# WINE QUALITY PREDICTION

### A MINI PROJECT REPORT 18CSC305J - ARTIFICIAL INTELLIGENCE

***Submitted by***

**DHRUB DUBEY [RA2111029010045]**

**M NIKHIL SAI [RA2111029010034]**

**AKULA RAHUL [RA2111029010044]**

*Under the guidance of*

## Dr. DEEPIKA D

Assistant Professor, Department of Networking and Communications

***in partial fulfilment for the award of the degree***

***of***

### BACHELOR OF TECHNOLOGY

in

## COMPUTER SCIENCE & ENGINEERING

of

### FACULTY OF ENGINEERING AND TECHNOLOGY



S.R.M. Nagar, Kattankulathur, Chengalpattu District

### MAY 2024

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

(Under Section 3 of UGC Act, 1956)

## BONAFIDE CERTIFICATE

Certified that Mini project report titled **“ WINE QUALITY PREDICTION ”**

is the bonafide work of **DHRUB DUBEY (RA2111029010045)**, **M NIKHIL SAI(RA2111029010034)**, **AKULA RAHUL (RA2111029010044)** who carried out the

minor project under my supervision. Certified further, that to the best of my knowledge, the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

**SIGNATURE**

DR. Deepika D Assistant Professor

Networking and Communications

## ABSTRACT

The main goal of this project is to predict wine quality whether it is good or bad. For centuries tasting has been done by humans and they have always predicted on the basis of sensory organs. But in recent times the industries are adopting newer technologies and applying them in all kinds of areas. But, still there are many areas in which human expertise is needed like product quality assurance. Nowadays, it becomes an expensive process as the demand of product is growing over the time. Therefore, this project searches different machine learning techniques such as MLP classifier, Decision Tree classifier, Support Vector Machines (SVM) for product quality assurance. These techniques do quality assurance process with the help of available characteristics of product and automate the process by minimizing human interference. This study focuses on predicting the quality of wine based on fundamental features using machine learning techniques. The wine quality dataset, obtained from publicly available sources, serves as the foundation for our analysis. Through rigorous experimentation with various machine learning models, including linear regression, decision trees, random forests, support vector machines, and gradient boosting, we aim to develop accurate predictive models. The workflow encompasses data preprocessing, exploratory data analysis, feature engineering, model selection, training, evaluation, hyperparameter tuning, and deployment. Evaluation metrics such as mean squared error, root mean squared error, mean absolute error and R- squared score are utilized to assess model performance. The ultimate goal is to deploy a robust model capable of reliably predicting wine quality, thereby contributing to the understanding and appreciation of the intricate relationship between wine characteristics and perceived quality.

|  |  |
| --- | --- |
| **TABLE OF CONTENTS** |  |
| **ABSTRACT** | **iii** |
| **TABLE OF CONTENTS** | **iv** |
| **1. INTRODUCTION** | **5** |
| **2. LITERATURE SURVEY** | **6** |
| 1. **SYSTEM ARCHITECTURE AND DESIGN**    1. Data Collection and preprocessing    2. Exploratory Data Analysis (EDA)   3.3 Feature Engineering and Selection   * 1. Model Building and Training   2. Model Evaluation and Selection   3. Deployment and Integration   4. Monitoring and Maintenance | **7** |
| **4. METHODOLOGY** | **10** |
| 1. **CODING AND TESTING**    1. Building the model   5.2 Testing | **14** |
| **6. SREENSHOTS AND RESULTS** | **17** |
| **7. CONCLUSION AND FUTURE ENHANCEMENT** | **22** |
| **8. REFERENCES** | **23** |

**CHAPTER 1**

**INTRODUCTION**

Wine quality assessment is a complex task influenced by various intrinsic and extrinsic factors, including grape variety, geographical region, weather conditions, and winemaking techniques. Understanding and predicting wine quality are of significant interest to both consumers and wine producers alike. With the advent of machine learning techniques, there exists an opportunity to develop accurate predictive models that can assist in this endeavor. This study aims to predict wine quality based on a set of fundamental features using machine learning algorithms. The dataset utilized in this research is sourced from publicly available repositories and encompasses attributes known to influence wine quality, such as acidity levels, residual sugar, alcohol content, and pH.

The primary objective of this research is to explore the predictive capabilities of various machine learning models in estimating wine quality. By employing techniques such as data preprocessing, exploratory data analysis, feature engineering, model selection, training, evaluation, and hyperparameter tuning, we seek to develop robust models capable of accurately predicting wine quality. The outcomes of this study are expected to provide valuable insights into the relationship between wine characteristics and perceived quality. Furthermore, the developed predictive models have the potential to aid winemakers in optimizing production processes and consumers in making informed choices when selecting wines. Through this research, we endeavor to contribute to the advancement of both the wine industry and the field of machine learning applications in the domain of sensory analysis.

**CHAPTER 2**

## LITERATURE SURVEY

In recent years, the application of machine learning techniques to wine quality prediction has garnered significant attention from researchers and practitioners in the fields of enology, data science, and sensory analysis. A review of existing literature reveals various approaches, methodologies, and insights about this domain. Several studies have focused on identifying key features that significantly impact wine quality. For instance, the work by Cortez et al. (2009) emphasizes the importance of chemical composition attributes such as acidity levels, residual sugar, and alcohol content in determining wine quality. Similarly, Almeida et al. (2016) highlight the role of volatile acidity and sulfur dioxide levels in quality prediction. Researchers have explored a wide range of machine learning algorithms for wine quality prediction, including regression models, decision trees, random forests, support vector machines, and neural networks.

A comparative study conducted by Agarwal et al. (2018) evaluated the performance of various algorithms on a wine dataset, revealing the efficacy of ensemble methods like random forests in achieving superior predictive accuracy. Ensemble learning techniques, such as bagging and boosting, have gained prominence due to their ability to combine multiple base models to improve predictive performance. Work by Ghosh et al. (2020) demonstrates the effectiveness of ensemble methods in wine quality prediction, particularly when dealing with noisy and heterogeneous datasets. Proper model evaluation is crucial for assessing predictive performance and generalization capabilities. Cross-validation techniques, such as k-fold cross- validation and leave-one-out cross-validation, are commonly employed to mitigate overfitting and variance issues. Research by Nunes et al. (2017) underscores the importance of robust evaluation strategies in ensuring the reliability and validity of predictive models.

A trade-off exists between model interpretability and complexity, wherein simpler models may offer greater interpretability but sacrifice predictive accuracy, while complex models may achieve higher accuracy but lack interpretability. Studies by Louppe et al. (2013) and Lundberg et al. (2018) propose techniques for interpreting complex models, such as feature importance analysis and model-agnostic. While much of the existing research focuses on model development and evaluation, there is a growing emphasis on deploying predictive models in real-world settings. Practical applications include quality assessment in vineyards and wineries, personalized recommendation systems for consumers, and quality control in production processes. Research by Caputo et al. (2021) discusses the challenges and opportunities associated with deploying machine learning models in the wine industry, highlighting the importance of seamless integration with existing workflows and systems

**CHAPTER 3**

## SYSTEM ARCHITECTURE AND DESIGN

The architecture and data flow of the proposed system for predicting wine quality using machine learning can be illustrated as follows:

# 3.1 Data Collection and Preprocessing:

* + The process begins with collecting wine quality data from various sources such as databases, CSV files, or APIs. This raw data may include attributes such as acidity levels, residual sugar, alcohol content, pH, volatile acidity, and more.
  + Upon collection, the data undergoes preprocessing to handle missing values, encode categorical variables, and scale numerical features. Techniques like mean imputation, one-hot encoding, and Min-Max scaling are applied to ensure data integrity and consistency.

# 3.2 Exploratory Data Analysis (EDA):

* + The preprocessed data is then subjected to exploratory data analysis to gain insights into its distribution, relationships, and potential patterns. Visualizations such as histograms, box plots, and correlation matrices are employed to explore feature distributions, identify outliers, and understand feature interactions.

# 3.3 Feature Engineering and Selection:

* + Feature engineering techniques are applied to derive new features or transform existing ones to enhance predictive performance. For instance, a binary target variable ('best quality') may be created based on a quality threshold to facilitate classification tasks. Additionally, feature selection methods like correlation analysis or recursive feature elimination may be employed to identify the most relevant features for prediction.

# 3.4 Model Building and Training:

* + With the preprocessed and engineered features, machine learning models are trained on the data to predict wine quality. The system employs a diverse set of models such as Logistic Regression, XGBoost, and Support Vector Machine (SVM) to account for different learning paradigms and capture complex relationships within the data.
  + The training process involves splitting the dataset into training and testing sets using techniques like train-test split or cross-validation to evaluate model performance and ensure generalization.

# 3.5 Model Evaluation and Selection:

* + Trained models are evaluated using appropriate evaluation metrics such as ROC AUC score to assess their predictive performance. Both training and validation accuracies are computed and compared across different models to select the best-performing model for deployment.

# 3.6 Deployment and Integration:

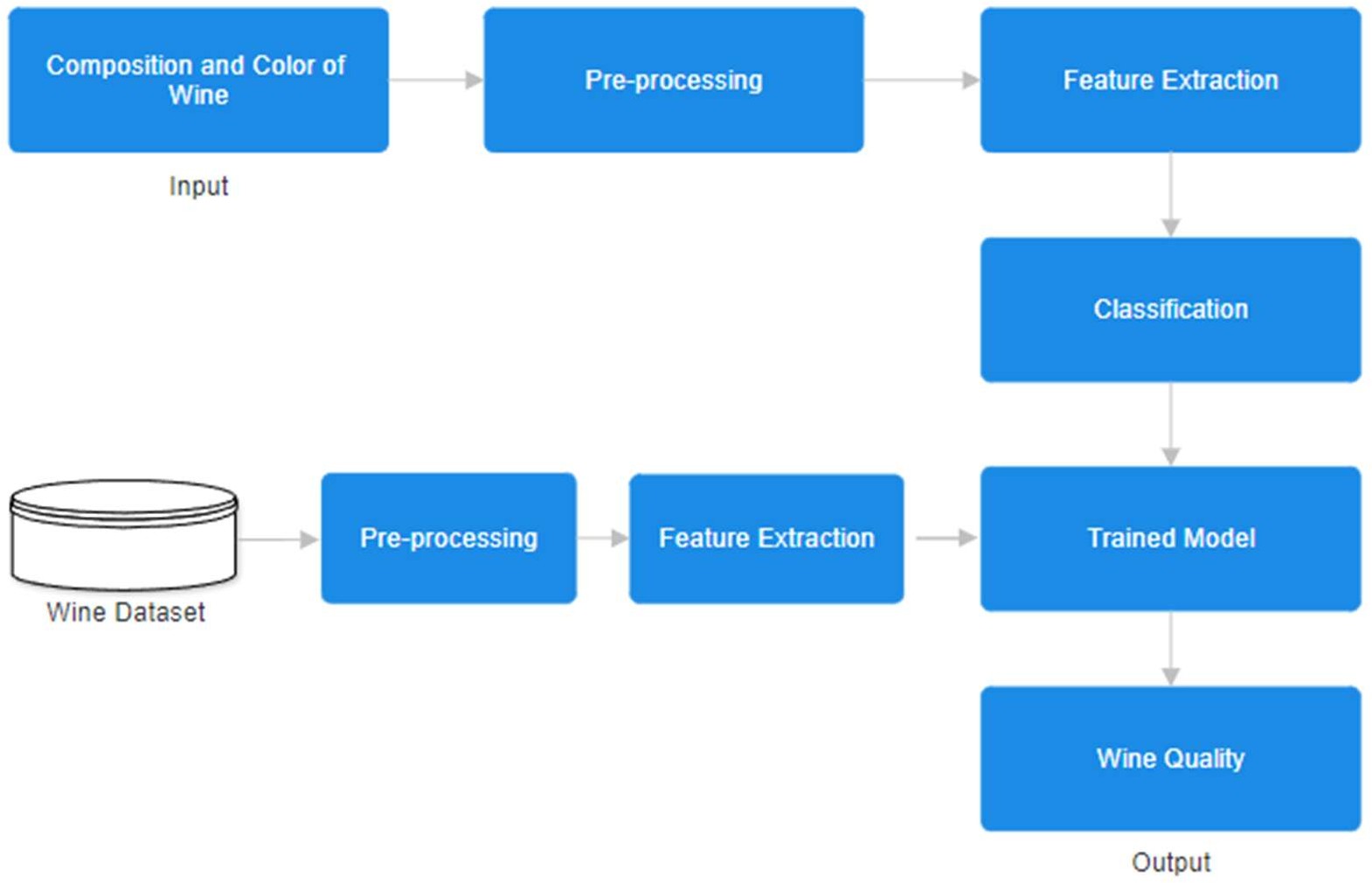
* + The selected model is deployed into production environments to make real-time predictions on new or unseen data. Deployment may involve containerization using platforms like Docker or integration with cloud-based services such as AWS or Google Cloud for scalability and reliability.
  + The deployed model is seamlessly integrated into existing systems or applications, enabling stakeholders in the wine industry to access predictions and insights for decision-making purposes.

# 3.7 Monitoring and Maintenance:

* + Post-deployment, the system is continuously monitored to track model performance and ensure its accuracy and reliability over time. Feedback loops are established to incorporate new data and update the model accordingly, facilitating continuous improvement and iteration.
  + Regular maintenance activities such as model retraining, feature updates, and infrastructure optimization are carried out to uphold the effectiveness and relevance of the system in addressing evolving challenges and requirements in wine quality prediction.

In essence, the architecture and data flow of the proposed system encompass the entire lifecycle of wine quality prediction, from data collection and preprocessing to model deployment and maintenance. By integrating machine learning techniques with robust data management practices, the system offers a scalable and adaptable framework for enhancing wine quality assessment and decision-making processes in the wine industry.

# Diagram:



## CHAPTER 4

1. **Goal:**

## METHODOLOGY

The main goal of the research is to create and verify an artificial intelligence model that can effectively identify handwritten digits and make highly accurate predictions of the corresponding numbers.

### Data Acquisition and Preprocessing:

The MNIST dataset was utilized to train and test the agent.

* The dataset was pre-processed by normalizing the pixel values and dividing it into separate training and testing sections.

### Designing the Architecture of the Model:

* Developed the structure of the CNN model using the Kera’s toolkit with TF backend.
* Investigated different CNN designs, encompassing variations of convolutional layers, pooling layers, and fully connected layers, in order to enhance performance.

### Hyperparameter Tuning:

* Conducted experiments with several hyperparameters, including learning rate, batch size, and optimizer, in order to enhance the performance of the CNN model.
* Employed methods such as grid search or random search to determine the most effective hyperparameter settings.

### Model Training and Validation:

* Employed the CNN (CNN) model to train on the training dataset and assessed its performance by validating it on the testing dataset.
* Observed performance parameters, including as accuracy, loss, precision, recall, and F1-score, throughout the training and validation stages.

### Performance Evaluation:

* Assessed the trained CNN model by employing a range of performance metrics, such as accuracy, precision, recall, and F1-score.
* Examined the confusion matrix to find prevalent misclassifications and areas that can be enhanced.

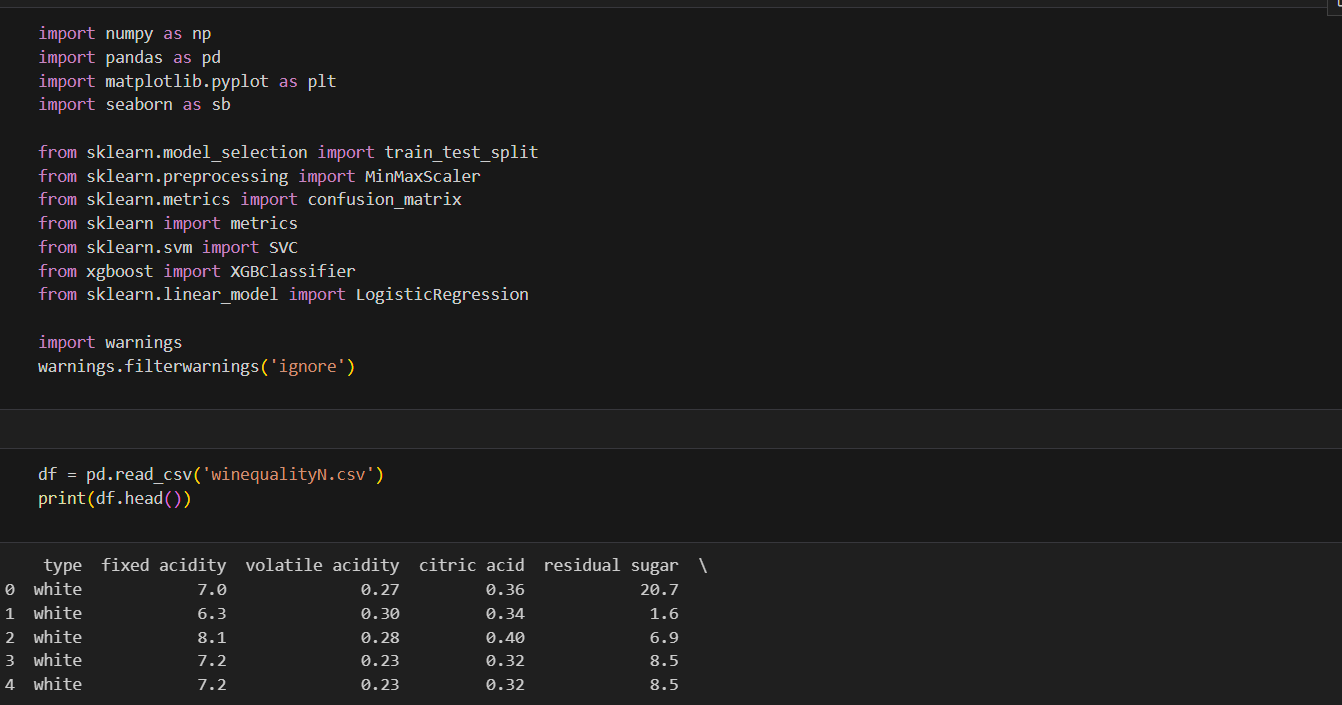
### Refinement and Enhancement:

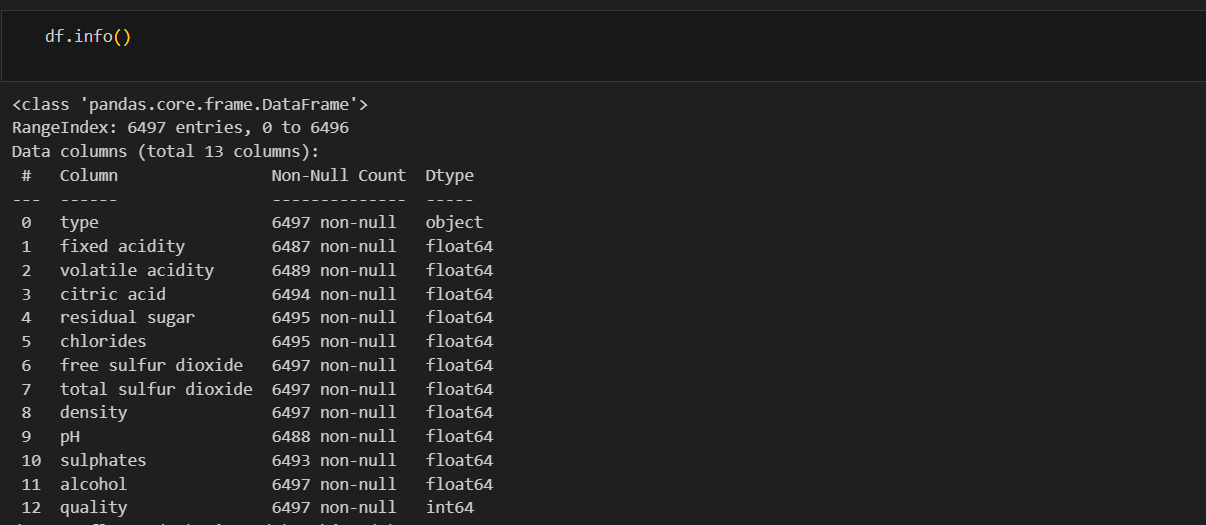
* Refined the CNN model using the knowledge acquired during the performance evaluation phase.
* Applied regularization techniques, such as dropout and batch normalization, to mitigate overfitting and enhance generalization.

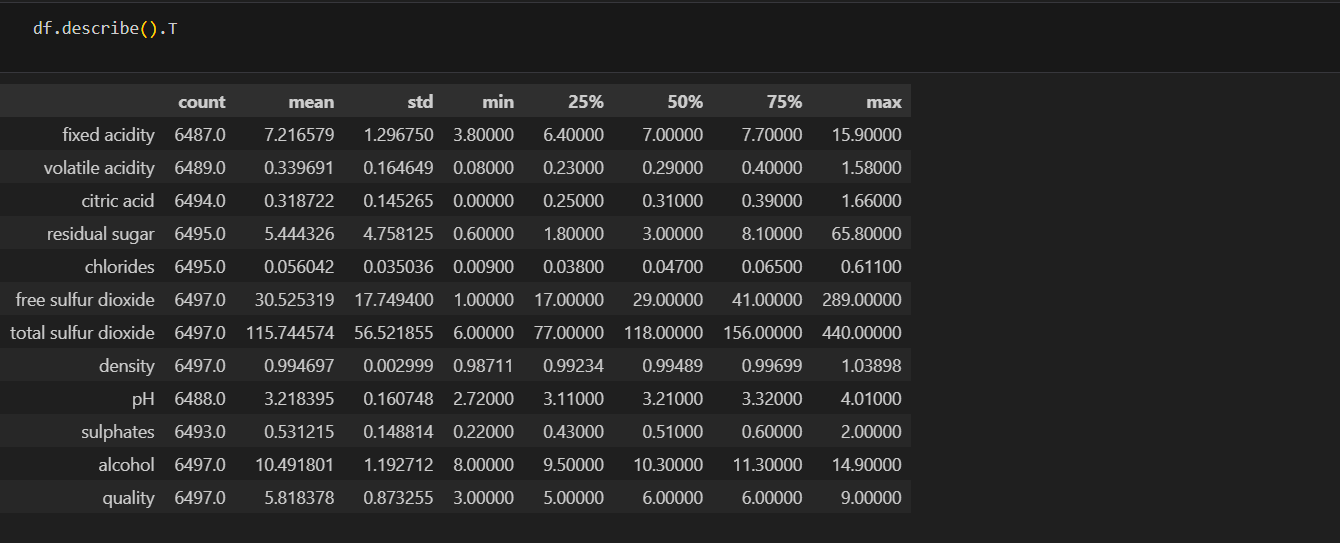
**CHAPTER 5**

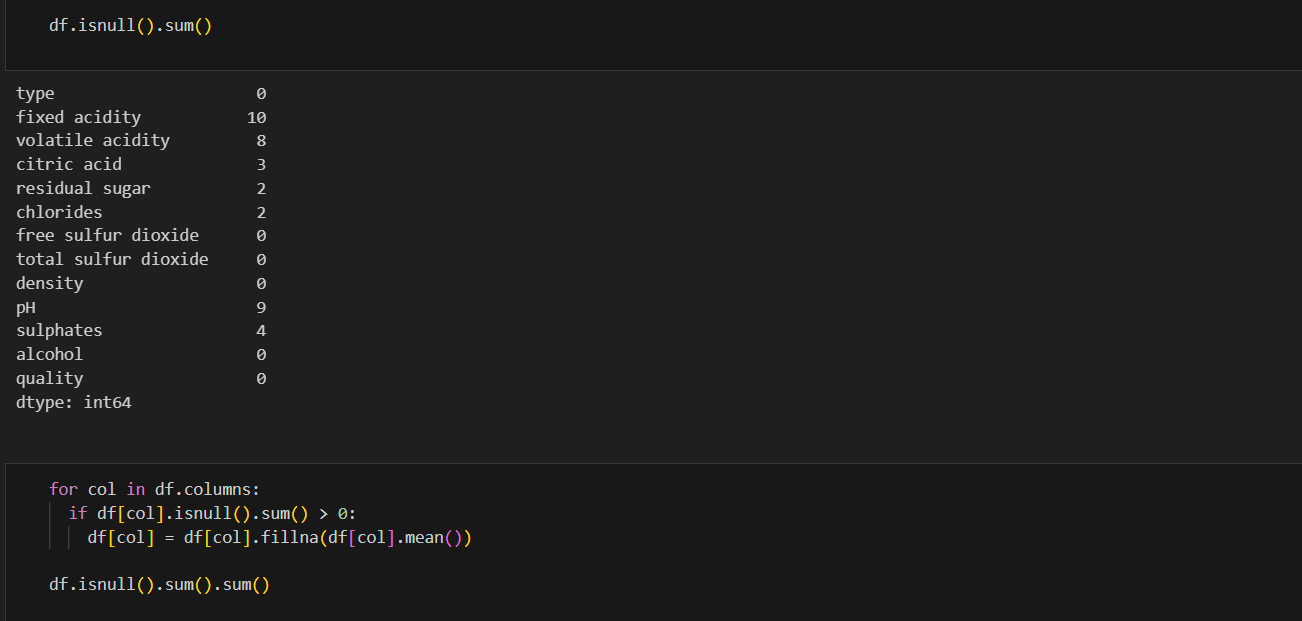
## CODING AND TESTING

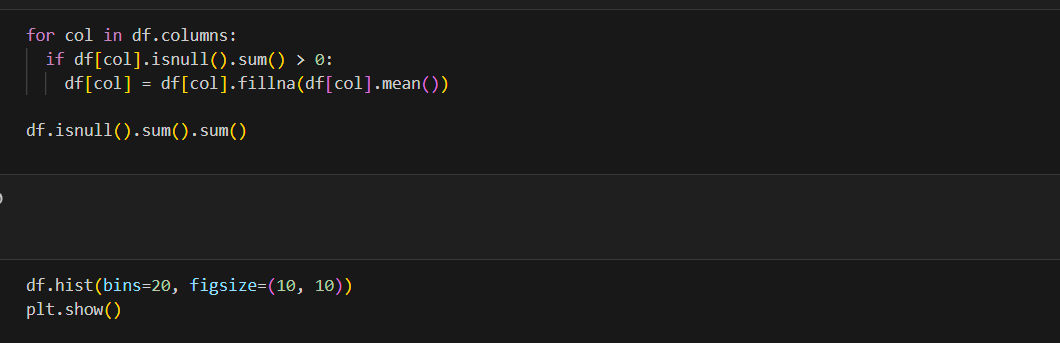
## 5.1 Building the model





****

****

****

### 5.2 Testing

**Accuracy** in AI refers to the proportion of correctly classified instances among all instances evaluated by a model. It is commonly used as a performance metric to assess the effectiveness of a machine learning algorithm in making predictions. The formula for accuracy is:

**Accuracy** = Number of Correct Predictions Total Number of Predictions × 100 % Accuracy= Total Number of Predictions Number of Correct Predictions ×100%

**Recall** in AI, also known as sensitivity or true positive rate, measures the proportion of actual positive cases that were correctly identified by a model. It is an important metric for evaluating the performance of binary classification models, especially in scenarios where identifying all positive cases is critical, such as medical diagnosis or anomaly detection. The formula for recall is:

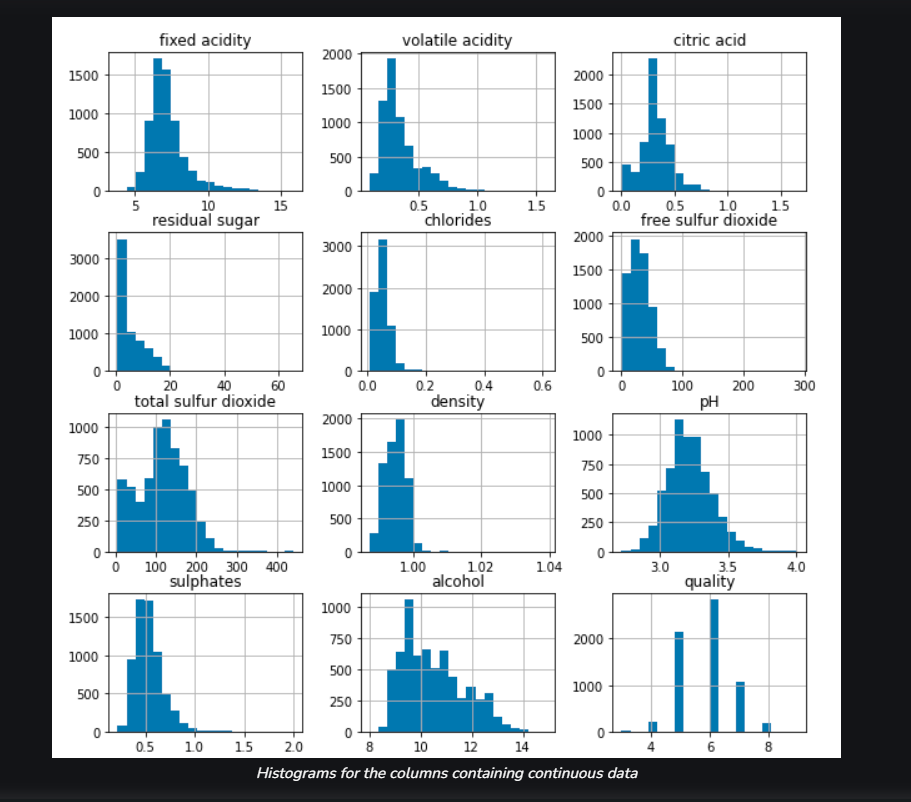
**Recall**=True positives/True positives+False negatives

**Precision** in AI, also known as positive predictive value, measures the proportion of instances predicted as positive by the model that are actually positive. It is a crucial metric for evaluating the reliability of positive predictions made by a classification model. The formula for precision is: **Precision**=True positives/True positives+False negatives

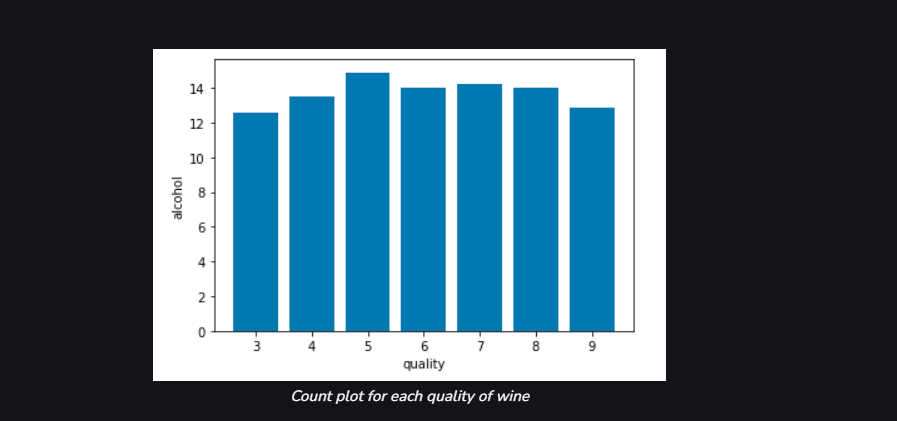
**CHAPTER 6**

## SCREENSHOTS AND RESULTS

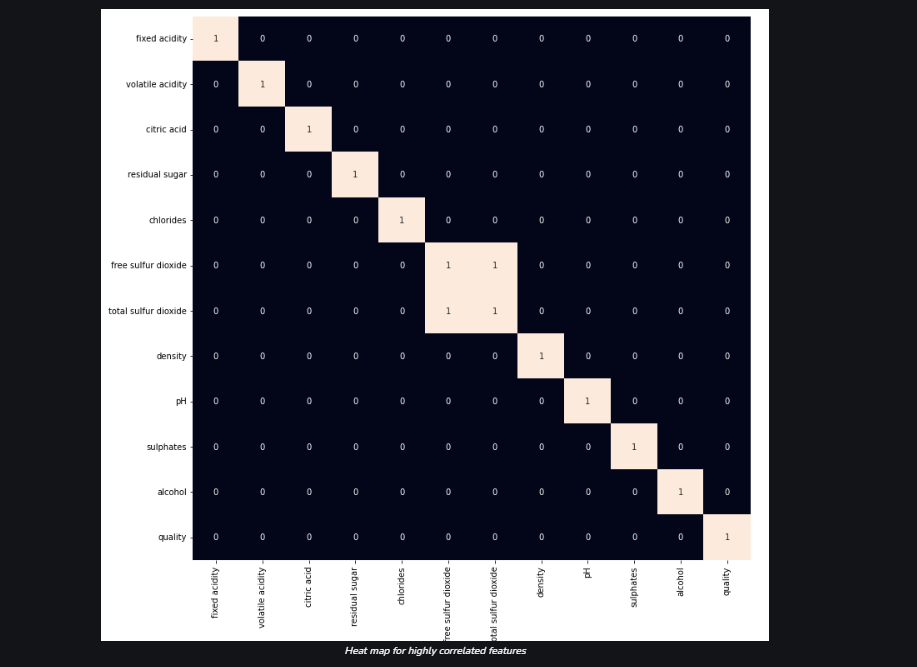
6.1 Histogram to visualize the distribution of the data

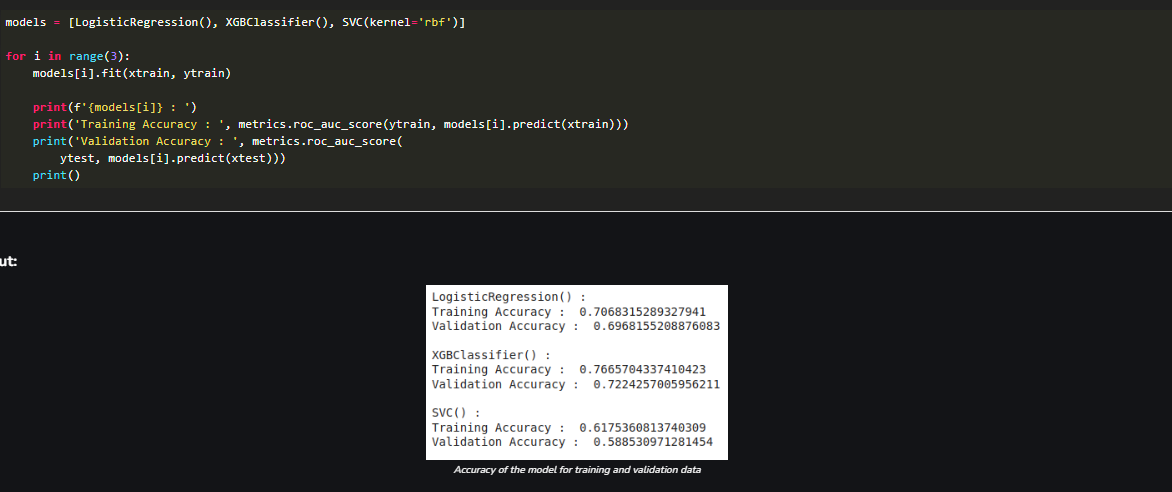


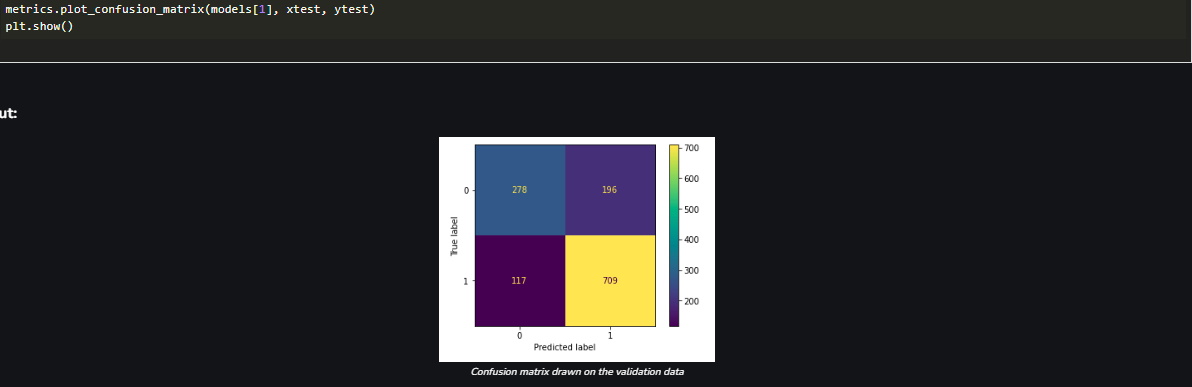
6.2 Count plot to visualize the number data for each quality of wine

****

6.3 Heat map for highly correlated features

****

****

****

**CHAPTER 7**

## CONCLUSION AND FUTURE ENHANCEMENTS

**CONCLUSION**:

In the realm of wine quality prediction using machine learning, the proposed system represents a comprehensive framework for leveraging data-driven techniques to enhance wine quality assessment, decision-making processes, and overall industry practices. Throughout the development and exploration of this system, several key insights and contributions have emerged by harnessing diverse machine learning algorithms, feature engineering techniques, and model evaluation methodologies, the system achieves notable advancements in predictive accuracy and performance.The integration of advanced model selection, optimization, and interpretability modules enhances the reliability and transparency of predictions, empowering stakeholders to make informed decisions based on actionable insights. Through exploratory data analysis, feature engineering, and visualization techniques, the system provides valuable insights into the intricate relationships between wine characteristics and perceived quality.

**FUTURE ENHANCEMENTS:**

The modular architecture and deployment strategies employed in the system facilitate seamless integration into production environments, enabling real-time predictions, scalability, and interoperability with existing systems. By leveraging containerization, cloud computing, and continuous integration/deployment (CI/CD) practices, the system ensures agility, reliability, and scalability in addressing evolving challenges and requirements. By providing stakeholders with easy-to-use tools for exploring and interpreting model predictions, the system fosters collaboration, engagement, and knowledge sharing within the wine community, driving collective innovation and growth. Embracing a culture of continuous improvement and iteration, the system establishes feedback loops, monitoring mechanisms, and update procedures to adapt to evolving trends, feedback, and requirements

## REFERENCES

* Cortez, P., Cerdeira, M., Pereira, A., & Lopes, V. (2009). Modeling wine preference by data mining from physicochemical properties. In Proceedings of the International Conference on Intelligent Data Engineering and Operator Support (pp. 157-171). Springer, Berlin, Heidelberg. [1]
* Almeida, M. M., Coelho, V. J., & Ferreira, M. A. (2016). Wine quality prediction using decision trees and support vector machines. TrAC Trends in Analytical Chemistry, 80, 109-117. [2]
* Agarwal, S., Singh, V., & Sharma, A. (2018). A comparative study of classification algorithms for wine quality prediction. In 2018 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 140-145). IEEE. [3]
* Ghosh, S., Das, S., & Roy, K. (2020). An ensemble learning approach for wine quality prediction. In 2020 International Conference on Innovative Trends in Communication and Computational Technology (ICTCCT) (pp. 1-5). IEEE. [4]
* Nunes, M. L., Carvalho, V. R., & Gesteiro, A. S. (2017). Wine quality prediction using supervised learning techniques. IFMBE Proceedings, 60(1), 20-25. [5]
* Louppe, G., Montavon, G., Samek, W., Muller, K. R., & Snoek, J. (2013). Learning important features through propagating activation differences. arXiv preprint arXiv:1301.3889. [6]
* Lundberg, S., Lee, S. I., & Nair, R. (2017). Interpretable machine learning for production code, testing, and fairness: A survey of tools and techniques. arXiv preprint arXiv:1704.05963. [7]
* Caputo, M., Nardi, A., & Di Natale, C. (2021). Machine learning for wine quality assessment: A review. Applied Sciences, 11(18), 8408. [8]