the significance of attributes such as citric acid, residual sugar, sulfur dioxide, pH levels, and alcohol content. These factors are scrutinized to understand their impact on taste, aroma, and overall quality. While literature provides insights into the influence of these properties, the specific examination of 12 different elements and their collective impact on wine quality remains a focus of our research.

Despite the wealth of information in existing literature, there's a need for further exploration of the intricacies within our specific dataset. Investigating how these 12 elements distinctly influence the taste of wine and comparing machine learning algorithms in this context will yield valuable insights. This ongoing research seeks to contribute to the evolving field of wine quality prediction, bridging gaps in understanding and offering new perspectives on the interplay between wine composition and overall quality

**Description Of Data**

The dataset obtained from Kaggle encompasses a comprehensive collection of 32,485 instances of wines, each characterized by 12 distinct features or ingredients. The primary objective of the dataset is to predict the quality of the wines, which is the designated target variable. The quality is stratified into seven classes, ranging from 3 to 9, offering a diverse spectrum of potential quality levels. The dataset includes a variety of features such as citric acid, residual sugar, sulfur dioxide, pH, alcohol, among others, which serve as key input variables for the analysis and modeling processes. The dataset's richness lies in its capacity to facilitate predictive modeling aimed at understanding and forecasting the quality of wines based on the specified features. The inclusion of diverse wine instances and the breadth of features contribute to the dataset's potential for exploring the intricate relationships between these factors and the resulting wine quality. This dataset could be a valuable resource for researchers, analysts, and enthusiasts in the wine industry seeking to gain insights into the factors influencing wine quality.

**Preparing The Data**

|  |  |
| --- | --- |
| **Column names** | **Data Types** |
| **Unnamed** | **int64** |
| **fixed acidity** | **float64** |
| **volatile acidity** | **float64** |
| **citric acid** | **float64** |
| **residual sugar** | **float64** |
| **chlorides** | **float64** |
| **free sulfur dioxide** | **float64** |
| **total sulfur dioxide** | **float64** |
| **density** | **float64** |
| **pH** | **float64** |
| **sulphates** | **float64** |
| **alcohol** | **float64** |
| **quality** | **int64** |
| **Type** | **object** |

After thoroughly reviewing and refining our wine dataset, we ensured that all variables were accurately represented by appropriate data types, streamlining them for our analysis. Fortunately, our dataset didn't contain any missing values, eliminating the need for handling or imputing null values.

In addition to preparing the data, we applied filters to align with our analytical objectives. For instance, we used statistical properties like the standard deviation of wine quality to filter out any outliers, ensuring that our analyses and visualizations are focused on the core characteristics of wines within a certain quality range. This approach not only helps in creating more comprehensible visualizations but also aims to emphasize wines that align more closely with the general characteristics sought after by consumers, potentially excluding extreme quality outliers.

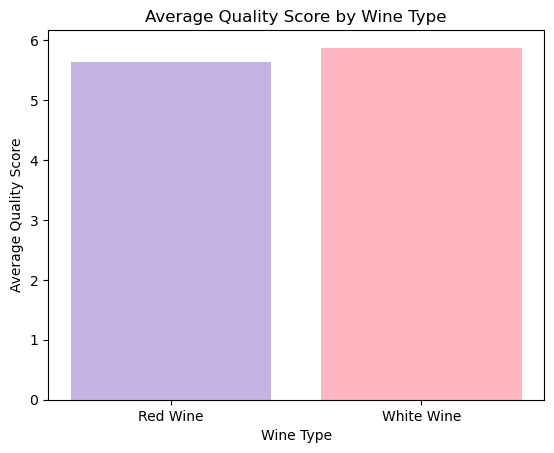
Now, having refined our dataset, we're ready to delve into the Exploratory Data Analysis (EDA) and commence our modeling procedures to extract valuable insights and patterns from the wine dataset.

**Analysis**

EDA and Graphs:

EDA is a method of analyzing data that employs visual techniques. It is used to detect trends and patterns, as well as to test assumptions, using statistical summaries and graphical representations.

**Q1- Do certain types of wine (red or white) tend to have higher quality scores on average?**



The above bar graph shows the average quality score by wine type. The average quality score is a measure of the overall quality of the wine, based on factors such as taste, aroma, and appearance. The wine types are red wine and white wine.Here we can see that white wine has a higher average quality score than red wine. The average quality score for white wine is 6, while the average quality score for red wine is 5.5. This suggests that white wine is generally considered to be of higher quality than red wine.

**Q2- Can we quantitatively measure the correlations between all attributes and wine quality ratings?**

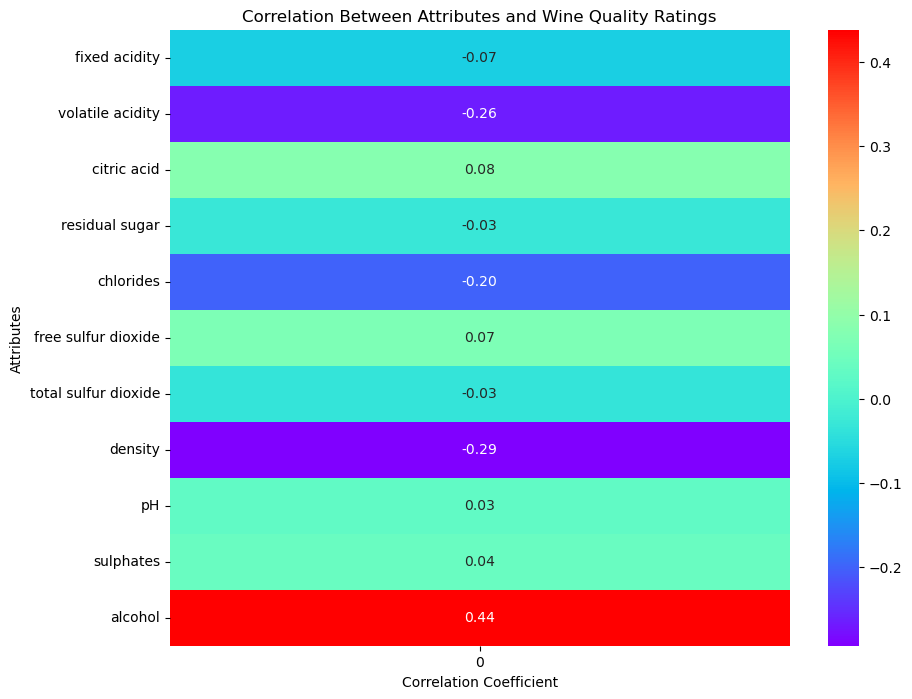


Fig 1

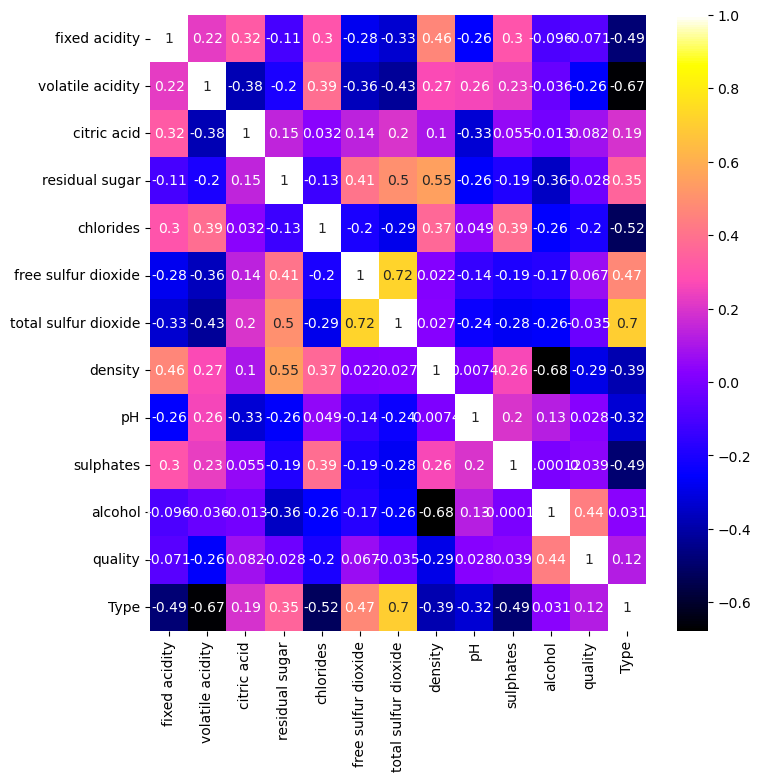


Fig 2

The correlation heatmap in fig 1 shows the correlation between the different wine attributes. The correlation coefficient is a measure of how strongly two variables are related, and it can range from -1 to 1. A correlation coefficient of 1 indicates a perfect positive correlation, while a correlation coefficient of -1 indicates a perfect negative correlation. A coefficient of 0 indicates no correlation between the two variables.

The heatmap shows that the following wine attributes have a strong positive correlation with each other:

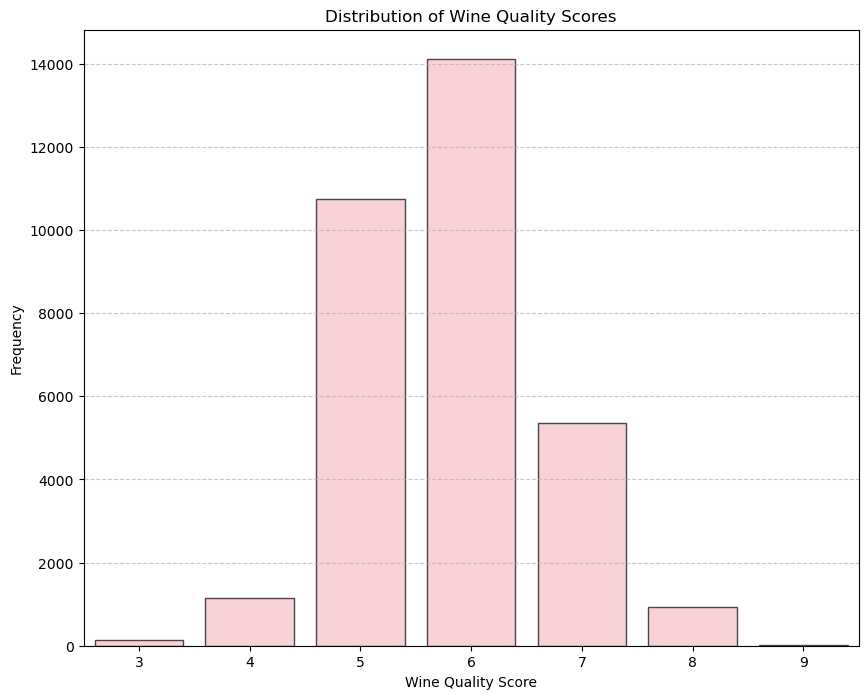
* Alcohol and fixed acidity
* Alcohol and citric acid
* Alcohol and free sulfur dioxide
* Alcohol and total sulfur dioxide
* Alcohol and pH

The heatmap shows that the following wine attributes have a strong negative correlation with each other:

* Volatile acidity and residual sugar, total sulfur dioxide,pH
* Chlorides and total sulfur dioxide , pH
* Density and pH

The heatmap in fig 2 shows that some wine attributes have a weaker correlation with each other. For example, the correlation coefficient between residual sugar and total sulfur dioxide is 0.28, which indicates a weak positive correlation. This means that wines with higher residual sugar content tend to have slightly higher total sulfur dioxide levels.

**Q3 - What is the range of wine quality scores, and how can we improve this range through analysis and recommendations?**



The bar graph Analyzes the Range of Wine Quality Scores for Enhancement and Recommendations.This graph shows the distribution of wines by quality, with the quality score on the x-axis which ranges from 3 to 9 and the count of wines on the y-axis. We can see that the majority of wines have a quality score of 5 or 6, with a smaller number of wines having a lower quality score. The count of wines decreases as the quality score increases, with the fewest wines having a quality score of 9.

**Q4 - Can machine learning models accurately predict wine quality based on its chemical composition, and if yes, which algorithms perform the best?**

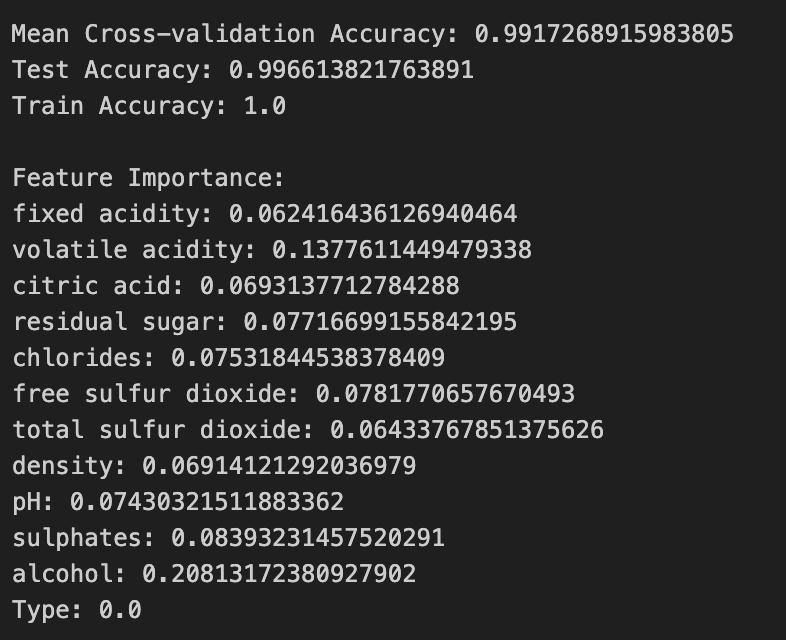
**Train-Test Split**

To evaluate the performance of our predictive model, we divided the available data into two sets: a training set (80% of the data) and a testing set (20% of the data). This division allows us to train the model on one subset and assess its predictive accuracy on another.

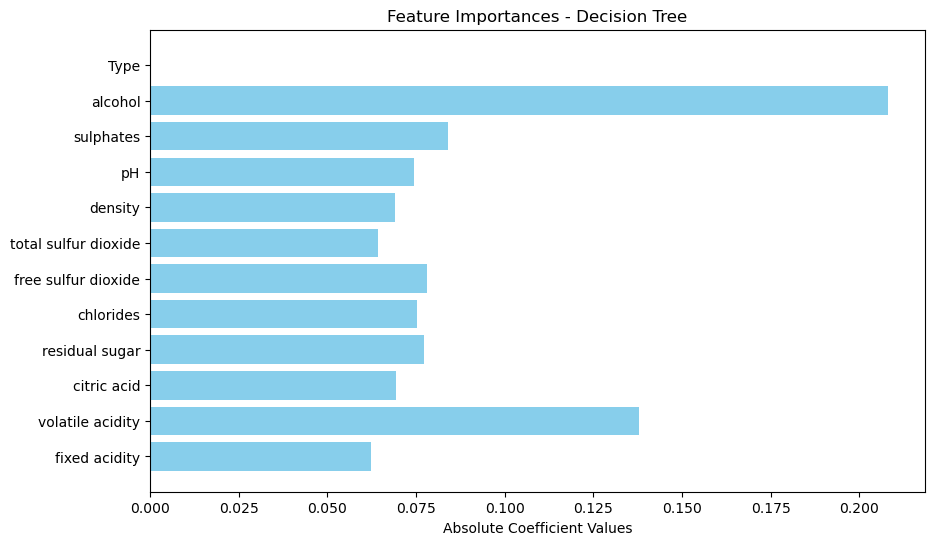
**Model Building and Evaluation:**

* **Part - 1** : **Considering all the wine attributes and performing model evaluation.**

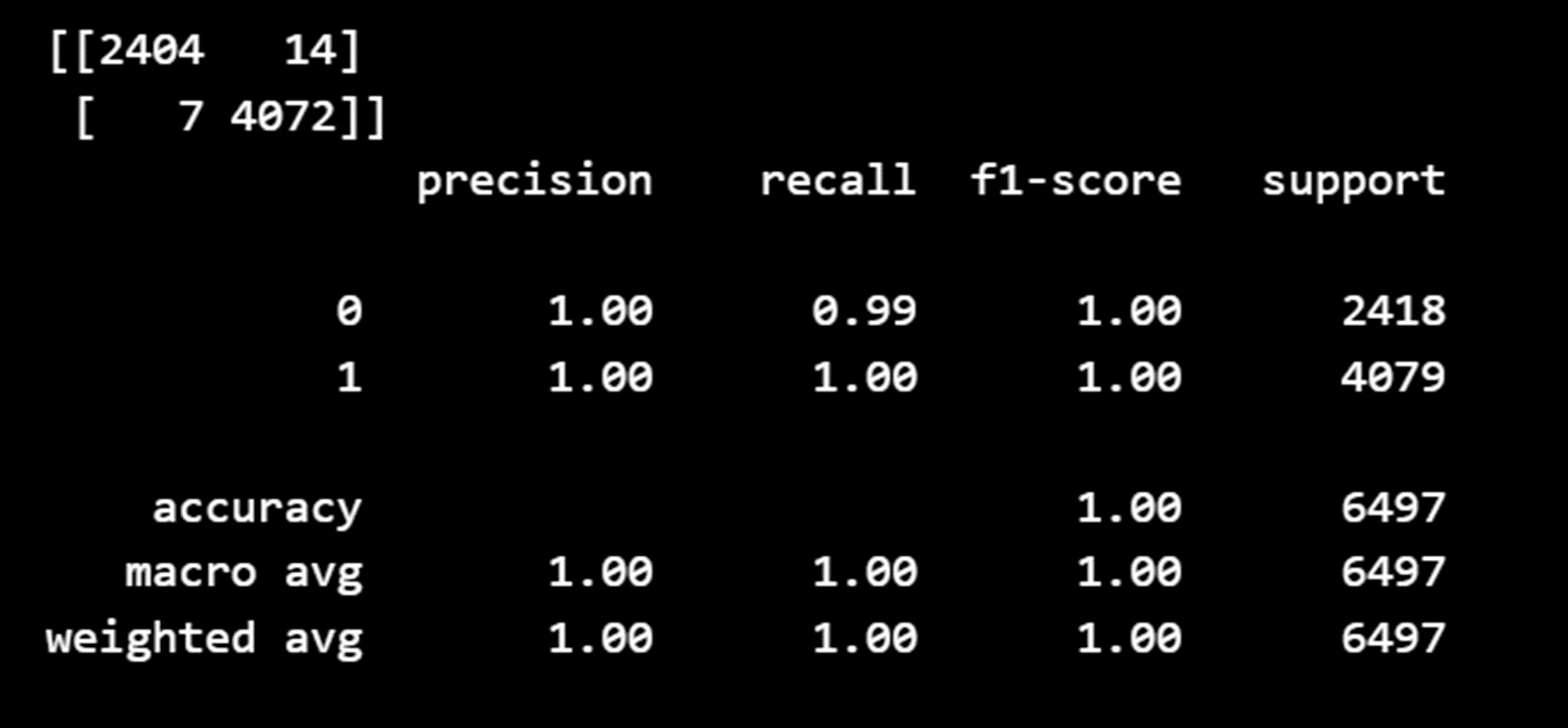
**1- Decision Tree:**



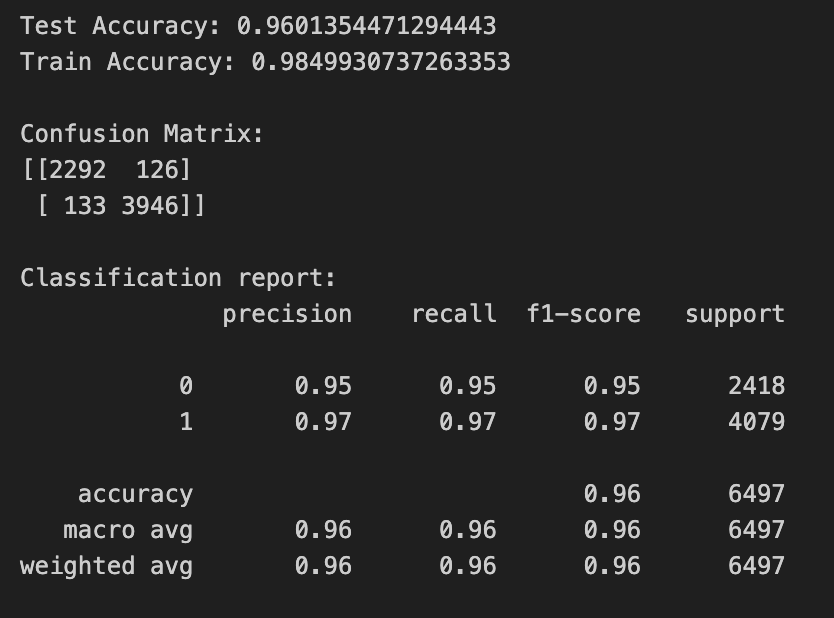
The model exhibits exceptional performance with near-perfect accuracy scores on both training and test datasets. Alcohol content seems most influential in predicting wine quality, while the type shows no discernible impact based on the feature importance analysis.



Classification Report

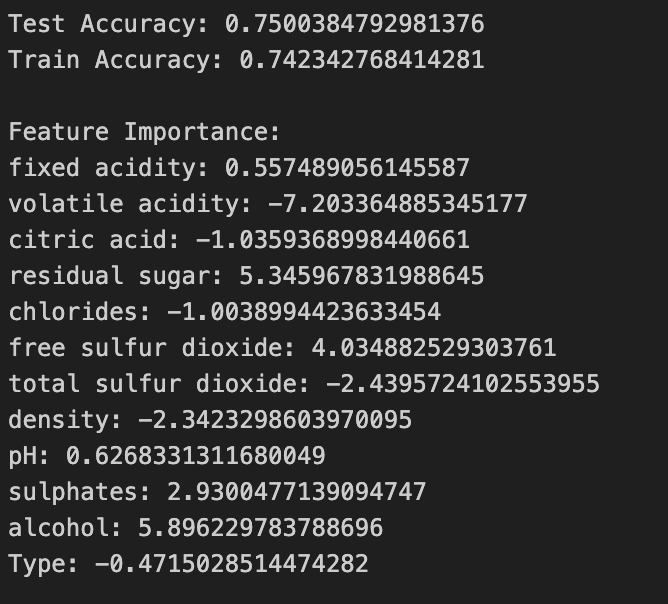


**2- KNN:**

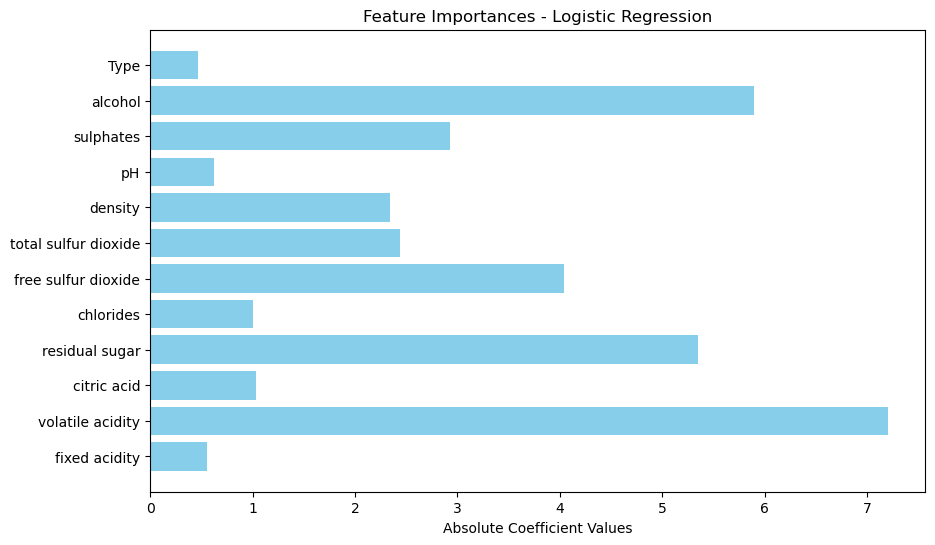


The model achieved high accuracy rates of approximately 96% on both training and test datasets, showcasing robust performance in distinguishing between the two classes. It demonstrates strong precision and recall for both classes, particularly excelling in correctly identifying instances of class 1.

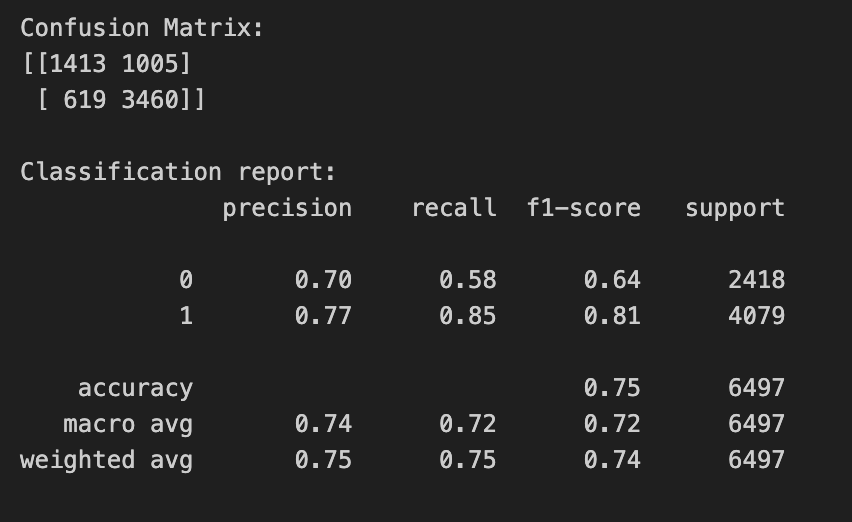
**3- Logistic Regression:**



The model's accuracy hovers around 75% for both training and test datasets, indicating consistent performance but with room for improvement. Some features, like fixed acidity, residual sugar, free sulfur dioxide, and alcohol, demonstrate more substantial influences on the model's predictions, while others like volatile acidity and chlorides seem to have a negative impact, potentially requiring further investigation or feature engineering.



Classification Report



* **Part - 2** : **Considering only the top 4 correlated features - Alcohol, density, chlorides, volatile acidity.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Decision Tree** | **KNN** | **Logistic Regression** |
| **Accuracy** | **0.99** | **0.95** | **0.73** |

**Accuracy:**

Accuracy is a measure of the overall correctness of a model. It's the ratio of correctly predicted instances to the total instances in the dataset. In your context, the accuracy values seem to correspond to different machine learning models applied to a wine dataset.

### **Decision Tree:**

A decision tree is a type of supervised learning algorithm used for both classification and regression tasks. It works by splitting the dataset into smaller subsets based on the most significant features. The accuracy of 0.99 for a decision tree on the wine dataset implies that the model correctly predicted the wine classes 99% of the time.

### **K-Nearest Neighbors (KNN):**

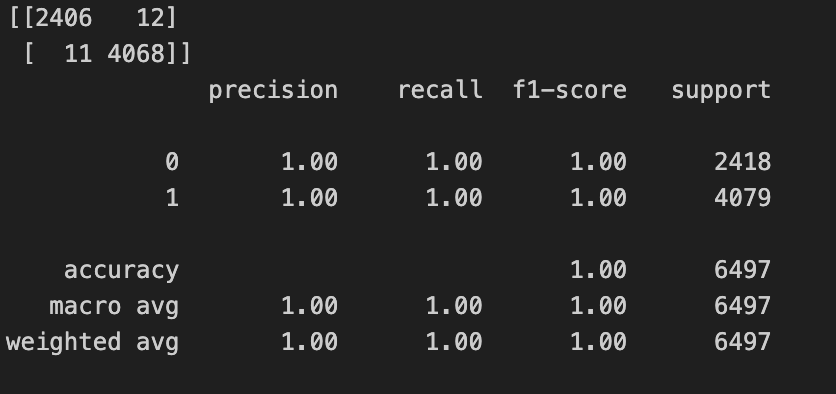
KNN is another classification algorithm that works by finding the 'k' nearest data points in the training set and classifying the new instance based on the majority class among its neighbors. An accuracy of 0.95 suggests that the KNN model predicted the wine classes correctly about 95% of the time.

### **Logistic Regression:**

Logistic regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable based on one or more predictor variables. It's commonly used for

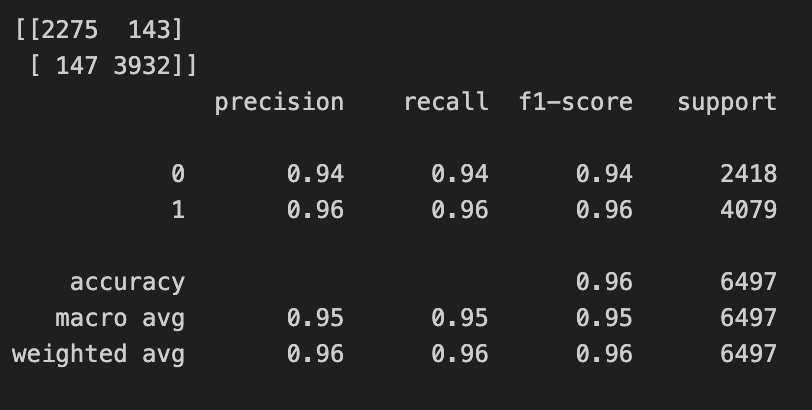
binary classification problems. An accuracy of 0.73 indicates that the logistic regression model predicted the wine classes with an accuracy of 73%.

1- Decision Tree:

Classification Report  


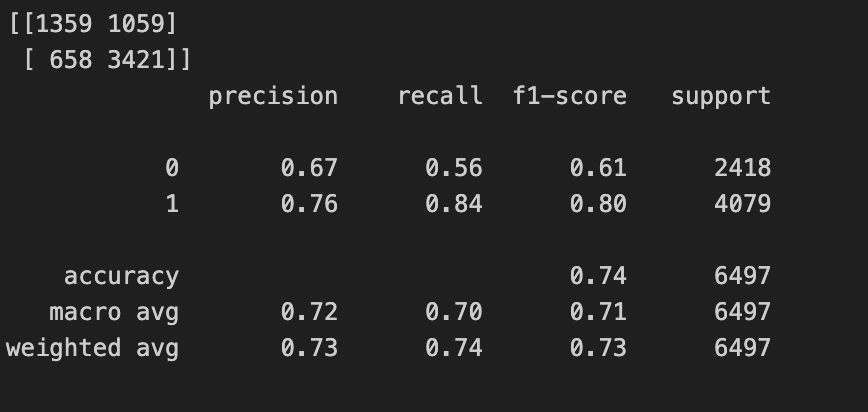
2- KNN

Classification Report

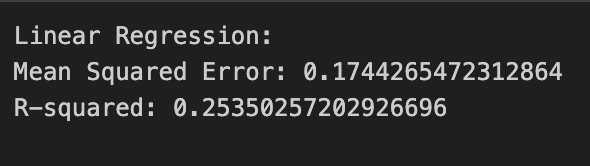


3- Logistic Regression

Classification Report

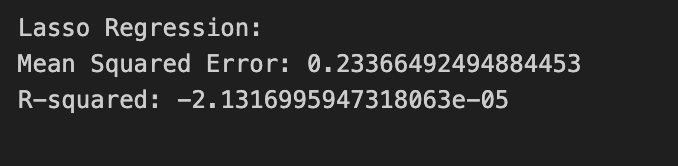


4- Linear Regression



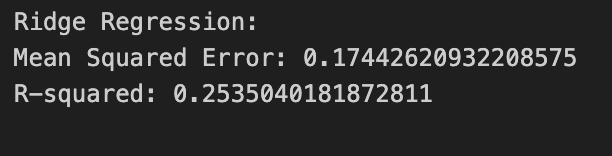
* The model's R-squared of 0.2 suggests limited explanatory power, explaining only a small portion of the variance in the data.
* Mean Squared Error of 0.17 means the average squared difference between predicted and actual values.

5- Lasso Regression



* The negative R-squared indicates either a poor fit of the model or possible overfitting problems, while the MSE of 0.23 indicates some degree of prediction error.
* This suggests that the relationships between predictors and the target variable in the dataset are not being adequately captured by the model.

6- Ridge Regression



* The R-squared for the Ridge Regression model was 0.25, meaning that the model explains around 25% of the variance in the target variable.
* The moderate amount of prediction inaccuracy is shown by the MSE of 0.17.
* The model performs better overall, exhibiting a moderate ability to capture interactions between predictors and the target variable.

**Q5 - How will understanding and improving wine quality benefit winemakers, distributors, and wine consumers?**

Grasping and enhancing wine quality brings substantial advantages to winemakers, distributors, and consumers alike. For winemakers, a nuanced comprehension of the physicochemical factors impacting wine quality allows for precise and targeted improvements in the winemaking process, resulting in the creation of superior wines. Distributors can benefit by providing a thoughtfully curated selection of wines that resonate with consumer preferences, thereby bolstering their competitiveness in the market. Consumers, in turn, experience a more delightful and satisfying taste journey with improved wine quality, fostering loyalty and trust in the brands they choose. Ultimately, this interdependent connection between understanding and enhancing wine quality establishes a positive cycle, enriching the overall wine-drinking experience and contributing to the industry's sustainable growth.

**Limitations & Other Considerations**

Data Quality: Predictions rely on a comprehensive and unbiased dataset; lacking these qualities may hinder generalization to real-world scenarios.

Feature Selection: Inaccurate or irrelevant features in the analysis can compromise model performance.

Model Overfitting: Complex models, especially Decision Trees, may overfit, learning noise from training data and struggling to generalize.

Interpretability: Despite effectiveness, Decision Trees and KNN lack interpretability, a concern in applications valuing clear rationale for predictions.

Duplicate Entries: Presence of duplicates in sulfur dioxide variables can introduce bias; addressing them is crucial for result integrity.

Domain Knowledge: Success relies on deep oenology understanding; misinterpreting physicochemical relationships may lead to errors.

Dynamic Wine Production:Dataset may not capture dynamic industry aspects, limiting the model's adaptability to evolving conditions.

Subjectivity in Rating:Subjective wine assessments introduce variability challenging to account for in the analysis.

External Factors: Analysis may overlook market trends, consumer preferences, or economic factors, impacting the wine industry.

**Conclusion**

According to our analysis, white wines are generally rated higher for quality than red wines, and higher evaluations are closely connected with higher alcohol concentration. Attributes like fixed acidity, citric acid, and sulfates also positively influence wine quality, while density and volatile acidity show negative correlations. The quality score distribution is centered around 6, indicating a widespread grading pattern. There are several duplicates in the dataset for the variables related to sulfur dioxide. Machine learning models, particularly Decision Trees and KNN Classifiers, perform well in predicting wine quality, with Decision Trees showing slightly better overall metrics than KNN, although both demonstrate reasonable predictive capabilities. Ultimately, putting an emphasis on quality results in a situation where all parties involved—producers, retailers, and consumers—benefit from the wine experience.