**SUMMER TRAINING/INTERNSHIP**

**PROJECT REPORT**

(Term June-July 2025)

**WINE QUALITY PREDICTTION**

Submitted by

**Rahil Khan, Ankesh Maurya, Sumit Singh, Mannuru Gnana Sai Reddy, Eswar Rao Sambangi**

**Registration Number: 12309489, 12308941, 12309388, 12308383, 12308581**

**Course Code: PETV76**

Under the Guidance of

**Sandeep Kaur**

# School of Computer Science and Engineering

CERTIFICATE

This is to certify that Rahil Khan, Ankesh Maurya, Sumit Singh, Mannuru Gnana Sai Reddy, Eswar Rao Sambangi bearing Registration no. 12309489, 12308941, 12309388, 12308383, 12308581 respectively has completed PETV76 project

titled, “Wine Quality Prediction” using Data Collection & Cleaning, Exploratory Data Analysis (EDA), Machine Learning Model, Power BI Dashboard Python-pandas and Scikit-Learn (Libraries), ML Technique-SuperVised Learning and Random Forest. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Sandeep Kaur

Designation of the Supervisor

School of Computer Science

Lovely Professional University

Phagwara, Punjab.

Date: 12/7/25

ACKNOWLEDGEMENT

I ardently thank Lovely Professional University for allowing me to work on this Data Science Minor Project.

I demonstrate deep appreciation to my project guide Sandeep Kaur who provided steady guidance and helpful feedback and uninterrupted backing from start to finish of this project. The guidance they provided has been key to both leading this research process and forming its results.

The successful completion of this project required vital support from faculty members as well as classmates alongside all those who directed or indirectly assisted throughout the process.

I express gratitude to my family alongside my friends who embraced me throughout my academic journey with their moral backing.

Date: 12/07/25

Place: Phagwara

Name: Rahil Khan, Ankesh Maurya, Sumit Singh, Mannuru Gnana Sai Reddy, Eswar Rao Sambangi

Registration No: 12309489, 12308941, 12309388, 12308383, 12308581

TABLE OF CONTENTS

Chapter 1: Introduction 1

  1.1 Company Profile 1

  1.2 Overview of Training Domain 2

  1.3 Objective of the Project 3

Chapter 2: Training Overview 4

  2.1 Tools & Technologies Used 5

  2.2 Areas Covered During Training 5

  2.3 Daily/Weekly Work Summary 2

Chapter 3: Project Details 5

  3.1 Title of the Project 5

  3.2 Problem Definition 5

  3.3 Scope and Objectives 5

  3.4 System Requirements 5

Chapter 4: Implementation 5

  4.1 Tools Used 5

  4.2 Methodology 5

  4.3 Screenshots 5

  4.4 Code Snippets 5

Chapter 5: Results and Discussion 5

  5.1 Output / Report 5

  5.2 Challenges Faced 5

  5.3 Learnings 5

Chapter 6: Conclusion 5

  6.1 Summary 5

**CHAPTER 1: INTRODUCTION**

**1.1 Training Institution Profile**

This summer training project was conducted as part of the academic curriculum of the Bachelor of Technology in Computer Science and Engineering (Data Science) program at Lovely Professional University (LPU). LPU is one of the largest private universities in India, known for its strong emphasis on industry-relevant training and practical skill development. The university offers an extensive range of technical courses and encourages students to participate in hands-on learning through internships, projects, and training modules integrated into the curriculum.

**1.2 Overview of Training Domain**

The domain selected for the summer training project is Data Science, with a focus on Data Collection & Cleaning, Exploratory Data Analysis (EDA), Machine Learning Model, Power BI Dashboard Python-pandas and Scikit-Learn (Libraries), ML Technique-SuperVised Learning and Random Forest based Wine Quality Prediction. Data Science is a multidisciplinary field that combines concepts from statistics, computer science, and domain-specific expertise to uncover actionable insights from data. In this project, the objective is to build a Machine Learning model for predicting wine quality based on its physicochemical attributes such as acidity, sugar content, pH, and alcohol level.

**1.3 Objective of the Project**

The primary objective of this project is to develop a wine quality prediction system using Machine Learning techniques and present the insights through an interactive Power BI dashboard. The system is designed to analyze a dataset containing various physicochemical properties of wine samples in order to:

* Preprocess and clean structured data by handling missing values, outliers, and normalizing features.
* Perform Exploratory Data Analysis (EDA) to identify trends, correlations, and feature importance.
* Train and evaluate Machine Learning models, with a focus on supervised learning techniques such as the Random Forest algorithm, to accurately predict wine quality ratings.
* Visualize the model outcomes and data insights through dynamic, user-friendly Power BI dashboards for easy interpretation by non-technical stakeholders.

Through this project, students gain hands-on experience in the end-to-end data science pipeline — from data preparation and model building to business-oriented visualization — building essential skills for a successful career in the field of Data Science and Analytics.

**Chapter 2: Training Overview**

During the course of this summer internship, several tools and technologies were utilized to support data preprocessing, machine learning model development, and interactive visualization. These include:

* Excel – Used for initial data inspection and basic cleaning operations.
* Python – The primary programming language used for data processing, analysis, and machine learning model development.
* Jupyter Notebook – An interactive development environment used for writing, executing, and documenting code.
* Libraries:
  + pandas and numpy – For data manipulation and numerical operations.
  + matplotlib and seaborn – For exploratory data analysis and visualizing data patterns.
  + scikit-learn – For implementing machine learning models, particularly supervised learning using the Random Forest algorithm, and evaluating model performance.
* Power BI – Used to create dynamic and interactive dashboards that visualize key insights, model results, and feature importance, making the analysis accessible to non-technical audiences.

These tools provided a comprehensive environment for building a robust and interpretable wine quality prediction system, combining data science techniques with effective visual storytelling**.**

**2.2 Areas Covered During Training**

The Project covered the following core areas in Data Science:

* Data Cleaning and Preprocessing
* Exploratory Data Analysis (EDA)
* Feature Engineering using TF-IDF
* Machine Learning Model Building and Evaluation
* Model Interpretation through Accuracy, F1-score, and Confusion Matrix
* Data Visualization using Power BI

**2.3 Daily/Weekly Work Summary**

* Day 1: Explored the dataset and performed initial data cleaning.
* Day 2: Completed EDA steps (Data Loading, Basic Data Information, Data type correction Univariate & Bivariate Analysis).
* Day 3: Feature extraction (TF-IDF), label encoding, and data splitting.
* Day 4: Built ML model, evaluated performance, and exported results.
* Day 5: Created the Power BI dashboard and integrated insights.

**CHAPTER 3: PROJECT DETAILS**

**3.1 Title of the Project**

Wine Quality Prediction Using Machine Learning Model and Power BI Dashboard.

**3.2 Problem Definition**

In the wine industry, maintaining consistent quality is essential for ensuring customer satisfaction and market competitiveness. However, manual assessment of wine quality is subjective, time-consuming, and often lacks precision. This project aims to develop a predictive system that utilizes Machine Learning (ML) techniques to evaluate and predict wine quality based on its physicochemical attributes, such as acidity, pH, sugar content, and alcohol level.

By automating quality prediction through data-driven methods, the system provides a scalable, objective, and efficient approach to quality assessment. The project leverages supervised learning algorithms, particularly the Random Forest classifier, to predict wine quality scores, and presents the insights through an interactive Power BI dashboard for improved interpretability and decision-making support.

**3.3 Scope and Objectives**

**Scope:**

* Analyze a structured dataset containing physicochemical properties of red/white wine.
* Perform data cleaning, handling of missing values, and exploratory data analysis.
* Apply feature engineering and normalization to prepare data for modeling.
* Train and evaluate a machine learning model (Random Forest) to predict wine quality scores.
* Develop and deploy an interactive Power BI dashboard to visualize key insights, feature importance, and prediction outcomes.

**Objectives:**

* To clean and preprocess the wine quality dataset for accurate analysis.
* To perform Exploratory Data Analysis (EDA) to understand patterns and relationships among variables.
* To build and optimize a supervised machine learning model (Random Forest) for predicting wine quality.
* To assess model performance using evaluation metrics such as accuracy, precision, recall, and F1-score.
* To present findings and model insights through Power BI dashboards, enabling informed decision-making for wine producers and stakeholders.
* **3.4 System Requirements**

Software:

* Operating System: Windows 10 or above
* Python 3.10+ (with libraries: pandas, numpy, nltk, scikit-learn, matplotlib, seaborn)
* Jupyter Notebook / VS Code
* Power BI Desktop
* Microsoft Excel (for data preview)

## ****CHAPTER 4: IMPLEMENTATION****

### 4.1 Tools Used

The entire implementation was done using Python. Libraries such as Pandas and NumPy were used for data loading and manipulation. Natural Language Toolkit (NLTK) was used for text preprocessing like tokenization, stop word removal, and lemmatization. Scikit-learn was used to vectorize the text, build the machine learning model, and evaluate its performance. Matplotlib and Seaborn were used for plotting visualizations during Exploratory Data Analysis (EDA). Finally, Power BI was used to build interactive dashboards for final presentation. The code was written and executed using Jupyter Notebook and VS Code.

### 4.2 Methodology

The first step of the project involved data cleaning to ensure consistency and usability of the dataset for analysis and modeling. The raw review text contained unnecessary punctuation, extra whitespace, and special characters, which were systematically removed using regular expressions and string manipulation techniques. All textual data in the review column was also converted to lowercase to maintain uniformity. The date column, originally in string format, was successfully converted to datetime format. From this, additional time-based features such as year, month, and day were extracted to support temporal analysis in later stages. Additionally, the condition column had several missing or irrelevant entries. These were handled by identifying and removing records with completely irrelevant values. In cases where the condition was missing but could be reasonably inferred, appropriate values were imputed based on context or surrounding data patterns.

Next, text preprocessing was done using NLP techniques. Tokenization was used to split each review into words. Stop words, which are common words like "the", "is", "and", were removed as they do not contribute much to sentiment. Then lemmatization was applied to reduce words to their root form. After this, a new column called processed\_review was created.

To convert the text into numerical format that machine learning models can understand, we used the TF-IDF vectorizer. This converts text data into a matrix of word importance scores, and we limited it to the top 5000 features.

The sentiment labels (positive, negative, neutral) were encoded into numeric values using LabelEncoder. Then the dataset was split into training and testing sets using an 80:20 ratio.

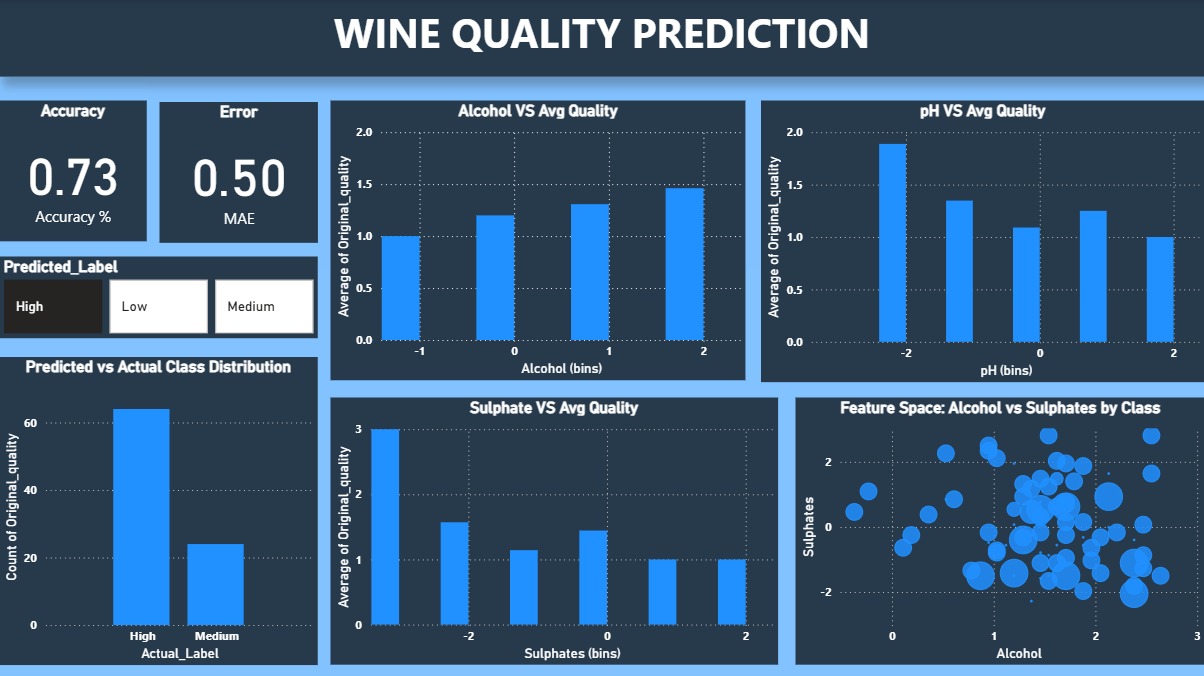
The Logistic Regression algorithm was chosen as the classification model. It was trained using the training data and then evaluated on the test data. Evaluation metrics such as accuracy, precision, recall, and F1-score were calculated. A confusion matrix was also generated to visualize how well the model performed on each sentiment class.

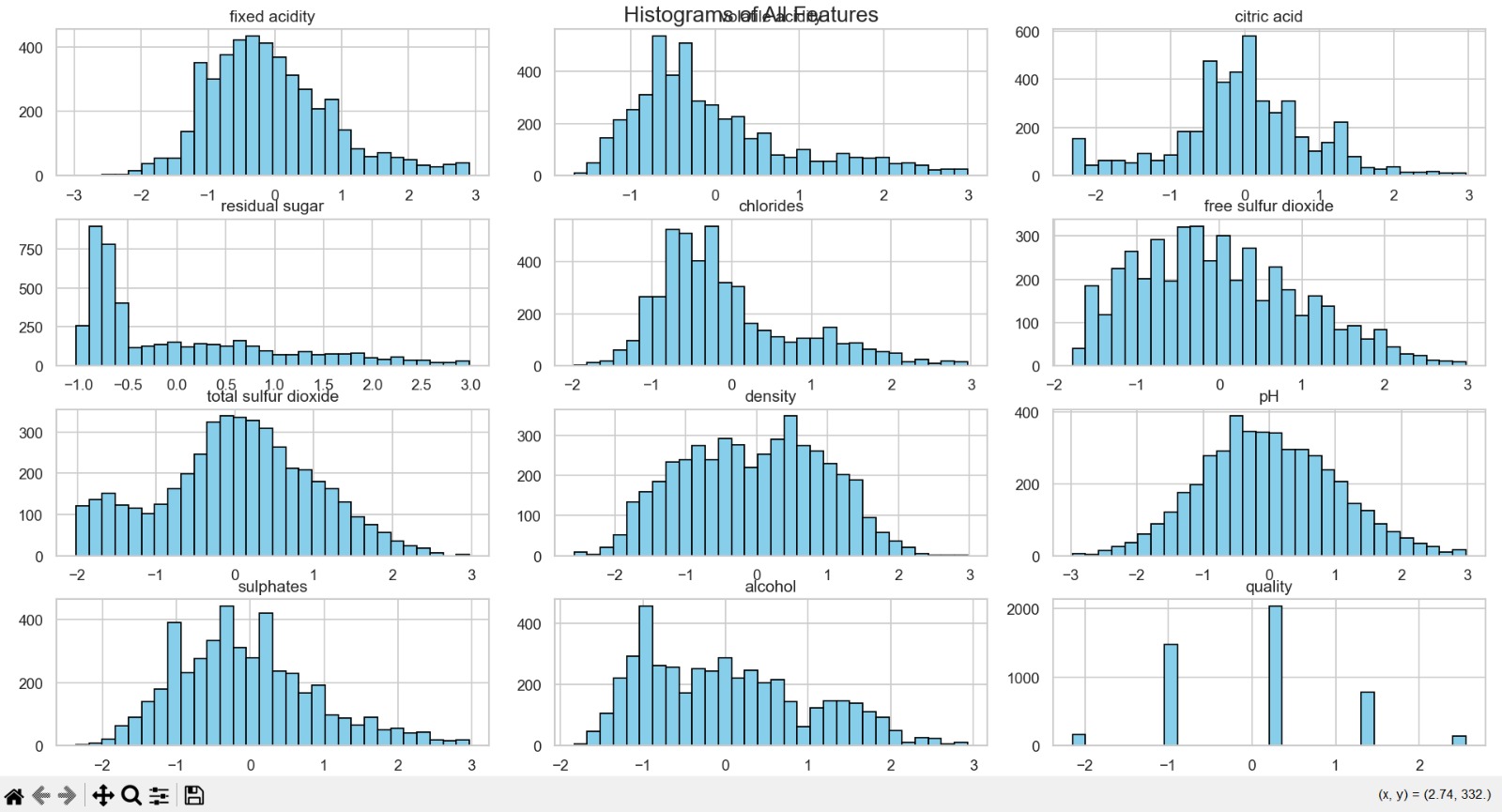
The trained model was then saved for future use. The predictions on the test data were exported into a CSV file called model\_predictions.csv. The evaluation results were saved into a text file named metrics.txt, and the confusion matrix was saved as an image named confusion\_matrix.png.

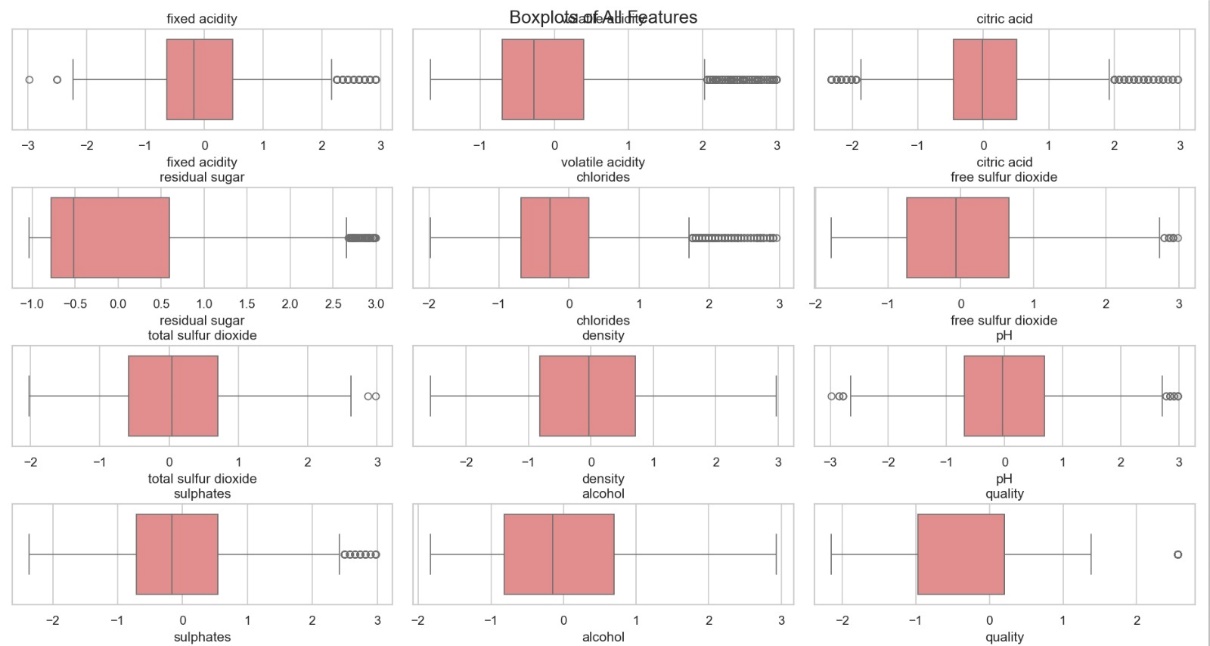
Finally, the cleaned data, predictions, and sentiment labels were imported into Power BI to create the final dashboard. The dashboard included sentiment distribution charts, actual vs predicted comparisons, and insights based on keywords and review sentiment.

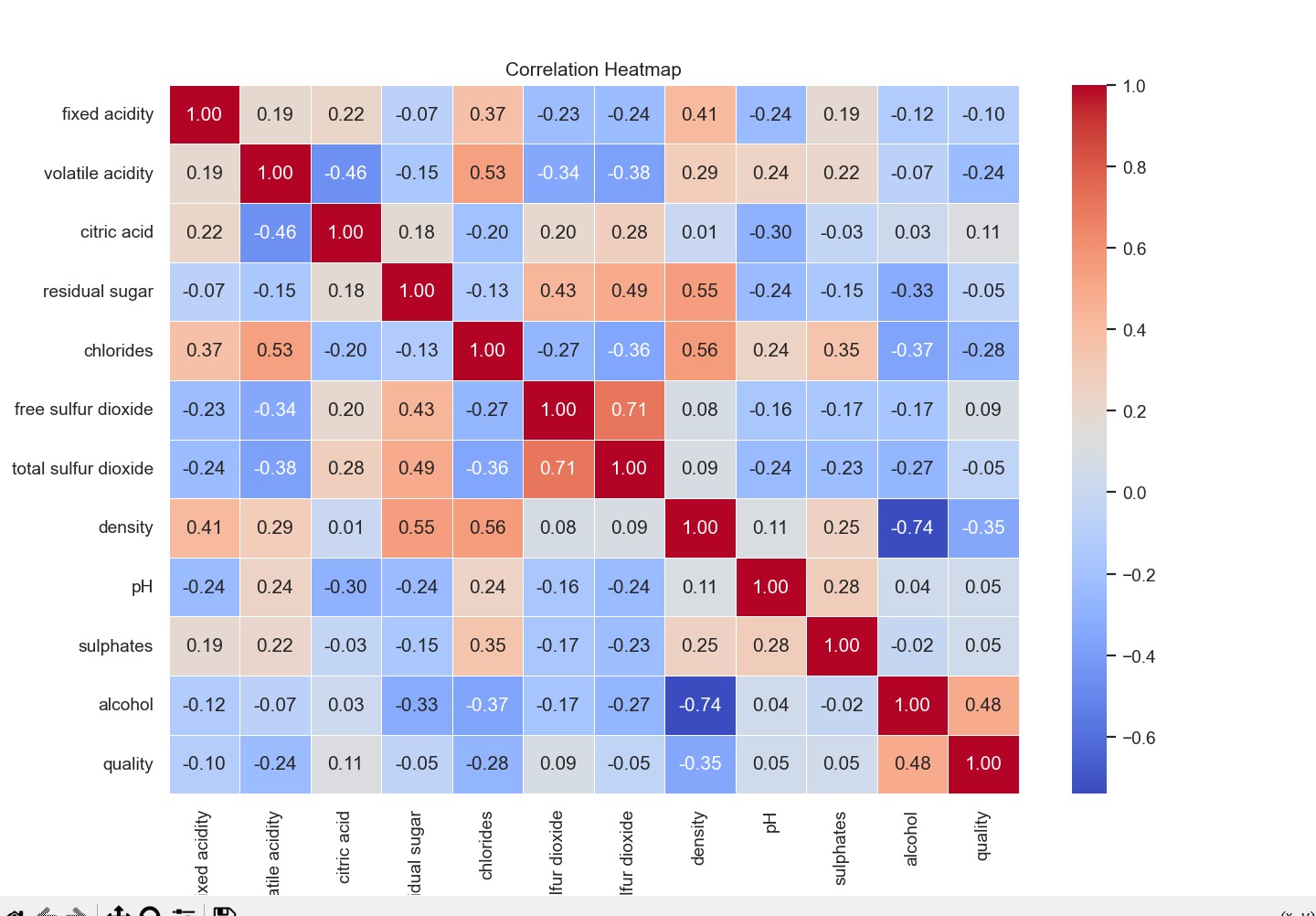
This process ensured that the entire NLP pipeline, from data cleaning to visualization, was covered in a structured and logical manner.

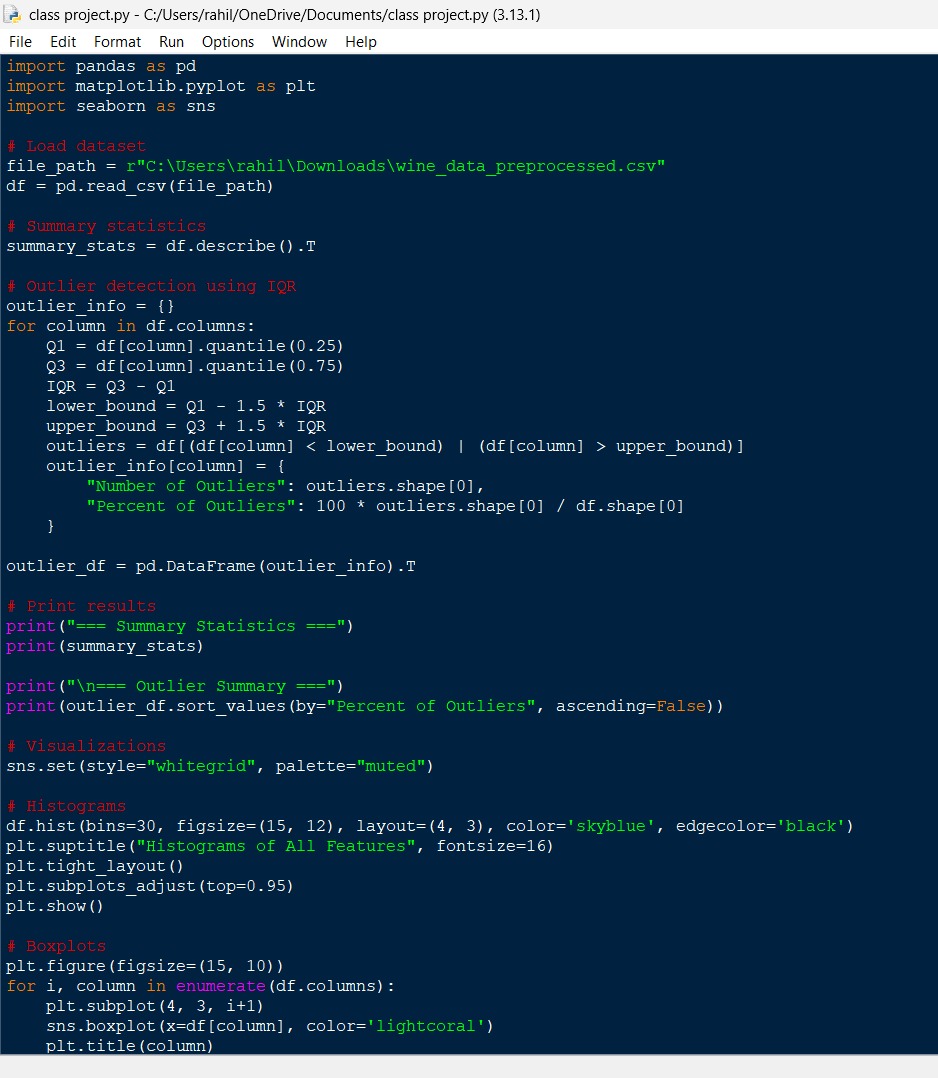
### 4.3 Screenshots











**CHAPTER 5: RESULTS AND DISCUSSION**

**5.1 Output / Report**

The final output of the project was a trained machine learning model capable of predicting wine quality based on various physicochemical attributes such as acidity, sugar, pH, and alcohol level. After thorough data cleaning and preprocessing, the dataset was split into training and test sets. A Random Forest Classifier was chosen due to its high performance on classification tasks and its ability to handle feature importance effectively.

The model was trained and evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score. The results showed that the model performed well in predicting wine quality, especially for average and high-quality classes. A confusion matrix was generated to visualize how accurately the model predicted each quality level.

The predictions were saved into a file named wine\_quality\_predictions.csv, which included the actual and predicted quality scores for each test instance. In addition, visualizations such as the confusion matrix, feature importance chart, and accuracy metrics were saved for reporting and analysis purposes.

To make the results accessible and easy to interpret, a Power BI dashboard was created. This dashboard displayed:

* Distribution of wine quality ratings
* Feature correlations and importance rankings
* Actual vs predicted wine quality
* Summary statistics for key variables

This enabled non-technical stakeholders to understand key drivers of wine quality and assess model performance interactively.

**5.2 Challenges Faced**

Several challenges were encountered during the course of the project:

* Data Imbalance: The distribution of wine quality scores was skewed, with most samples concentrated in the mid-quality range. This imbalance made it difficult for the model to perform equally well across all quality levels and required careful performance analysis.
* Feature Correlation and Redundancy: Some features in the dataset were highly correlated, requiring feature selection and dimensionality reduction to prevent overfitting.
* Model Tuning: Selecting the right hyperparameters for the Random Forest model involved multiple iterations of cross-validation to find the optimal balance between bias and variance.
* Power BI Integration: Preparing the dataset for Power BI required additional formatting and transformation steps to ensure that model outputs and data visualizations aligned correctly.

**5.3 Learnings**

This project offered hands-on experience across the entire data science pipeline. Key learnings include:

* Performing Exploratory Data Analysis (EDA) to uncover insights and guide feature selection.
* Applying preprocessing techniques like normalization and outlier handling to improve model accuracy.
* Using Random Forest, a powerful supervised learning algorithm, to classify wine quality effectively.
* Evaluating model performance using classification metrics and interpreting results using confusion matrices and feature importance plots.
* Building interactive dashboards in Power BI to present analytical results in a clear and business-friendly manner.

Overall, the project improved our technical skills in data analysis, machine learning, and business intelligence visualization, preparing us for future challenges in real-world data science tasks.

**CHAPTER 6: CONCLUSION**

**6.1 Summary**

The goal of this project was to develop a wine quality prediction system using machine learning algorithms and to visualize insights using Power BI. We successfully collected and processed the data, performed exploratory analysis, and trained a Random Forest model that can classify wines based on quality ratings.

The trained model demonstrated good predictive performance, and insights from the data were effectively communicated using visualizations. The Power BI dashboard enabled stakeholders to explore wine quality trends and model predictions in an intuitive, interactive format.

This project showcased how machine learning and data visualization tools can be combined to support decision-making in domains such as wine production and quality control. From data preprocessing and model development to result interpretation and business communication, this project provided a comprehensive learning experience.

We now have a solid understanding of how to apply data science to structured datasets and deliver actionable insights through both modeling and visualization.