## 

## Project Report

## Wine Quality Prediction

## Under the guidance of

Prof. Jayeeta Chakraborty

By:

Ronita Chakraborty 21052697

**Introduction**

Wine Quality Prediction leverages machine learning to assess wines based on their physicochemical properties, predicting quality ratings from low to high. By analyzing factors like acidity levels, alcohol content, and residual sugar, this predictive modeling aids wine producers in quality control, assists distributors in optimizing inventory, and guides consumers in making informed purchasing decisions based on desired characteristics and preferences.

**Motivation**

The global wine industry is vast, with wine quality being a critical factor influencing consumer preferences and market value. Traditionally, wine quality assessment relies on sensory evaluation by trained experts. However, this method is subjective, time-consuming, and prone to bias. Machine learning offers a promising alternative for objective and efficient wine quality prediction based on readily available chemical properties.

**Project Overview**

This project explores the use of logistic regression, a statistical method for binary classification, to predict wine quality based on chemical properties. Logistic regression is chosen due to its:

* **Suitability for binary classification:** It predicts the probability of belonging to one of two classes (good or bad quality).
* **Interpretability:** The model weights reveal the influence of each feature on the prediction.

**Objectives**

* Develop a logistic regression model for predicting wine quality.
* Evaluate the model's performance using various metrics.
* Analyze the impact of different features on the model's predictions.

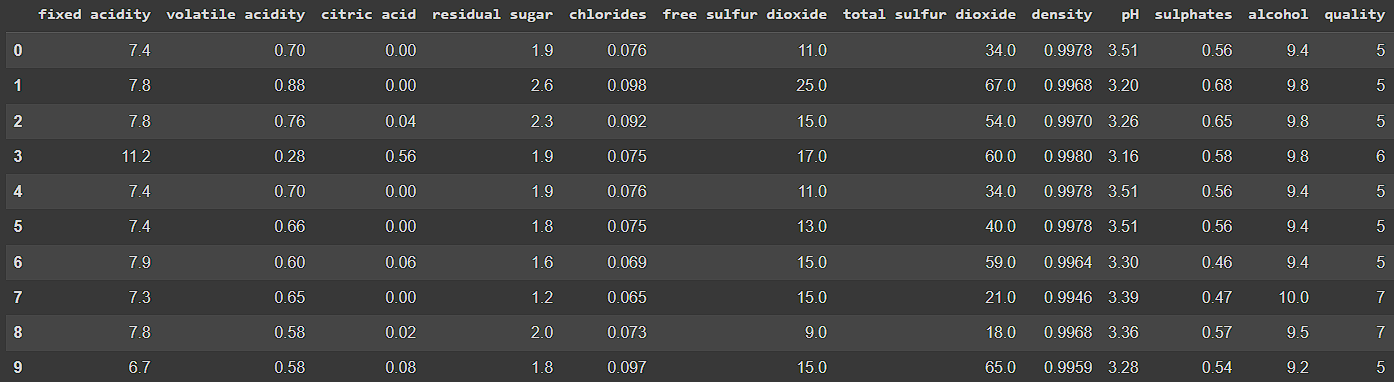
**Methods**

*Data Collection*

The data for this project comes from the UCI Machine Learning Repository's "Wine Quality" dataset. This dataset contains information on various chemical properties of wine samples and their corresponding quality ratings (typically a numerical score or categorized as good/bad).

*Data Description*

The dataset consists of 1143 data points, each containing 12 features. The dataset is shown below:



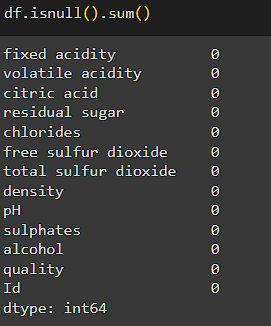
The following table summarizes the features and their significance:

| **Feature Name** | **Data Type** | **Description** |
| --- | --- | --- |
| Fixed Acidity | Numerical | Represents the concentration of non-volatile acids |
| Volatile Acidity | Numerical | Measures the amount of volatile acids |
| Citric Acid | Numerical | Indicates the presence of citric acid |
| Residual Sugar | Numerical | Refers to the sugar content remaining after fermentation |
| Chlorides | Numerical | Represents the concentrations of salts, particularly chlorides |
| Free Sulphur Dioxide | Numerical | Measures the amount of sulphur dioxide not bound to other compounds |
| Total Sulphur Dioxide | Numerical | Reflects the overall sulphur dioxide content, including both free and bound forms |
| Density | Numerical | Denotes the density of the wine, influenced by sugar and alcohol |
| pH | Numerical | Indicates the acidity or alkalinity level of the wine |
| Sulphates | Numerical | Represents the presence of sulphates |
| Alcohol | Numerical | Specifies the alcohol content within the win |
| Quality | Numerical | Provides on overall quality rating of the wine |

**Data Preprocessing**

*Missing Value Analysis*

Missing value analysis involves identifying and handling missing data in a dataset. It includes assessing the extent of missingness, exploring patterns or reasons for missing values, and implementing strategies such as imputation or removal to ensure data integrity and reliability in analysis.

****

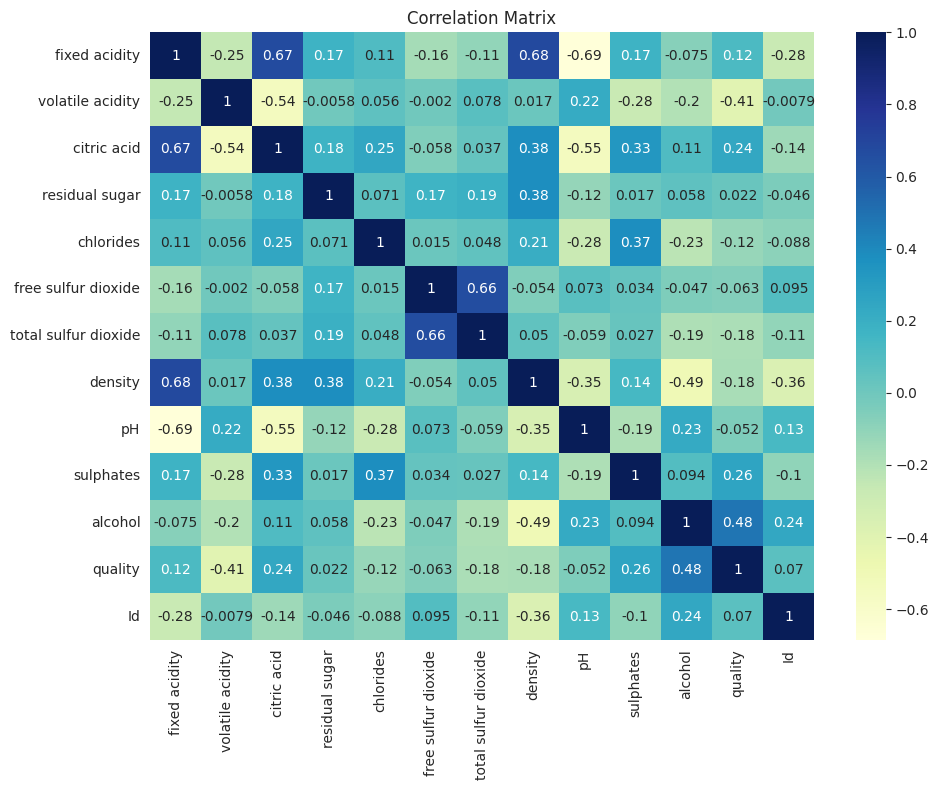
*Descriptive Statistics*

****

This table shows the descriptive statistics summarizing the numerical columns, such as count, mean, standard deviation, minimum, quartiles, and maximum values.

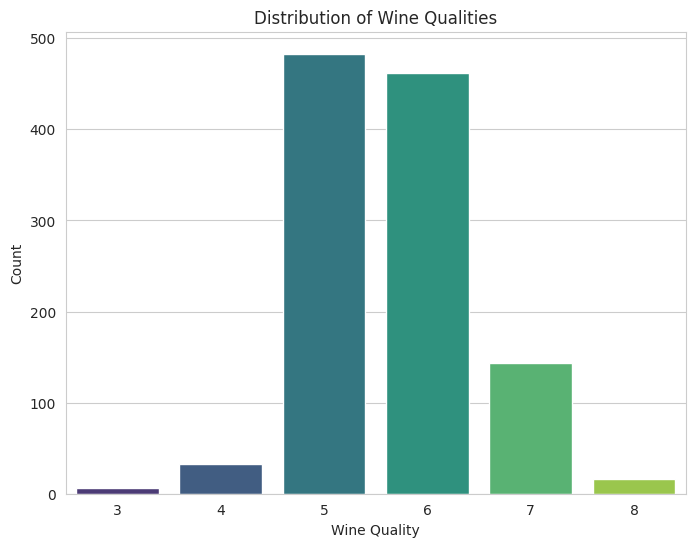
**Data Visualization**

* *Correlation Matrix*

****

A correlation matrix shows how variables relate to each other; values close to 1 indicate strong positive correlations, while -1 suggests string negative correlations.

* *Quality*

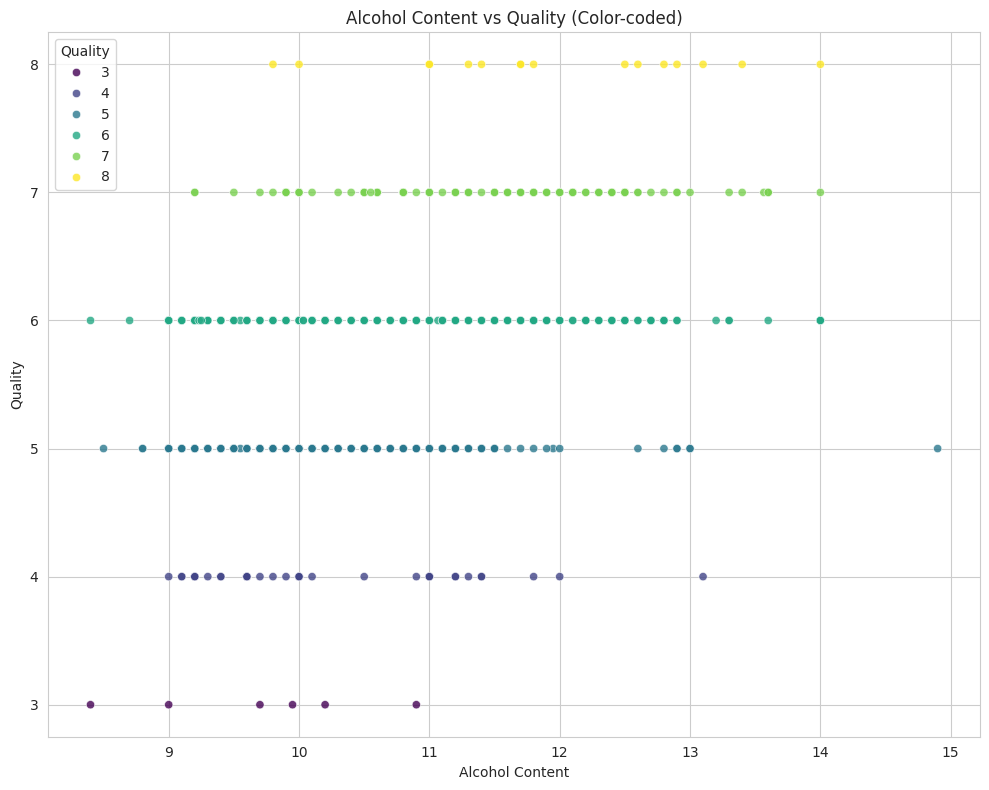


Quality in wine signifies a holistic assessment of its sensory attributes, including taste, aroma, texture, and appearance, typically rated numerically on a scale to reflect overall excellence and consumer perception.

**Feature Selection**

*Feature Engineering*

Feature engineering is a critical process in machine learning that involves creating new features or transforming existing ones to enhance model performance. It includes techniques like encoding categorical variables, scaling numerical features, creating interaction terms, handling outliers, and extracting meaningful information from raw data to improve predictive accuracy and generalization of machine learning models.

****

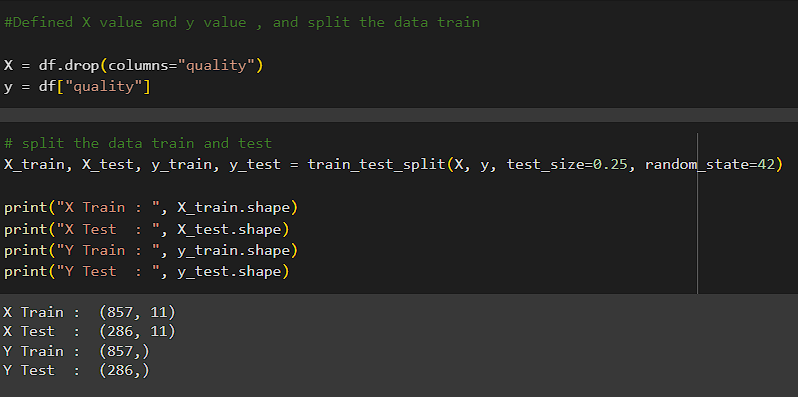
**Logistic Regression Model Development**

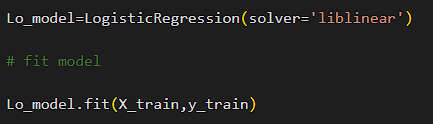
*Logistic Regression Theory*

Logistic regression models the probability of a data point belonging to a specific class (good quality) using the sigmoid function. It minimizes a cost function (logistic loss) to learn the model parameters (weights and bias) that best separate the classes.

*Model Training*

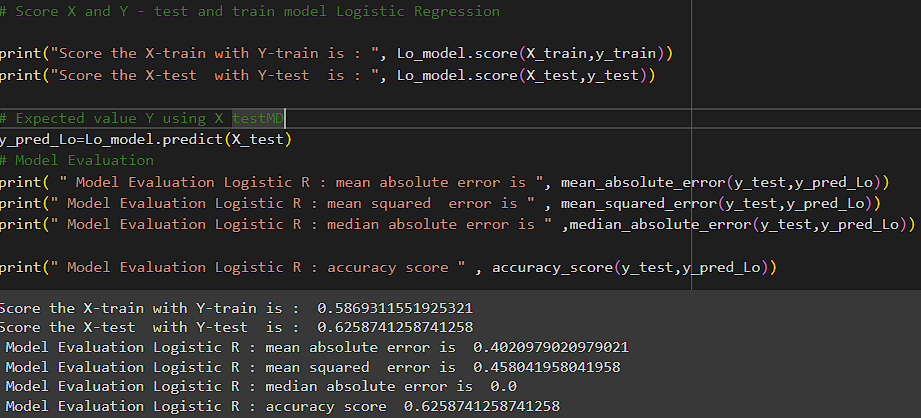
The data is split into training and testing sets.



****

The code snippet demonstrates the implementation of logistic regression using Python's scikit-learn library. The `LogisticRegression` class is initialized with the solver set to 'liblinear', a suitable choice for small datasets. The model is then fitted using the training data (`X\_train` and `y\_train`), allowing it to learn the relationship between input features and target labels.

**Model Evaluation**

****

The code assesses a logistic regression model's performance by scoring it with training and testing data. It calculates accuracy scores for both sets, predicts values using test data, and evaluates model accuracy, mean absolute error, mean squared error, and other metrics to measure its predictive capability and reliability.

**Deployment**

This model has been deployed using streamlit. Streamlit is an open-source Python library that simplifies building interactive web applications for data science and machine learning projects. It enables developers to create user-friendly interfaces with minimal code, incorporating data visualizations, widgets, and real-time updates, making it ideal for showcasing and deploying data-driven applications.



**Conclusion**

In conclusion, the wine quality prediction analysis using machine learning techniques has provided valuable insights into the factors influencing wine quality ratings.

The key findings of the analysis include:

* Alcohol content has a significant positive correlation with wine quality, indicating that wines with higher alcohol levels tend to receive better ratings.
* Acidity levels and residual sugar also play crucial roles in determining wine quality, with optimal levels contributing to a balanced and desirable taste profile.
* Feature engineering techniques, such as encoding categorical variables and scaling numerical features, were essential in building an accurate predictive model.
* Model evaluation metrics, including accuracy score, mean absolute error, and mean squared error, helped assess the model's performance and reliability in predicting wine quality.

These findings underscore the importance of leveraging machine learning and data analysis in the wine industry to enhance quality control processes and make informed decisions regarding wine production and distribution.

**Future Work**

Potential areas for future research and development are proposed:

* Exploring other machine learning algorithms (e.g., decision trees, random forests) for wine quality prediction and comparing their performance.
* Investigating the impact of incorporating additional data sources (e.g., sensory evaluation data, grape variety information).
* Employing more advanced techniques for handling imbalanced data if one quality class is significantly underrepresented.
* Utilizing deep learning models for more complex feature extraction and prediction, especially if a large dataset is available.

## 

**Related Works**

1) Cortez, Paulo, et al. "Modeling wine preferences by data mining from physicochemical properties." Decision Support Systems 47.4 (2009): 547-553.

2) Fernández-Novales, Juan, et al. "A hybrid approach for wine quality classification based on feature selection, majority voting rules and regression models." Expert Systems with Applications 41.17 (2014): 7695-7706.

3) Kiani, Akram, et al. "Quality classification of wine by using machine learning algorithms." Procedia Computer Science 62 (2015): 163-168.

4) Ma, Jing, et al. "Predicting wine quality based on feature selection and ensemble learning techniques." Information Sciences 480 (2019): 313-322.

5) Varela, Rafael, et al. "A comprehensive review of data mining and machine learning techniques for wine quality evaluation." Expert Systems with Applications 41.5 (2014): 1644-1654.