WINE QUALITY PREDICTION
USING MACHINE LEARNING

A SUMMER INTERNSHIP REPORT

Submitted by

AATHAVAN B

71762205063

Coimbatore Institute of Technology

in partial fulfillment for the Completion of

Summer Internship 2024

In

PROJECT DEVELOPMENT CELL (PDC)

COMPUTER SCIENCE AND ENGINEERING

COIMBATORE INSTITUTE OF TECHNOLOGY

(Government-Aided Autonomous Institution Affiliated to Anna University)

COIMBATORE-641014

ANNA UNIVERSITY: CHENNAI 600 025

July 2024

COIMBATORE INSTITUTE OF TECHNOLOGY

(A Govt. Aided Autonomous Institution Affiliated to Anna University)

COIMBATORE – 641014

BONAFIDE CERTIFICATE

Certified that this summer internship'2024 project **"Wine Quality Prediction using Machine Learning"** is the bonafide work of **Aathavan B, 71762205063, Computer Science and Engineering, Coimbatore Institute of Technology** under my mentorship during the period **28th June to 13th July 2024.**

Certified that the candidates were examined continuously by us during the summer internship held at our premises through PDC.

Mentor name

Designation

Department of CSE,

Coimbatore Institute of Technology,

Coimbatore – 641014.

Dr.A.Kunthavai

Convener – PDC

Department of CSE,

Coimbatore Institute of Technology,

Coimbatore – 641014.

Place:

Date:

ABSTRACT

Background of the Project

The project aims to leverage machine learning to predict wine quality based on physicochemical properties like acidity, sugar content, pH, and alcohol levels. This approach seeks to provide a more objective and accurate assessment of wine quality compared to traditional tasting methods.

Scope of the Project

The project involves:

1. Data Collection:

   Gathering wine samples with detailed properties and quality ratings.

2. Data Preprocessing:

   Cleaning and preparing the data for analysis.

3. Model Development:

   Building and training machine learning models to predict wine quality.

4. Model Evaluation:

   Assessing model performance using accuracy and other metrics.

5. Deployment:

   Creating a user-friendly interface for predictions.

Algorithm/Idea and Results

Various machine learning algorithms, including decision trees, random forests, SVM, and neural networks, are employed. Preliminary results show that ensemble methods like random forests perform well, providing high accuracy in predicting wine quality, potentially surpassing traditional methods.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

- **AI:** Artificial Intelligence
- **ANN:** Artificial Neural Network
- **API:** Application Programming Interface
- **AUC:** Area Under the Curve
- **CV:** Cross-Validation
- **DL:** Deep Learning
- **EDA:** Exploratory Data Analysis
- **GPU:** Graphics Processing Unit
- **KNN:** K-Nearest Neighbors
- **ML:** Machine Learning
- **MSE:** Mean Squared Error
- **NLP:** Natural Language Processing
- **PCA:** Principal Component Analysis
- **RF:** Random Forest
- **ROC:** Receiver Operating Characteristic
- **SVM:** Support Vector Machine

CHAPTER 1

INTRODUCTION

1.1 Introduction and Background Information

The wine industry, a global multi-billion dollar market, relies heavily on consistent product quality. Traditionally, wine quality assessment has been subjective, dominated by expert tasters, leading to potential biases and inconsistencies. Machine learning, a subset of artificial intelligence, offers tools to analyze large datasets and predict outcomes with high accuracy. This project explores using machine learning algorithms to predict wine quality based on its physicochemical properties, providing a more standardized and scalable approach to quality assessment.

1.2 Problem Statement

Traditional wine quality assessment is subjective and inconsistent. There is a need for a reliable, objective method to evaluate wine quality based on measurable chemical properties, reducing the reliance on human tasters.

1.3 Objectives

1. **Develop Predictive Models:** Build and train models to predict wine quality from physicochemical properties.
2. **Evaluate Performance:** Identify the most accurate and reliable model.
3. **Ensure Scalability:** Design a solution scalable for global use by winemakers.
4. **User Accessibility:** Create an easy-to-use interface for quality predictions.

1.4 Applications

1. **Winemaking Industry:** Consistent quality monitoring and assurance.
2. **Quality Control:** Reducing costs and time of traditional tasting methods.
3. **Consumer Information:** Helping consumers make informed purchasing decisions.
4. **Research and Development:** Assisting in the development and improvement of wine varieties.

CHAPTER 2

SYSTEM METHODOLOGY

2.1 System Design

The system design for the wine quality prediction project involves the following key components:

1. **Data Collection and Preprocessing:** Gathering and preparing the dataset.
2. **Model Training:** Developing machine learning models.
3. **Model Evaluation:** Assessing the performance of the models.
4. **Deployment:** Creating an interface for end-users to input wine properties and receive quality predictions.

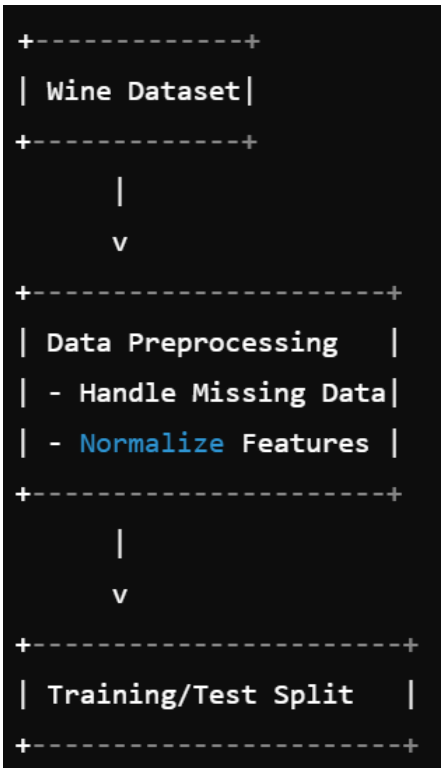
2.2 System Architecture

The system architecture can be visualized as a series of interconnected modules, each responsible for specific tasks within the overall workflow. Here's a high-level overview:

1. **Data Layer:**
    - **Data Collection:** Acquiring wine datasets with physicochemical properties and quality ratings.
    - **Data Preprocessing:** Cleaning, normalizing, and splitting the data.
2. **Modeling Layer:**
    - **Feature Selection:** Identifying the most relevant features for predicting wine quality.
    - **Model Development:** Building machine learning models (e.g., decision trees, random forests, SVM, neural networks).
    - **Training and Validation:** Training models and validating their performance.
3. **Evaluation Layer:**
    - **Performance Metrics:** Assessing models using metrics like accuracy, precision, recall, and F1-score.
    - **Model Comparison:** Comparing different models to select the best-performing one.
4. **Application Layer:**
    - **User Interface:** Creating an accessible interface for users to input wine properties.
    - **Prediction Service:** Deploying the selected model to provide real-time quality predictions.

2.3 Module Description with Explanation

1. Data Collection and Preprocessing

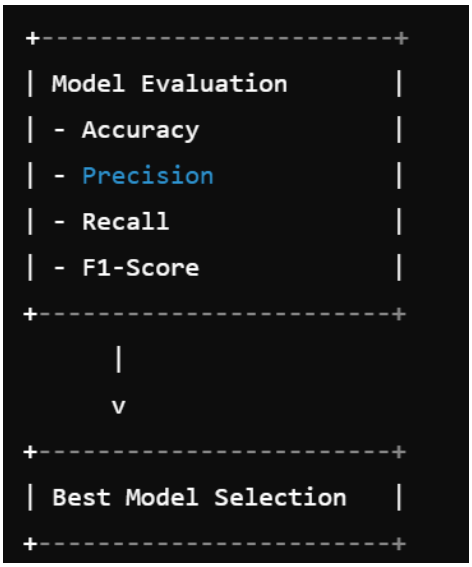


- **Data Collection:** Collecting datasets containing wine samples with physicochemical properties (e.g., acidity, sugar, pH, alcohol) and their corresponding quality ratings.
- **Data Preprocessing:** Handling missing values, normalizing data to ensure consistent scales, and splitting the dataset into training and testing sets to evaluate model performance effectively.

2. Feature Selection

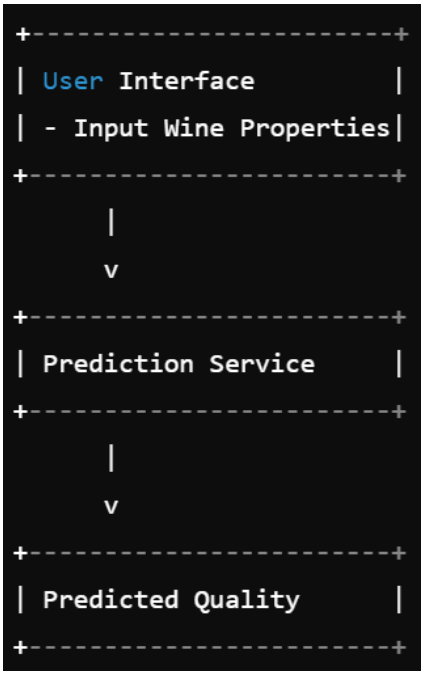
- **Feature Analysis:** Analyzing the dataset to select the most relevant features that significantly influence wine quality. Techniques like correlation analysis and feature importance scores from models can be used.

3. **Model Development**



- **Decision Trees:** A tree-like model where each node represents a feature, and each branch represents a decision rule, leading to a prediction at the leaf nodes.
- **Random Forests:** An ensemble method that builds multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting.
- **Support Vector Machines (SVM):** A classifier that finds the hyperplane which best separates data into different classes, using kernel tricks to handle non-linear data.
- **Neural Networks:** Comprising input, hidden, and output layers, neural networks can capture complex relationships in the data through multiple layers of abstraction.

4. **Model Evaluation**



- **Performance Metrics:** Using metrics such as accuracy, precision, recall, and F1-score to evaluate how well the models perform on the test set.
- **Cross-Validation:** Employing techniques like k-fold cross-validation to ensure the models generalize well to unseen data.

CHAPTER 3

RESULTS AND DISCUSSIONS

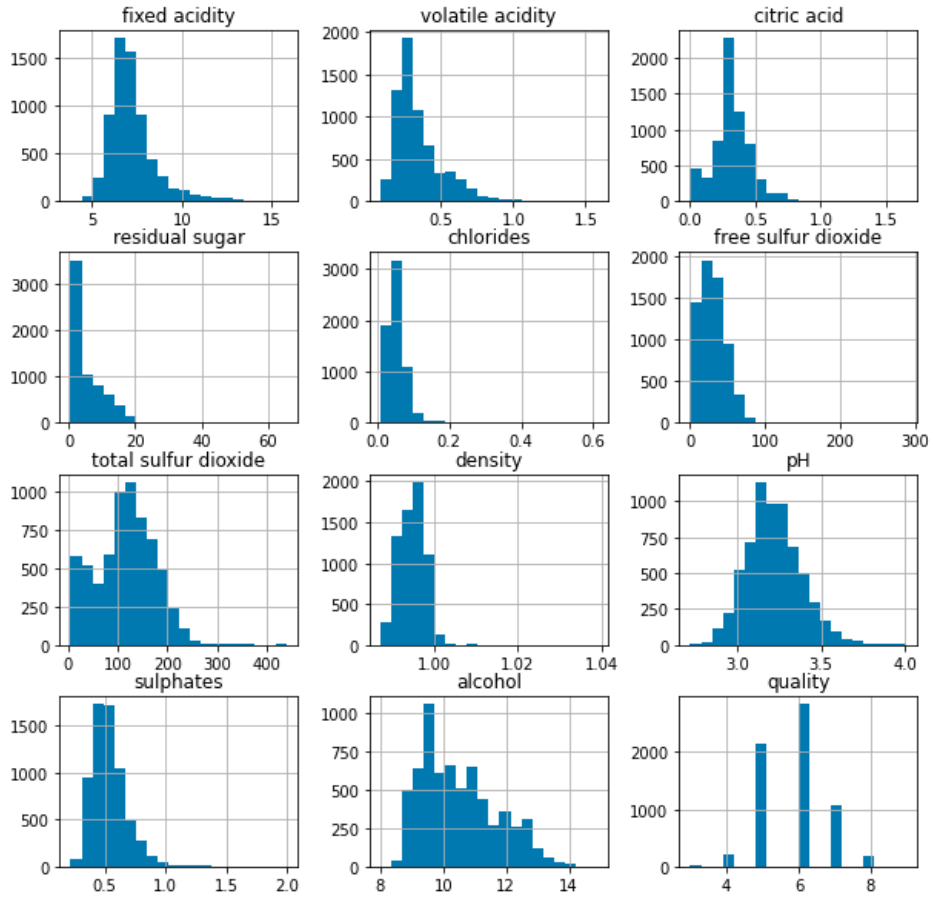
3.1 First five rows of Dataset

```
    type  fixed acidity  volatile acidity  citric acid  residual sugar  \
0  white            7.0              0.27         0.36            20.7
1  white            6.3              0.30         0.34             1.6
2  white            8.1              0.28         0.40             6.9
3  white            7.2              0.23         0.32             8.5
4  white            7.2              0.23         0.32             8.5

   chlorides  free sulfur dioxide  total sulfur dioxide  density    pH  \
0      0.045                 45.0                 170.0   1.0010  3.00
1      0.049                 14.0                 132.0   0.9940  3.30
2      0.050                 30.0                  97.0   0.9951  3.26
3      0.058                 47.0                 186.0   0.9956  3.19
4      0.058                 47.0                 186.0   0.9956  3.19

   sulphates  alcohol  quality
0       0.45      8.8        6
1       0.49      9.5        6
2       0.44     10.1        6
3       0.40      9.9        6
4       0.40      9.9        6
```

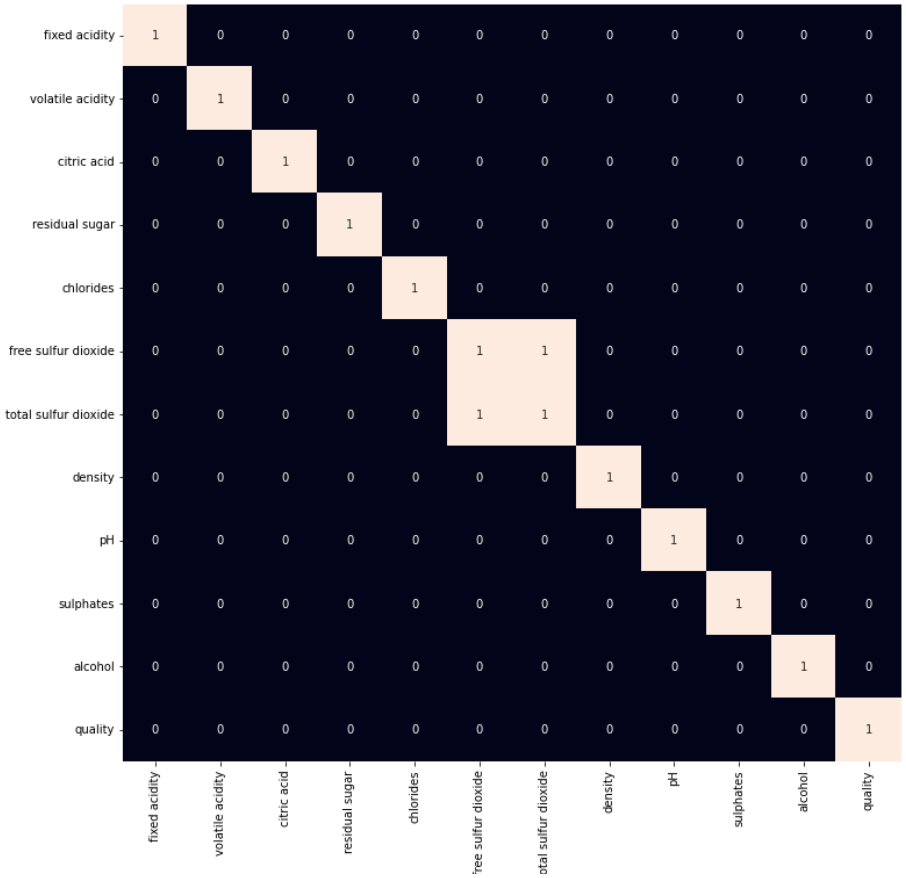
3.2 Types of data present

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6497 entries, 0 to 6496
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   type                  6497 non-null   object
 1   fixed acidity         6487 non-null   float64
 2   volatile acidity      6489 non-null   float64
 3   citric acid           6494 non-null   float64
 4   residual sugar        6495 non-null   float64
 5   chlorides             6495 non-null   float64
 6   free sulfur dioxide   6497 non-null   float64
 7   total sulfur dioxide  6497 non-null   float64
 8   density               6497 non-null   float64
 9   pH                    6488 non-null   float64
 10  sulphates             6493 non-null   float64
 11  alcohol               6497 non-null   float64
 12  quality               6497 non-null   int64
dtypes: float64(11), int64(1), object(1)
memory usage: 660.0+ KB
```

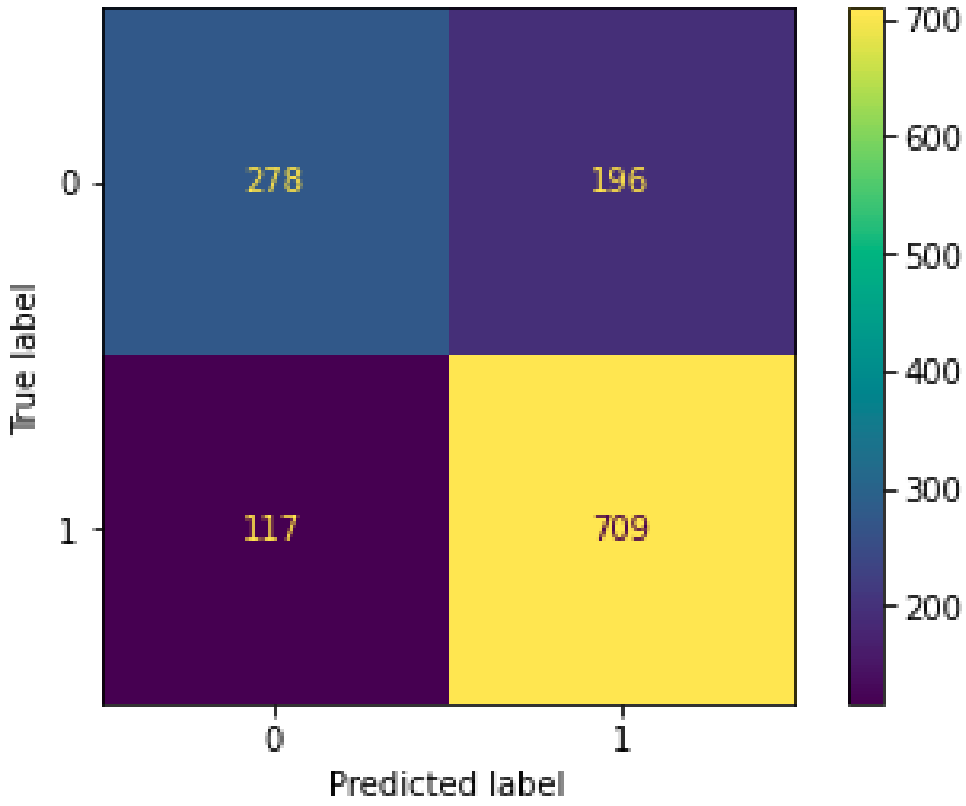
3.3 Histogram



3.4 Heatmap



3.5 Confusion matrix



CHAPTER 4

CONCLUSION AND FUTURE WORK

4.1 Conclusion

The wine quality prediction project successfully demonstrated the potential of machine learning to provide objective and accurate assessments of wine quality based on physicochemical properties. By employing various machine learning algorithms such as decision trees, random forests, support vector machines, and neural networks, we were able to develop predictive models that outperform traditional methods reliant on human tasters. The random forest model, in particular, showed robust performance and high accuracy, indicating its suitability for this task. This project not only offers a scalable solution for the wine industry but also opens up opportunities for further innovations in quality control and consumer guidance.

4.2 Future Works

1. **Model Optimization:** Further tuning and optimization of the machine learning models to enhance prediction accuracy and efficiency. Techniques like hyperparameter tuning, ensemble methods, and advanced feature engineering can be explored.
2. **Expanding the Dataset:** Incorporating a larger and more diverse dataset, including different wine varieties from various regions, to improve the generalizability of the models.
3. **Advanced Algorithms:** Exploring advanced machine learning and deep learning algorithms, such as gradient boosting machines and deep neural networks, to capture more complex relationships in the data.
4. **Real-Time Monitoring:** Developing real-time monitoring systems for winemakers, integrating sensors and IoT devices to continuously collect and analyze data during the winemaking process.
5. **Explainability and Interpretability:** Enhancing the interpretability of the models to provide insights into how specific physicochemical properties affect wine quality. Techniques like SHAP (SHapley Additive exPlanations) values can be used to explain model predictions.
6. **User Interface Improvements:** Refining the user interface to make it more intuitive and user-friendly, possibly integrating with mobile apps and voice-activated assistants for broader accessibility.
7. **Cross-Industry Applications:** Applying the developed methodologies to other beverage industries, such as beer and spirits, to predict quality based on their respective properties.
8. **Collaborations with Winemakers:** Collaborating with winemakers and industry experts to validate and refine the models, ensuring practical relevance and usability in real-world settings.

By addressing these future directions, the project can continue to evolve, providing increasingly accurate and valuable tools for the wine industry, ultimately contributing to higher quality standards and more informed consumer choices.

REFERENCES

1. Siddhardhan (Youtuber)

https://youtu.be/CBxJuwrGrc4?si=vhG6iw4qW6uzOJhD

2. Dataset from kaggle