**Artificial Intelligence and Machine**

**Learning**

Project Report

Semester-IV (Batch-2022)

**WINE QUALITY PREDICTION MODEL**



**Supervised By: Submitted By:**

Mr. Kirandeep Singh Avikam Bir (2210990198)

Daksh Chawla (2210990243)

Ashish Karakoti (2210990186)

# Department of Computer Science and Engineering

Chitkara University Institute of Engineering & Technology, Chitkara University, Punjab

|  |  |  |
| --- | --- | --- |
| Sr. No. | Content | Remarks |
| 1. | Introduction |  |
| 2. | Background |  |
| 3. | Significance |  |
| 4. | Problem Definition and Requirements |  |
| 5. | Dataset |  |
| 6. | Features |  |
| 7. | Proposed Methodology |  |
| 8. | Result |  |
| 9. | Conclusion |  |
| 10. | References |  |

**Index**

# Introduction

Wine quality assessment is a significant aspect of the wine industry, influencing consumer preferences and market dynamics. Traditional methods of evaluating wine quality involve sensory analysis, which can be subjective and time-consuming. In recent years, machine learning techniques have emerged as powerful tools for predicting wine quality based on chemical and physical characteristics. This project aims to explore the application of machine learning algorithms to predict wine quality and assess its feasibility and accuracy.

|  |
| --- |
| **Background:** |

The quality of wine is influenced by various factors such as grape variety, geographical location, ph level, citric acid level, density, volatile acidity etc. Traditionally, wine quality assessment relied on expert judgment through sensory evaluation, including taste, aroma, colour, and mouthfeel. However, this approach is subjective and may lack consistency. With advancements in technology and the availability of large datasets containing chemical composition and sensory attributes of wines, machine learning has become increasingly popular for predicting wine quality.

Machine learning algorithms, such as regression and classification can analyse complex relationships between input features (e.g., pH, alcohol content, acidity) and wine quality scores. By training models on historical data, these algorithms can learn patterns and make predictions on new, unseen data. This project leverages machine learning techniques to build predictive models for wine quality assessment.

**Objectives**

The primary objectives of this project are as follows:

* Collect and preprocess a dataset containing chemical properties and sensory scores of wines.
* Explore and analyse the dataset to understand the relationships between input features and wine quality.
* Implement various machine learning algorithms, including regression and classification models, for predicting wine quality.
* Evaluate the performance of the models using appropriate metrics such as mean squared error, accuracy, and F1-score.
* Fine-tune the models and compare their performance to identify the most effective approach for wine quality prediction.

**Significance**

The significance of this project lies in its potential impact on the wine industry and related research fields:

* Improving Quality Assessment: Machine learning models can provide objective and consistent evaluations of wine quality, complementing traditional sensory analysis methods.
* Cost and Time Efficiency: Automated prediction of wine quality using machine learning algorithms can save time and resources compared to manual evaluations, especially for large-scale wine producers.
* Insights for Winemaking: By identifying key chemical factors influencing wine quality, this project can provide insights for winemakers to optimize production processes and enhance product consistency.
* Research Advancement: The application of machine learning techniques to wine quality prediction contributes to the advancement of research in both the fields of data science and oenology.

**Problem Definition and Requirements**

**Problem Statement**

The objective of this project is to develop machine learning models capable of accurately predicting the quality of wine based on its chemical composition and sensory attributes. Given a dataset containing features such as acidity, pH level, alcohol content, and sensory scores, the goal is to train regression or classification models that can predict the quality rating of wines on a numerical or categorical scale.

**Software Requirements:**

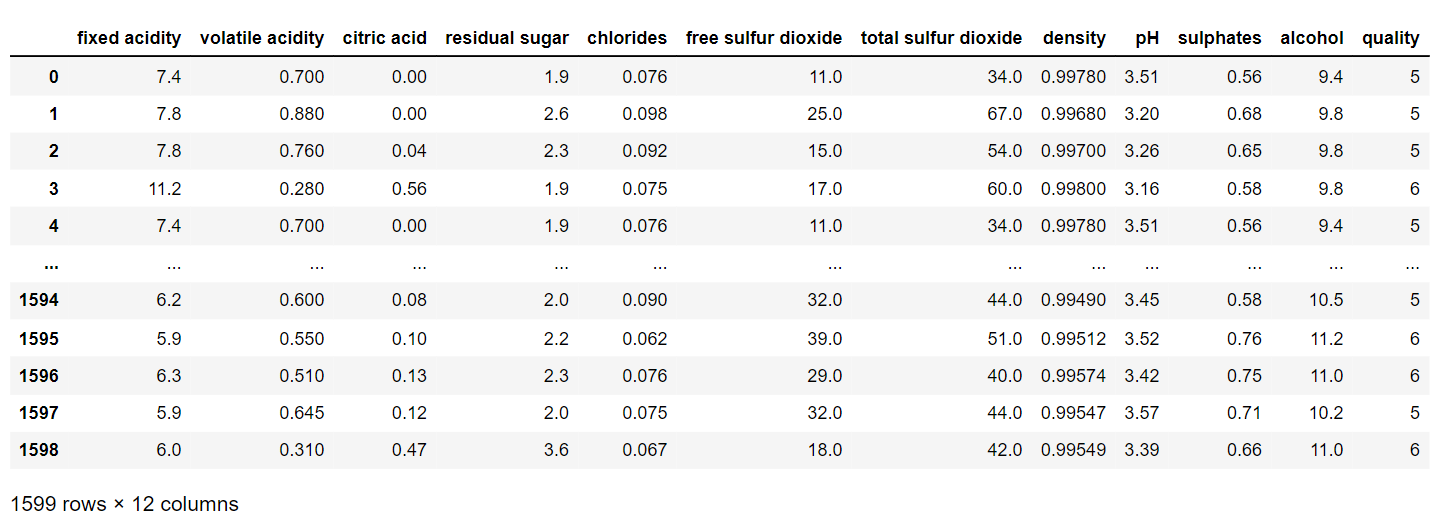
* **Programming Language:** Python will be the primary programming language due to its extensive libraries for machine learning and data analysis, including scikit-learn, pandas, and NumPy.
* **Development Environment:** Anaconda or a similar Python distribution will be used to manage dependencies and create virtual environments.
* **Machine Learning Libraries:** Scikit-learn will be utilized for building and evaluating machine learning models, while additional libraries such as TensorFlow or PyTorch may be considered for advanced modelling techniques.
* **Data Visualization Tools:** Matplotlib and Seaborn will be employed for data visualization to gain insights into the dataset and model performance.
* **Text Editor or Integrated Development Environment (IDE):** Jupyter Notebooks or IDEs like PyCharm will be used for coding and experimentation.
* **Version Control:** Git and platforms like GitHub will facilitate collaboration, version control, and project management.

**Hardware Requirements:**

* **Processor:** A multi-core processor to handle data preprocessing and model training efficiently.
* **Memory (RAM):** At least 8 GB of RAM to accommodate large datasets and machine learning algorithms.
* **Storage:** Sufficient storage space for storing datasets, code files, and model artifacts. SSD storage is preferable for faster data access and model training.
* **Graphics Processing Unit (GPU) (Optional):** While not mandatory, a GPU (NVIDIA GeForce or AMD Radeon) can significantly accelerate model training, especially for deep learning algorithms.
* **Operating System:** The project can be executed on Windows, macOS, or Linux-based systems, ensuring compatibility across different platforms

**Dataset**

* The dataset consists of 1599 samples with 11 physicochemical attributes and a quality rating.
* Exploratory data analysis (EDA) techniques, including histograms, KDE plots, boxplots, and correlation heatmaps, are utilized to understand data distribution, identify outliers, and explore feature relationships.



**Fig**: Dataset

**Features**

* **fixed acidity:** Primary fixed acids found in wine are tartaric, succinic, citric, and malic
* **volatile acidity:** Volatile acidity is the gaseous acids present in wine.
* **citric acid:** It is weak organic acid, found in citrus fruits naturally.
* **residual sugar:** Amount of sugar left after fermentation.
* **Chlorides:** Amount of salt present in wine.
* **free sulphur dioxide:** So2 used for prevention of wine by oxidation and microbial spoilage.
* **total sulphur dioxides:** Sum of free and combined SO2
* **density:** thickness of wine
* **pH:** pH is used for checking acidity
* **sulphates:** Added sulphites preserve freshness and protect wine from oxidation, and bacteria.
* **Alcohol:** Percent of alcohol present in wine.
* **Quality:** Target label with 6 classes in range 3-8

- This is a multiclass classification problem, as labels are discrete.

- There are 12 features (attributes).

- There are no null values, so we don't need to do Data manipulation

**Proposed Methodology**

1. **Data Collection and Preprocessing:**

* **Data Collection:** Gather a dataset containing information on the chemical properties and sensory attributes of wines, including features such as acidity levels, pH values, alcohol content, residual sugar, and sensory scores.
* **Data Preprocessing:**
* Handle missing values: Impute missing values using techniques such as mean or median imputation.
* Outlier detection: Identify and handle outliers using methods such as z-score, IQR (Interquartile Range), or domain-specific knowledge.
* Feature scaling: Scale numerical features to a similar range using techniques such as Min-Max scaling or standardization.
* Feature encoding: Encode categorical features using techniques such as one-hot encoding or label encoding.
* Train-test split: Split the dataset into training and testing sets to evaluate model performance.

1. **Model Development:**

* **Support Vector Machines (SVMs):**
* Implement SVM models with different kernels (e.g., linear, polynomial, radial basis function) to capture complex relationships between features.
* Experiment with different hyperparameters such as C (regularization parameter) and kernel coefficients to optimize model performance.
* **Decision Trees:**
* Develop decision tree models to partition the feature space based on the most informative features.
* Tune hyperparameters such as maximum depth, minimum samples split, and criterion (e.g., Gini impurity or entropy) to control model complexity and prevent overfitting.
* **Random Forests:**
* Construct random forest models consisting of multiple decision trees trained on random subsets of the data.
* Explore hyperparameters such as the number of trees, maximum depth, and minimum samples split to balance model complexity and accuracy.
* **Logistic Regression:**
* Build logistic regression models to predict the probability of wine quality classes.
* Experiment with regularization techniques to prevent overfitting.

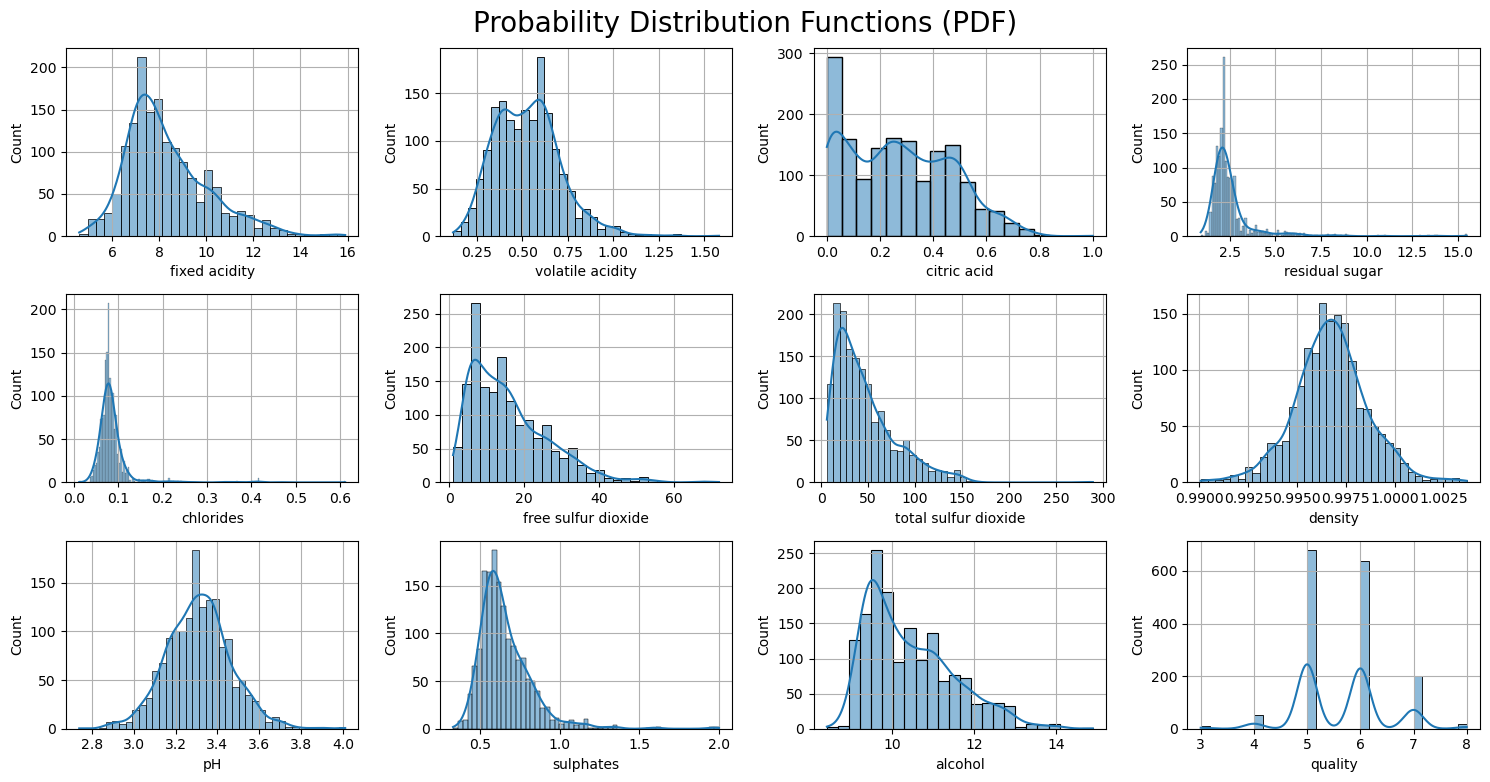
1. **Model Evaluation and Hyperparameter Tuning:**

* **Model Evaluation:**
* Evaluate the performance of each model using appropriate metrics such as accuracy, precision, recall, F1-score, and ROC-AUC (Receiver Operating Characteristic Area Under Curve).
* Use cross-validation techniques such as k-fold cross-validation to assess model generalization and variance.
* **Hyperparameter Tuning:**
* Conduct grid search or random search to identify the optimal hyperparameters for each model.
* Use techniques such as nested cross-validation to prevent overfitting during hyperparameter tuning.

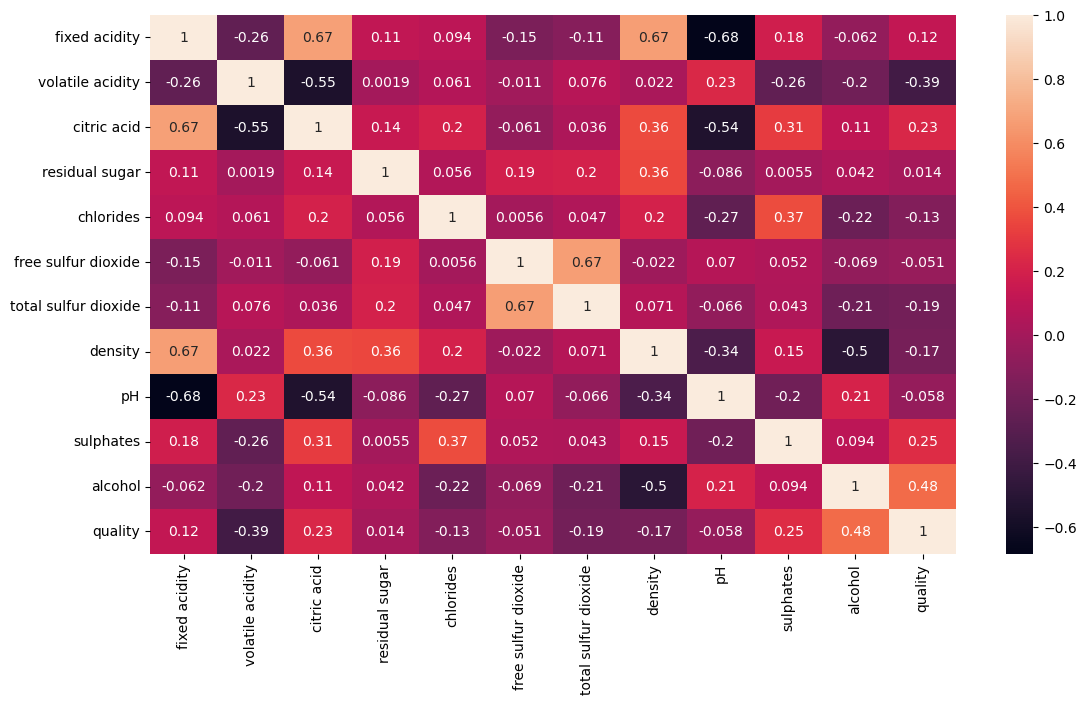
1. **Model Comparison and Selection:**

* Compare the performance of SVMs, Decision Trees, Random Forests, and Logistic Regression models based on evaluation metrics and computational efficiency.
* Select the best-performing model(s) for predicting wine quality based on overall performance and suitability for the task.

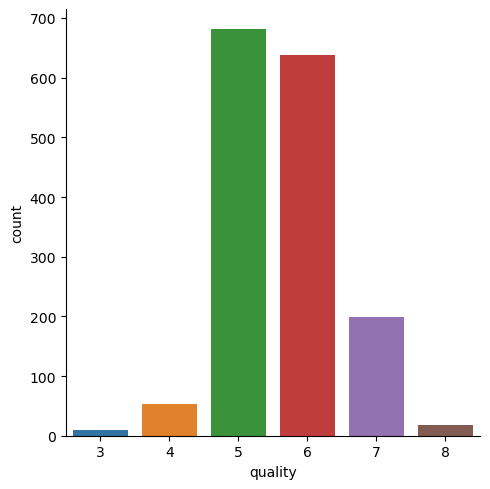
**Result**



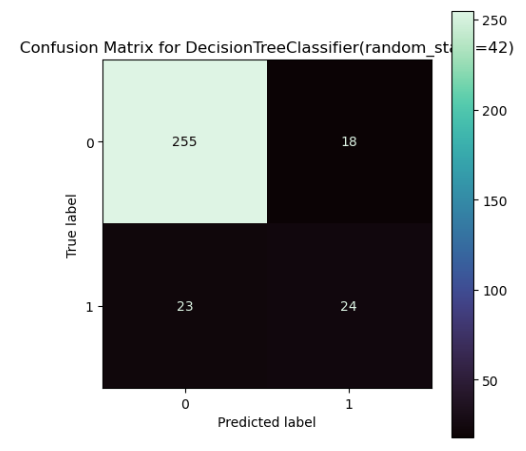
**Fig:** Probability Distribution Function



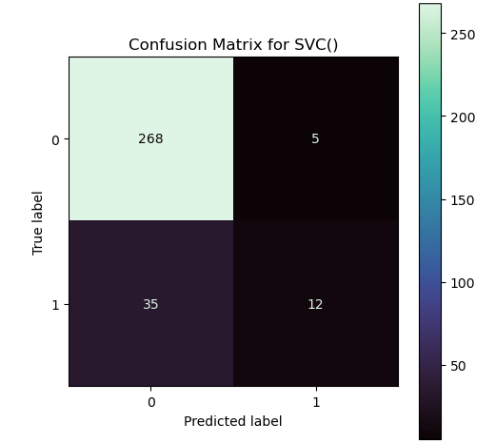
**Fig:** Correlation Matrix



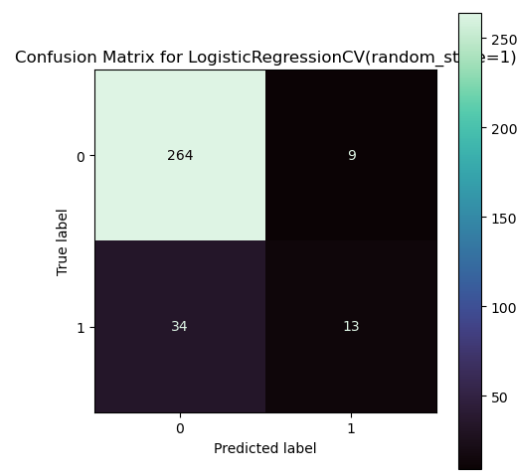
**Fig:** Visualizing the count of all 6 classes of target label



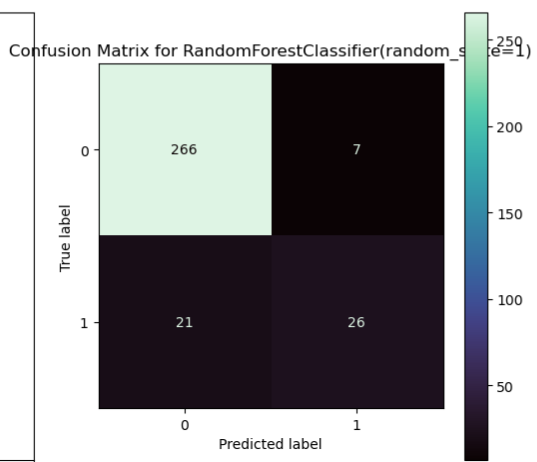
**Fig:** Confusion matrix for Decision Tree



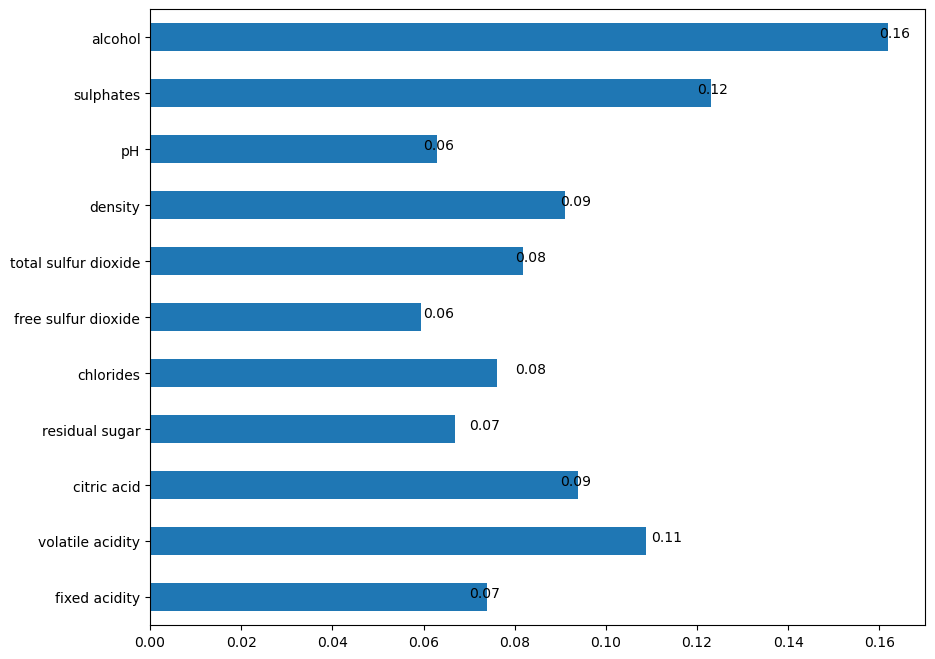
**Fig:** Confusion matrix for SVC



**Fig:** Confusion matrix for Logistic Regression



**Fig:** Confusion matrix for Random Forest



**Fig:** Selecting Best Features

**\**

**Conclusion:**

In this study, we explored the development of a wine quality prediction model utilizing machine learning techniques, including Random Forest, Support Vector Machines (SVM), and Logistic Regression. Through meticulous data preprocessing, feature engineering, and model selection, we achieved substantial insights into wine quality prediction.

* Random Forest, Support Vector Machines (SVM), and Logistic Regression were employed in developing the wine quality prediction model.
* Each algorithm contributed distinct strengths: Random Forest handled complex interactions, SVM effectively managed high-dimensional data, and Logistic Regression provided interpretable insights.
* The ensemble of these algorithms resulted in a robust predictive model with promising accuracy and reliability.
* Winemakers can leverage this model to optimize production processes and enhance wine quality, thereby improving decision-making.
* Further refinement and exploration of advanced machine learning techniques can unlock greater potential for accurate wine quality prediction.
* Integration of additional data sources and fine-tuning feature selection methodologies are avenues for future improvement.
* Ultimately, this model stands to benefit both the wine industry and consumers by optimizing processes and elevating product quality.

**References:**

* [**https://www.kaggle.com/datasets**](https://www.kaggle.com/datasets)
* [**https://www.geeksforgeeks.org/**](https://www.geeksforgeeks.org/)
* **https://stackoverflow.com/**