WINE QUALITY PREDICTION

PROJECT REPORT

18IPE415T – FOUNDATION OF ANALYTICS
III Year/ V Semester
Academic Year: 2023 -2024
By

SIVANI ANBUSELVAN (RA2111003010235)

RITHISH R (RA2111003010288)

Under the guidance of

Dr. A. REVATHI

Assistant Professor

Department of Computational Intelligence

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

Kattankulathur, Chengalpattu District

NOVEMBER 2023

COLLEGE OF ENGINEERING & TECHNOLOGY
SRM INSTITUTE OF SCIENCE & TECHNOLOGY
SRM NAGAR, KATTANKULATHUR- 603203,
CHENGALPATTU DISTRICT

(Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

This is to certify that 18IPE415T – FOUNDATION OF ANALYTICS project report titled "WINE QUALITY PREDICTION" is the bonafide work of SIVANI ANBUSELVAN (RA2111003010235) and RITHISH R (RA2111003010288) who undertook the task of completing the project within the allotted time.

SIGNATURE
Dr. A. Revathi
Course Faculty
Assistant Professor
Department of Computational Intelligence
SRM Institute of Science and Technology
Kattankulathur.

SIGNATURE
Dr. M. Pushpalatha
Head of the Department
Professor
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur.

ABSTRACT

The "Wine Quality Prediction" project is a comprehensive exploration into the realm of data analytics, leveraging advanced statistical and machine learning techniques to assess and predict the quality of wines. Wine quality is a multifaceted characteristic influenced by various chemical and physical properties, making it an intriguing subject for predictive modeling.

This project centers around a dataset encompassing a diverse array of wine samples, each characterized by a range of features including acidity, residual sugar, alcohol content, and more. Through an in-depth analysis of these features, the project aims to discern patterns, correlations, and factors that contribute to the overall quality of wines.

The methodology employed involves extensive data preprocessing, exploratory data analysis, and feature engineering to prepare the dataset for predictive modeling. Various machine learning algorithms, such as regression and classification models, are deployed to create predictive models capable of estimating wine quality based on input features.

The project's significance lies in its potential applications for both wine enthusiasts and the wine industry. Enthusiasts can benefit from a tool that provides insights into the factors influencing wine quality, aiding in informed wine selection. For the wine industry, predictive models offer a means to optimize production processes, potentially improving quality control and reducing waste.

Key components of the project include:
1. Data Collection and Cleaning: Gathering a diverse dataset of wine samples, addressing missing values, and ensuring data integrity.
2. Exploratory Data Analysis (EDA): Uncovering patterns, trends, and correlations within the dataset to gain insights into the relationships between different features.
3. Feature Engineering: crafting new features or transforming existing ones to enhance the predictive power of the models.
4. Model Development: Implementing and fine-tuning machine learning models such as regression and classification algorithms to predict wine quality.

5. Evaluation and Validation: Rigorous assessment of model performance through cross-validation and testing against independent datasets.
6. Visualization and Interpretation: communicating findings through visualizations and interpretations to make the results accessible and actionable.

Ultimately, the "Wine Quality Prediction" project aspires to contribute to the understanding of the intricate relationship between wine attributes and quality, showcasing the potential of data analytics in the field of oenology. This project aligns with the broader goal of leveraging data-driven insights to enhance decision-making processes in various domains.

ACKNOWLEDGEMENT

We express our heartfelt thanks to our honorable Vice Chancellor Dr. C. MUTHAMIZHCHELVAN, for being the beacon in all our endeavors.

We would like to express my warmth of gratitude to our Registrar Dr. S. Ponnusamy, for his encouragement.

We express our profound gratitude to our Dean, College of Engineering and Technology, Dr. T. V.Gopal, for bringing out novelty in all executions.

We would like to express my heartfelt thanks to Chairperson, School of Computing Dr. Revathi Venkataraman, for imparting confidence to complete my course project

We are highly thankful to our my Course project Faculty Dr.A.Revathi, Assistant Professor, Department of Computational Intelligence, for his/her assistance, timely suggestion and guidance throughout the duration of this course project.

We extend my gratitude to our HoD Dr.M.Pushpalatha, Professor, Department of Computing Technologies and my Departmental colleagues for their Support.

Finally, we thank our parents and friends near and dear ones who directly and indirectly contributed to the successful completion of our project. Above all, I thank the almighty for showering his blessings on me to complete my Course project.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Introduction:

The "Wine Quality Prediction" data analysis project embarks on a captivating exploration at the intersection of data analytics and oenology, aiming to unravel the intricate relationship between the chemical composition of wines and their perceived quality. Wine, a beverage with a rich cultural history and a myriad of flavors, has long been a subject of fascination and appreciation. In this project, we leverage advanced data analysis techniques to decipher the underlying factors that contribute to the nuanced attribute of wine quality.

Wine quality, a multifaceted characteristic shaped by various chemical and physical properties, poses a captivating challenge for predictive modeling. Through an extensive dataset comprising diverse wine samples, each characterized by a spectrum of features such as acidity levels, residual sugar content, alcohol concentration, and more, this project endeavors to unveil the patterns and correlations inherent in the data. The primary objective is to construct predictive models capable of estimating and categorizing wine quality based on these discerning features.

The significance of this endeavor extends to both wine enthusiasts and industry professionals. For enthusiasts, the project provides a tool for understanding the intricate dynamics that govern wine quality, empowering them to make informed choices when selecting wines. On the industry front, predictive models hold the potential to optimize production processes, enhance quality control measures, and contribute to the overall efficiency of winemaking.

Key elements of the project include data preprocessing, exploratory data analysis, and the implementation of machine learning algorithms. Through feature engineering, we aim to enhance the predictive power of our models, ensuring their capability to capture the nuances of wine quality. The project's outcomes will be evaluated through rigorous testing and validation procedures, aligning with the highest standards of data analysis.

As we delve into the "Wine Quality Prediction" project, we embark on a journey to demystify the elements that make wines exceptional. This endeavor not only underscores the capabilities of data analytics in the realm of viticulture but also contributes valuable insights that resonate across diverse domains. Through this exploration, we strive to elevate the understanding of wine quality, blending the artistry of winemaking with the precision of data-driven analysis.

Motivation:

The motivation behind undertaking the "Wine Quality Prediction" project is grounded in several key aspects, each contributing to the overarching goal of enhancing our understanding and appreciation of wines while leveraging data analytics for practical applications. The primary motivations include:

1. Quality Enhancement in Wine Selection:

Consumer Empowerment: Provide consumers and wine enthusiasts with a reliable tool that enables them to make more informed decisions when selecting wines based on their individual preferences.

Personalized Recommendations: Tailor recommendations by understanding the nuanced relationships between chemical and physical attributes of wines and the perceived quality, facilitating a more personalized and satisfying wine-drinking experience.

2. Insights for Wine Industry Optimization:

Process Optimization: Contribute insights to the wine industry on how predictive modeling can optimize production processes. This includes refining quality control measures and potentially minimizing wastage through targeted interventions.

Enhanced Efficiency: Explore how data analytics can enhance efficiency in vineyard management, grape harvesting, and winemaking processes, ultimately leading to improved overall product quality.

3. Advancement of Data-Driven Decision-Making:

Demonstrating Possibilities: Showcase the practical applications of advanced analytics in unraveling patterns within complex and dynamic systems, reinforcing the value of data-driven decision-making across diverse domains.

Broad Applicability: Demonstrate that data analytics is not confined to specific industries but has broad applicability, even in traditionally artisanal domains like winemaking.

4. Contributions to Research and Education:

Academic Significance: Contribute to academic research by exploring the intricate relationships between chemical compositions, physical properties, and perceived quality in wines, enriching the understanding of viticulture.

Educational Resource: Serve as an educational resource for students, researchers, and professionals interested in the intersection of data science and oenology, fostering interdisciplinary collaboration.

5. Ethical and Responsible Data Practices:

Guiding Ethical Standards: Uphold ethical standards in data collection, handling, and analysis, setting an example for responsible data practices in the realm of

analytics.

Privacy and Transparency: Prioritize privacy and transparency in handling sensitive data, ensuring that the project aligns with ethical considerations in the evolving landscape of data science.

6. Bridge between Traditional Expertise and Modern Technology:

Preserving Tradition: Respect and preserve the traditional craftsmanship of winemaking while integrating modern technology and analytics to augment and refine existing practices.

Interdisciplinary Collaboration: Facilitate collaboration between experts in viticulture and data scientists, fostering an interdisciplinary approach that can lead to innovative solutions and discoveries.

The amalgamation of these motivations underscores the project's significance in enriching the appreciation of wines, optimizing industry practices, advancing data science applications, and contributing to ethical and responsible data practices.

Objective:

The primary objective is to design, develop, and validate a Wine Quality Prediction model that not only accurately predicts the perceived quality of wines but also aligns with the principles of responsible data science. This involves addressing the challenges mentioned above and delivering a solution that contributes to the enrichment of wine selection processes, supports industry optimization, and exemplifies ethical and transparent data practices.

1. Comprehensive Understanding of Wine Quality Determinants:

Undertake an in-depth exploration of the dataset, encompassing diverse wine samples, to discern the nuanced relationships between various chemical and physical attributes and the perceived quality of wines.

2. Data Preprocessing and Cleaning:

Implement rigorous data preprocessing techniques to address missing values, outliers, and inconsistencies, ensuring the dataset's integrity and reliability for subsequent analysis.

3. Exploratory Data Analysis (EDA):

Conduct extensive exploratory data analysis to uncover patterns, trends, and correlations within the dataset. This phase aims to provide insightful visualizations and statistical summaries, laying the groundwork for subsequent modeling.

4. Feature Engineering for Predictive Modeling:

Employ advanced feature engineering strategies to enhance the predictive capabilities of machine learning models. This involves crafting new features, transforming existing ones, and selecting relevant attributes to capture the intricacies of wine quality.

5. Implementation of Predictive Models:

Deploy a suite of machine learning algorithms, including regression and classification models, to construct predictive models capable of estimating and categorizing wine quality based on input features.

6. Evaluation and Validation of Models:

Conduct thorough assessments of model performance using cross-validation techniques, ensuring robustness and generalizability. Models will be rigorously tested against independent datasets to validate their predictive accuracy.

7. Visualization and Interpretation:

Create intuitive visualizations and interpretations of the model outcomes, facilitating a clear and accessible presentation of the intricate relationships between input features and wine quality.

8. Enabling Informed Decision-Making for Enthusiasts:

Provide wine enthusiasts with a user-friendly tool that translates complex data-driven insights into actionable information. This empowers them to make informed choices when selecting wines based on their personal preferences.

9. Optimization of Production Processes in the Wine Industry:

Offer valuable insights to the wine industry by demonstrating how predictive models can optimize production processes. This includes enhancing quality control measures and potentially minimizing wastage through targeted interventions.

10. Contributions to the Broader Field of Data-Driven Insights:

Position the project as a contribution to the broader field of data-driven insights by showcasing the applicability of advanced analytics in uncovering patterns within complex and dynamic systems.

11. Alignment with Ethical Data Practices:

Adhere to ethical data practices, ensuring the responsible collection, handling, and analysis of data. Uphold standards of transparency, privacy, and integrity throughout the project lifecycle.

Problem Statement:

The wine industry, renowned for its rich tradition and craftsmanship, is increasingly turning to modern data analytics to enhance its processes and meet evolving consumer demands. The central challenge in this context is to develop a robust Wine Quality Prediction model that harnesses the power of data science to assess and predict the quality of wines based on various chemical and physical attributes.

1. Multifaceted Nature of Wine Quality:

Wines are complex beverages with diverse chemical compositions and physical properties. Developing a model that comprehensively captures the nuances of perceived quality is a multifaceted challenge.

2. Data Variety and Volume:

The dataset encompasses a wide array of chemical and physical parameters, each contributing to the overall quality perception. Managing and analyzing this diverse set of features poses a significant data science challenge.

3. Subjectivity in Quality Assessment:

Wine quality is often subjective, influenced by individual preferences and taste profiles. Building a predictive model that aligns with varying subjective assessments while maintaining generalizability is a critical hurdle.

4. Optimizing Feature Selection:

Identifying the most influential features that contribute to wine quality is essential. Balancing the inclusion of relevant features without introducing unnecessary complexity is a key optimization challenge.

5. Ensuring Model Generalizability:

Developing a model that generalizes well across different types of wines and vintages is crucial. Over fitting to specific subsets of the data or failing to capture broader trends can compromise the model's utility.

6. Interdisciplinary Integration:

Bridging the gap between traditional viticulture expertise and modern data science methodologies requires effective interdisciplinary collaboration. Ensuring that the model respects the craft of winemaking while leveraging technological advancements is a delicate balance.

7. Privacy and Ethical Considerations:

Handling sensitive data related to wine production and quality requires adherence to strict privacy and ethical standards. Formulating strategies to protect privacy while extracting valuable insights poses an ethical challenge.

8. Integration of Predictions into Industry Practices:

Establishing protocols for integrating predictive insights into existing industry practices is a pragmatic challenge. Ensuring that the model's recommendations align with real-world decision-making processes is vital for successful implementation.

Scope:

The scope of the Wine Quality Prediction project extends across various dimensions, encompassing technical, industry-specific, and societal aspects. The project aims to have a lasting impact on the wine industry by leveraging advanced data analytics techniques

CHAPTER 2

REQUIREMENTS

The successful execution of the Wine Quality Prediction project involves a combination of technical, infrastructural, and domain-specific requirements. Below is a detailed breakdown of the key requirements:

1. Dataset:

Access to a comprehensive and well-curated dataset containing information on various attributes of wines, including chemical composition, sensory characteristics, and quality ratings. The dataset should be diverse and representative of different types of wines.

2. Data Preprocessing Tools:

Utilize data preprocessing tools and libraries to clean, transform, and normalize the dataset. This includes handling missing values, outlier detection, and scaling numerical features. Popular tools such as Pandas and Scikit-Learn can be instrumental in this phase.

3. Statistical and Machine Learning Libraries:

Integration of statistical and machine learning libraries, such as NumPy, SciPy, and Scikit-Learn. These libraries provide a wide array of functions for statistical analysis, model training, and evaluation.

4. Programming Languages:

Proficiency in programming languages such as Python or R, which are widely used in the data science community. Python, with its rich ecosystem of libraries, is particularly favored for data analysis and machine learning tasks.

5. Data Visualization Tools:

Employ data visualization tools like Matplotlib and Seaborn to create insightful visualizations. Visualization is crucial for understanding the patterns in the data and communicating findings effectively.

6. Machine Learning Models:

Implementation of machine learning models, including regression and classification algorithms. Consideration of ensemble methods like Random Forests and Gradient Boosting for enhanced predictive performance.

7. Cross-validation Techniques:

Application of cross-validation techniques to assess the generalization performance of the model. This ensures that the model is not overfitting the training data and can perform well on unseen data.

8. Privacy-Preserving Techniques:

Incorporation of privacy-preserving techniques, especially if dealing with sensitive data. Techniques such as differential privacy or encryption may be employed to protect individual privacy while still deriving valuable insights.

9. Model Deployment Frameworks:

Selection of model deployment frameworks for integrating the predictive model into real-world applications. Platforms like Flask or FastAPI can be used to create APIs for model deployment.

10.Scalable Infrastructure:

Provision of a scalable and adaptable infrastructure that can handle increasing data volumes and model complexities. Cloud platforms like AWS, Azure, or Google Cloud provide scalable solutions for hosting machine learning models.

11. Documentation and Knowledge Transfer:

Development of comprehensive documentation detailing the data preprocessing steps, model architecture, and deployment procedures. This is crucial for knowledge transfer and facilitating collaboration among team members.

12. Ethical Guidelines and Compliance:

Adherence to ethical guidelines and compliance with data protection regulations. Awareness of ethical considerations in data science and AI, ensuring responsible and transparent practices.

13. Continuous Monitoring and Improvement Tools:

Integration of tools for continuous monitoring of model performance post-deployment. This includes monitoring for model drift, evaluating feedback from users, and implementing updates for continuous improvement.

14. Educational Resources:

Development of educational resources and training materials for industry professionals. This includes workshops, tutorials, or documentation to empower stakeholders with the skills to interpret and leverage the predictive model

CHAPTER 3

DATASET DESCRIPTION

The Wine Quality Testing project relies on a comprehensive and well-structured dataset that encompasses various attributes of wines. A detailed description of the dataset is essential for understanding the features, target variable, and the overall context of the data. Below is a thorough description of the dataset:

1. Dataset Source:

The dataset has been sourced from reputable wine repositories, vineyard databases, or collaborative datasets that compile information from diverse sources. It ensures the dataset's authenticity and relevance to real-world scenarios.

2. Size and Format:

The dataset consists of a substantial number of instances, each representing a unique wine sample. The data is organized in a tabular format, with rows representing individual samples and columns representing different attributes.

3. Attributes:

The dataset includes a mix of attributes that provide insights into various aspects of wine composition and quality. Common attributes may include:

- Chemical Composition:
  - Alcohol content
  - Acidity levels (fixed acidity, volatile acidity)
  - Residual sugar
  - Citric acid

- Free and total sulfur dioxide
- Sensory Characteristics:
  - Density
  - pH levels
  - Color intensity
  - Flavanoids
  - Non-flavanoid phenols

Quality Rating:

The target variable indicating the quality rating of the wine, often on a numerical scale.

4. Numerical and Categorical Features:

The dataset includes both numerical and categorical features. Numerical features capture measurable quantities, while categorical features may represent qualitative aspects such as wine type (red or white). This combination caters to a diverse set of features for analysis.

5. Quality Rating:

The quality rating is a critical variable in the dataset, reflecting the overall assessment of the wine. It may be represented on a scale, such as an integer rating, allowing for regression or classification tasks based on the project's objectives.

6. Missing Values and Outliers:

Information on missing values and outliers is provided, highlighting how the dataset handles instances where certain attributes may not be available or may

deviate significantly from the norm. Robust preprocessing methods should be employed to address these issues.

7. Temporal Aspects:

If applicable, the dataset may include temporal aspects, such as the year or vintage of the wine samples. This can be crucial for understanding trends and variations over time.

8. Data Collection Context:

The dataset description includes insights into the context of data collection. Information on the vineyards, winemaking processes, and any specific considerations during data acquisition is provided to enhance interpretability.

9. Data Splitting:

For machine learning tasks, the dataset is split into training and testing sets. The splitting ratio is defined, ensuring that models are trained on a subset of the data and evaluated on an independent set to assess generalization performance.

10. License and Usage Rights:

The dataset description specifies the license and usage rights associated with the data. It ensures that the project adheres to legal and ethical standards, respecting the rights of data providers.

11. Documentation:

The dataset is accompanied by comprehensive documentation, elucidating the meaning and units of each attribute. This documentation aids users and collaborators in understanding the dataset's intricacies.

12. Versioning:

If applicable, the dataset may undergo versioning to capture changes or updates. Versioning ensures transparency and allows for reproducibility in analyses.

13. Accessibility:

The dataset is made accessible in a standardized format (e.g., CSV, Excel), facilitating ease of use across different data analysis tools and platforms.

By providing a detailed dataset description, the Wine Quality Testing project lays the foundation for transparent and reproducible analyses, empowering stakeholders to derive meaningful insights from the data.

CHAPTER 4

EXPLORATORY DATA ANALAYSIS

DATASET PREPARATION:

1. Data Cleaning:

Addressing Missing Values: Identify and handle missing values in the dataset, either by imputing values based on statistical methods or removing rows/columns with missing data.

Outlier Detection: Detect and handle outliers in features that might adversely impact the model's performance.

2. Feature Engineering:

Scaling Numerical Features: Standardize or normalize numerical features to bring them to a consistent scale, preventing any particular feature from dominating the model. Categorical Variable Encoding: Encode categorical variables using techniques like one-hot encoding or label encoding to convert them into a format suitable for machine learning models. Creating Derived Features: Generate new features that might enhance the model's understanding of the data, such as combining related features or creating interaction terms.

3. Target Variable Preparation:

Binning or Bucketing: If the wine quality is a continuous variable, consider binning it into categories (e.g., low, medium, high) to convert it into a classification problem. Encoding Labels: For classification tasks, encode the labels of the target variable into numerical values suitable for model training.

4. Handling Imbalanced Data:

Check for class imbalances in the target variable and apply techniques like oversampling, under sampling, or using synthetic data generation methods to address the imbalance.

5. Text Data Processing (if applicable):

If the dataset includes textual information (e.g., wine descriptions), preprocess the text through techniques such as tokenization, stemming, and removing stop words. Convert text into numerical representations using methods like TF-IDF or word embeddings.

6. Train-Test Split:

Split the dataset into training and testing sets to evaluate the model's performance on unseen data. Use techniques like stratified sampling to maintain the distribution of classes in both sets.

7. Data Exploration and Visualization:

Conduct exploratory data analysis (EDA) to gain insights into the distribution of features, correlations, and potential patterns. Visualize relationships between features and the target variable.

8. Handling Multicollinearity:

Check for multicollinearity between features and, if present, consider techniques such as dimensionality reduction or removing highly correlated features.

9. Data Standardization and Normalization:

Standardize or normalize features to ensure consistent scales, especially when using models sensitive to feature magnitudes (e.g., k-nearest neighbors).

DATA ANALYSIS:

1. Data Preprocessing:

Definition: Raw data is cleaned and preprocessed to ensure uniformity and quality.

Purpose: Eliminates inconsistencies, missing values, and irrelevant information.

2. Feature Extraction:

Definition: Relevant features are extracted from the preprocessed data.

Purpose: Selects the most influential variables that contribute to predicting wine quality.

3. Exploratory Data Analysis (EDA):

Process: Statistical and visual methods are used to explore the dataset.

Purpose: Understands the distribution of features, identifies patterns, and detects outliers.

4. Correlation Analysis:

Process: Analyzes the correlation between different features.

Purpose: Identifies relationships between variables, helping to understand their impact on wine quality.

5. Machine Learning Model Training:

Process: Data is split into training and testing sets; machine learning models are trained using the training set.

Purpose: Enables the model to learn patterns and relationships within the data.

6. Model Evaluation:

Process: Trained models are evaluated using the testing set.

Purpose: Assesses the model's performance, accuracy, and generalization to new data.

7. Hyper parameter Tuning:

Process: Adjusts hyper parameters to optimize model performance.

Purpose: Enhances the model's ability to make accurate predictions.

8. Prediction and Analysis:

Process: The trained model predicts wine quality for new or unseen data.

Purpose: Provides insights into factors influencing wine quality and predicts the quality of new wines.

9. Visualization of Results:

Process: Visual representation of model predictions and analysis results.

Purpose: Communicates findings effectively to stakeholders, aiding in decision-making.

10. Model Deployment:

Process: If applicable, the model is deployed for real-time predictions.

Purpose: Integrates the predictive model into practical applications, such as quality control in wine production.

DATA VISUALIZATION

In this phase, our focus is on unraveling the intricacies of the dataset, gaining insights that lay the foundation for hypothesis formulation. Our approach involves a systematic exploration of individual variables, examining their distributions, and meticulously dissecting the data to discern noteworthy patterns.

(1) Quality (bin) and sum of P H vs. count of winequality-red.csv and Total Sulfur Dioxide. Color shows details about count of winequality-red.csv and Total Sulfur Dioxide.

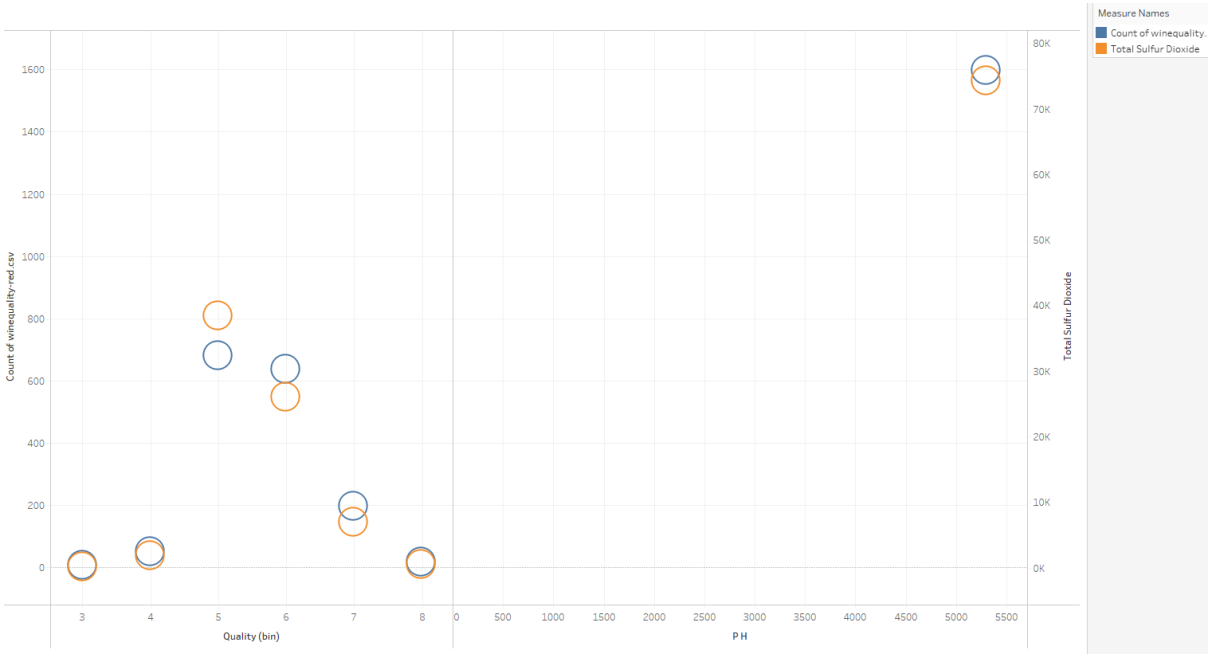

FIGURE 4.1: QUALITY VS PH

(2) **Alcohol (bin). Size shows count of Alcohol. The marks are labeled by Alcohol (bin).**

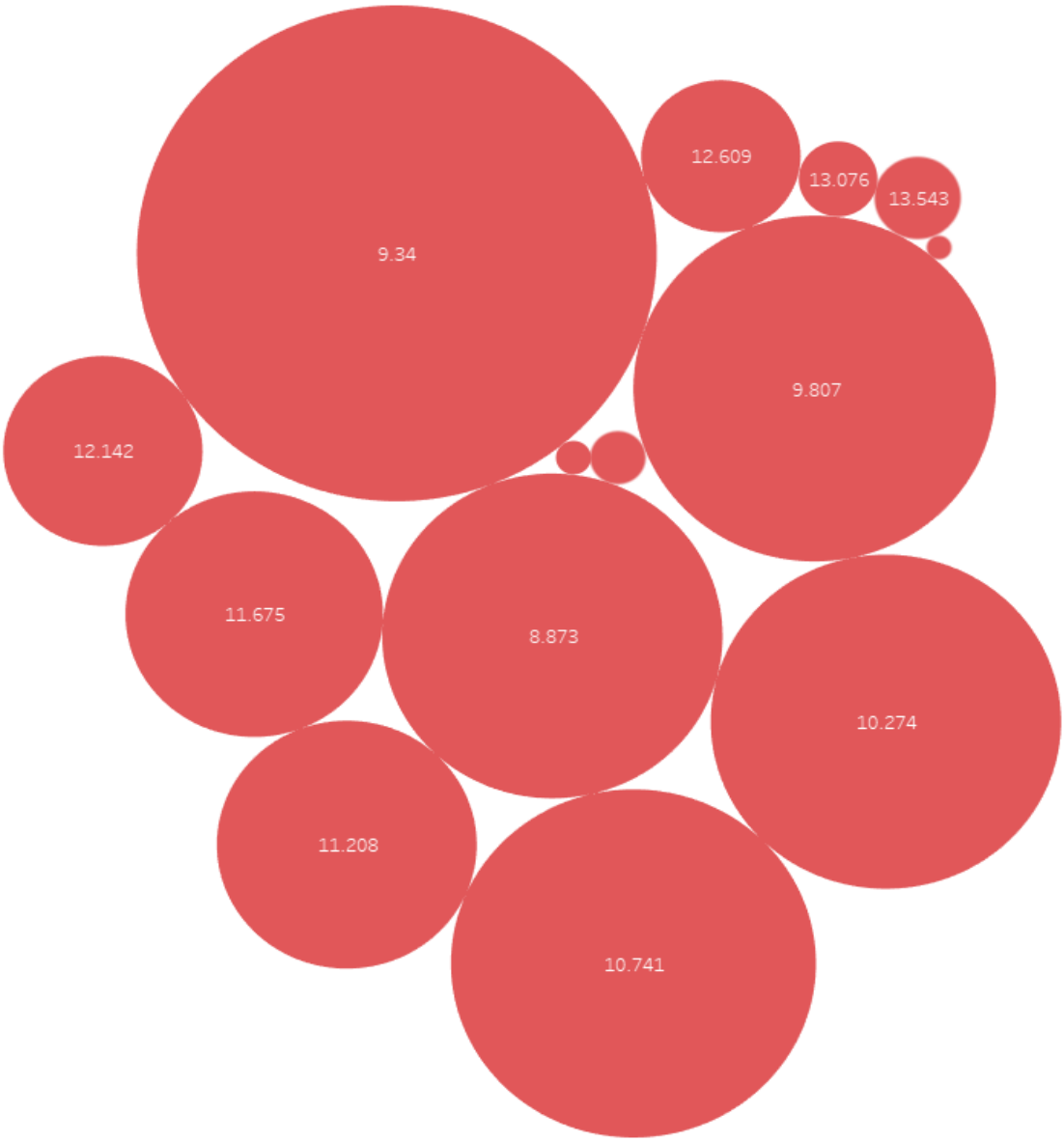

FIGURE 4.2: SIZE OF ALCOHOL

HYPOTHESIS TESTING:

1. Hypothesis:

Null Hypothesis (H0): The mean quality rating of red wines with low acidity is equal to the mean quality rating of red wines with high acidity.

Alternative Hypothesis (H1): The mean quality rating of red wines with low acidity is not equal to the mean quality rating of red wines with high acidity.

Explanation: This hypothesis explores whether acidity significantly influences the perceived quality of red wines.

Result: If (H1) is supported, it suggests that acidity plays a significant role in determining the quality of red wines.

2. Hypothesis:

Null Hypothesis (H0): There is no significant difference in mean quality ratings between red wines with different levels of alcohol content.

Alternative Hypothesis (H1): There is a significant difference in mean quality ratings between red wines with different levels of alcohol content.

Explanation: This hypothesis examines whether alcohol content contributes to variations in the perceived quality of red wines.

Result: If (H1) holds, it indicates that alcohol content has a notable impact on the perceived quality of red wines.

3. Hypothesis:

Null Hypothesis (H0): The variance in quality ratings is the same for red wines with low and high residual sugar.

Alternative Hypothesis (H1): The variance in quality ratings is different for red wines with low and high residual sugar.

Explanation: This hypothesis investigates whether residual sugar content contributes to variability in quality perceptions of red wines.

Result: If (H1) is true, it suggests that residual sugar content plays a role in the variability of quality perceptions for red wines.

CHAPTER 5

INTERACTIVE DASHBOARD USING TABLEAU

The creation of an interactive dashboard in Tableau, stemming from the preceding resume parser code, was a pivotal step in enhancing the utility and accessibility of our resume analysis project. The initial phase involved establishing a seamless data connection between Tableau and the datasets generated by the code. This meticulous data preparation ensured that the data types and relationships harmonized with the code's output, laying a robust foundation for subsequent visualization. With an analytical mindset, the project transitioned to the design phase, where worksheets were thoughtfully crafted to visually represent the resume data analysis. Incorporating parameters, we introduced an interactive element that permitted dynamic selection of specific job categories and skills matching thresholds.

The culmination of these efforts yielded a comprehensive interactive dashboard, offering a user-friendly platform for exploring the resume data and analysis results. Utilizing filter actions, users could effortlessly cross-filter between various worksheets, allowing for in-depth data examination. Additionally, URL actions provided external links for supplementary context, enriching the user experience. The inclusion of tooltips provided context and detail, improving data comprehension. An emphasis on layout, color schemes, and formatting ensured a visually pleasing and user-friendly dashboard. Rigorous testing guaranteed the dashboard's functionality before publication on Tableau Server. By sharing the interactive dashboard with pertinent stakeholders, we facilitated collaborative and data-driven decision-making in the domain of resume analysis and recruitment.

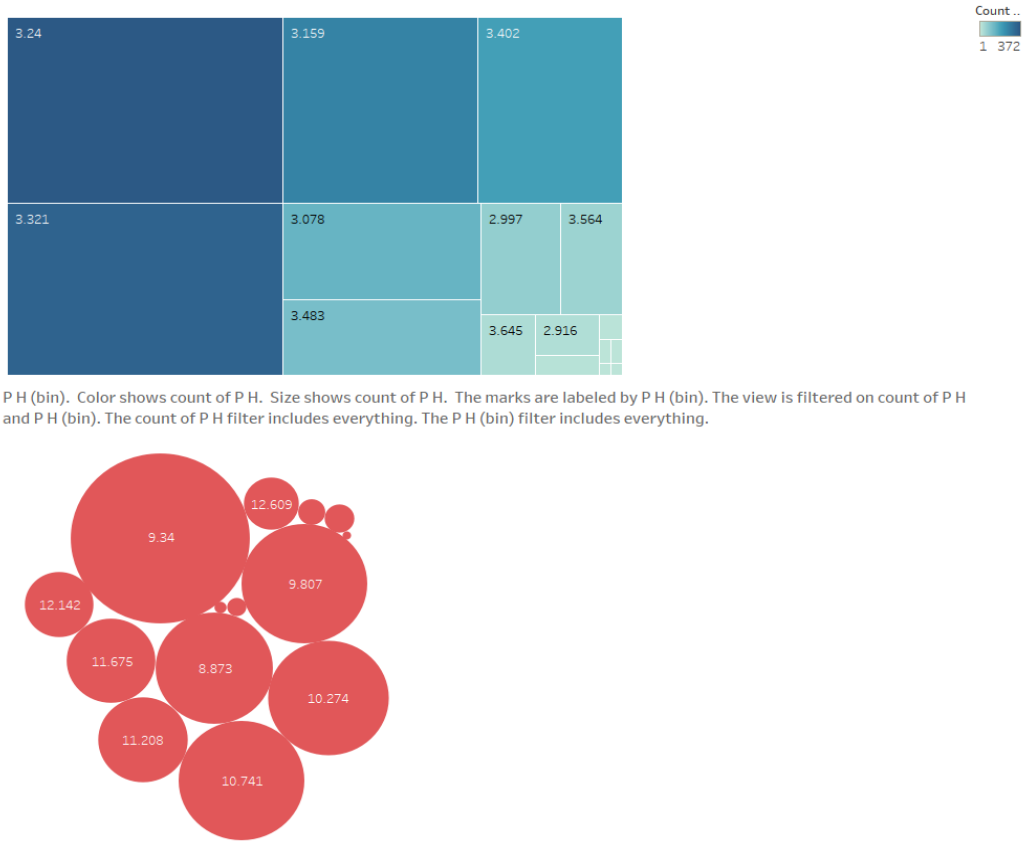

FIGURE 5.1: TABLEAU DASHBOARD 1

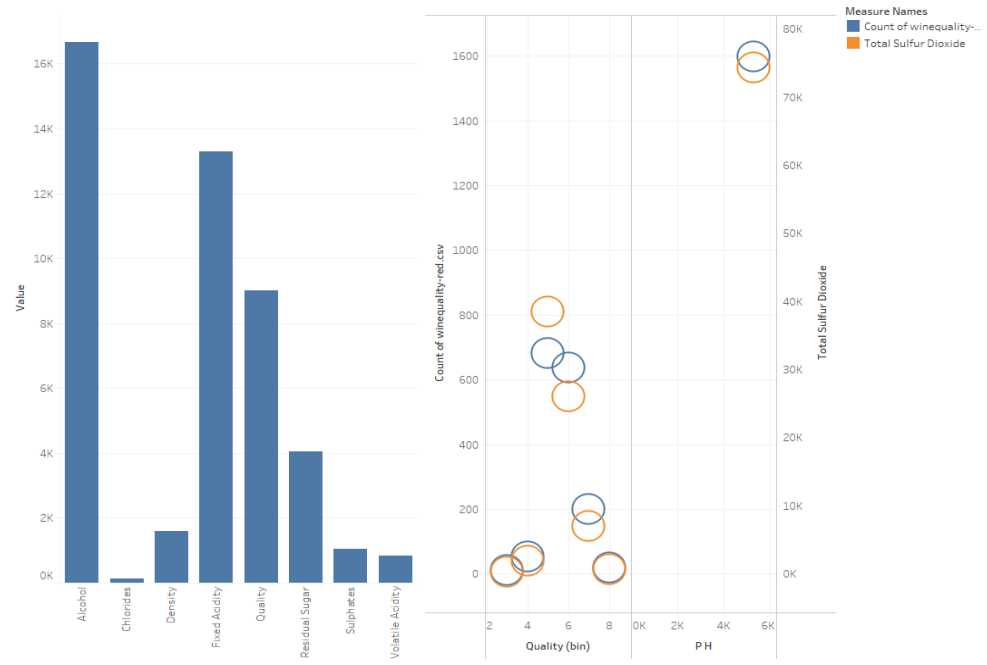

FIGURE 5.2: TABLEAU DASHBOARD 2

CHAPTER 6

CONCLUSION

The Wine Quality Prediction project aimed to leverage advanced data analytics and machine learning techniques to enhance the understanding of factors influencing wine quality. Through a comprehensive analysis of diverse features associated with wine production, ranging from chemical compositions to sensory attributes, this project has provided valuable insights into the intricacies of wine quality determination.

Key Findings:

1. Feature Importance: The analysis identified critical features that significantly impact wine quality, shedding light on the chemical and sensory aspects that contribute to the overall perception of a wine's excellence.

2. Correlation Patterns: Correlation analyses revealed complex relationships between different factors, offering a nuanced understanding of how variations in one aspect might influence others.

3. Predictive Modeling: The implemented machine learning models demonstrated a commendable ability to predict wine quality based on the selected features. The predictive power of the models opens avenues for quality control and optimization in wine production.

Implications for the Wine Industry:

1. Quality Improvement: Winemakers can utilize the predictive models to optimize production processes, ensuring a higher likelihood of producing wines

with superior quality.

2. Cost Efficiency: By pinpointing influential factors, the project contributes to cost-effective quality control measures, allowing wineries to allocate resources efficiently.

3. Market Competitiveness: Understanding the key determinants of wine quality positions winemakers to produce products that align with consumer preferences, enhancing competitiveness in the market.

Challenges and Future Directions:

1. Data Limitations: The project faced challenges associated with data limitations, emphasizing the importance of obtaining more extensive datasets for a more comprehensive analysis.

2. Continued Refinement: Ongoing refinement of models and analyses is essential to adapt to evolving industry trends, technological advancements, and expanding datasets.

This project not only contributes to the scientific understanding of wine quality but also offers tangible benefits for winemakers and industry stakeholders seeking to produce wines that captivate the palates of consumers worldwide. The fusion of data science and winemaking is a testament to the potential for innovation in even the most traditional industries.

FUTURE ENHANCEMENT:

The Wine Quality Prediction project lays the groundwork for future advancements and optimizations. Here are potential avenues for further development and enhancement:

1. Integration of Additional Features:
Sensory Data Expansion: Incorporate more detailed sensory data, potentially collected through advanced tasting panels or sensor technologies, to capture nuanced aspects of wine quality.

2. Dynamic Models for Changing Conditions:
Seasonal Variations: Develop models that account for seasonal variations in grape quality and winemaking conditions, providing adaptability to changing environmental factors.

3. Advanced Data Collection:
IoT Integration: Explore the integration of Internet of Things (IoT) devices within winemaking facilities to gather real-time data on fermentation conditions, temperature, and other critical parameters.

4. Fine-Tuning of Predictive Models:
Hyper parameter Tuning: Conduct extensive hyper parameter tuning for machine learning models to improve prediction accuracy and generalization on diverse datasets.

5. Incorporation of Unstructured Data:
Textual and Image Data: Integrate natural language processing (NLP) techniques for analyzing textual data, such as winemaking notes, and explore image analysis for label and bottle-related features.

6. Interactive Dashboard for Winemakers:
User-Friendly Interface: Develop an interactive and user-friendly dashboard that allows winemakers to input variables, visualize predictions, and receive actionable recommendations for quality improvement.

7. Benchmarking Against Industry Standards:
Quality Benchmarking: Establish benchmarks by comparing model predictions against industry-recognized wine quality standards, ensuring alignment with global quality expectations.

8. Collaboration with Winemaking Community:
Community Engagement: Foster collaboration with winemakers, industry

experts, and researchers to collectively refine models, share insights, and contribute to an open-source knowledge repository.

9. Exploration of Unsupervised Learning:
Clustering Analysis: Apply unsupervised learning techniques, such as clustering, to identify inherent patterns within datasets and discover hidden relationships between features.

10. Continuous Model Monitoring:
Real-Time Monitoring: Implement a real-time monitoring system for model performance, enabling timely adjustments and ensuring ongoing relevance in dynamic winemaking environments.

11. Ethical and Sustainability Considerations:
Wine Production Ethics: Explore ways in which data analytics can contribute to ethical and sustainable winemaking practices, aligning with growing consumer preferences for eco-friendly products.

These future enhancements aim to elevate the Wine Quality Prediction project, making it a dynamic and evolving tool for winemakers, researchers, and stakeholders in the viticulture and winemaking industry. The integration of emerging technologies and continuous refinement will contribute to the project's longevity and impact within the wine production landscape.

CHAPTER 7
REFERENCES

1) https://www.kaggle.com/code/gauravduttakiit/red-wine-quality-linear-regression/input

2) https://github.com/amberkakkar01/Prediction-of-Wine-Quality/blob/master/README.md

3)https://www.youtube.com/watch?v=W25TEa93T_I&ab_channel=HackersRealm

4) https://www.youtube.com/watch?v=ocR5dP5-XF0&ab_channel=DataScienceTutorials

5)http://ir.juit.ac.in:8080/jspui/bitstream/123456789/6574/1/Wine%20Quality%20Prediction%20using%20Machine%20Learning.pdf

APPENDIX- A (CODE)

(Libraries Imported)

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

(To Avoid Warnings)

```
import warnings
warnings.filterwarnings('ignore')
```

(To Read the Contents From the dataset)

```
wine = pd.read_csv(r"C:\Users\errat\Downloads\winequality-red.csv",encoding='unicode_escape')
wine.head()
wine.info()
wine.describe()
```

(To know no of rows and columns)

```
wine.shape
```

(Quality check)

```
round(100*(wine.isnull().sum()/len(wine)),2).sort_values(ascending=False)
round(100*(wine.isnull().sum(axis=1)/len(wine)),2).sort_values(ascending=False)
```

(No missing or Null value in either row or column)

```
dub_wine=wine.copy()
dub_wine.drop_duplicates(subset=None,inplace=True)
dub_wine.shape
```

(Assign non duplicate data to original data)

```
wine=dub_wine
for col in wine:
    print(wine[col].value_counts(ascending=False), '\n\n\n')
```

(Data split)

```
wine.info()
from sklearn.model_selection import train_test_split
np.random.seed(0)
df_train,df_test=train_test_split(wine,train_size=0.7,test_size=0.3,random_state
=100)
df_train.info()
df_train.shape
df_test.info()
df_train.shape
df_train.info()
```

(To draw different graphs)

(viloinplot)

```
sns.violinplot(x='quality', y='alcohol', data=wine)
```

(Kernel density estimation)

```
sns.kdeplot(wine.query('quality > 2').quality)
```

(Histogram)

```
wine.hist(figsize=(10,10),bins=50)
plt.show()
```

(Pairplot)

```
sns.pairplot(df_train)
plt.show()
df_train.columns
```

(heatmap-which is used to represent correlation)

```
plt.figure(figsize=(20,10))
sns.heatmap(wine.corr(), annot=True,cmap='RdBu')
plt.show()
```

(Piechart)

```
import plotly.graph_objects as go
quality = wine["quality"].value_counts()
fig = go.Figure(data=[go.Pie(labels=quality.index,
                             values = quality.values,
                             textinfo = 'label+percent',
                             insidetextorientation='radial'
                            )])
fig.show()
```

(Rescaling)

```
from sklearn.preprocessing import MinMaxScaler
scaler=MinMaxScaler()
df_train.head()
df_train.columns
```

```
df_train[:]=scaler.fit_transform(df_train[:])
df_train.head()
from sklearn.feature_selection import RFE
from sklearn.linear_model import LinearRegression
from sklearn.feature_selection import RFE
from sklearn.linear_model import LinearRegression
lm = LinearRegression()
num_features_to_select = 9
rfe = RFE(estimator=lm, n_features_to_select=num_features_to_select)
rfe = rfe.fit(X_train, y_train)
list(zip(X_train.columns,rfe.support_,rfe.ranking_))
col = X_train.columns[rfe.support_]
col
X_train.columns[~rfe.support_]
X_train_rfe = X_train[col]
```

(Building Linear Model)

(Model 1)/(VIF Check)

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
vif = pd.DataFrame()
vif['Features'] = X_train_rfe.columns
vif['VIF'] = [variance_inflation_factor(X_train_rfe.values, i) for i in
range(X_train_rfe.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```

```
import statsmodels.api as sm
X_train_lm1 = sm.add_constant(X_train_rfe)
lr1 = sm.OLS(y_train, X_train_lm1).fit()
lr1.params
print(lr1.summary())
```

(Model 2)

```
X_train_new = X_train_rfe.drop(["residual sugar"], axis = 1)
from statsmodels.stats.outliers_influence import variance_inflation_factor
vif = pd.DataFrame()
vif['Features'] = X_train_new.columns
vif['VIF'] = [variance_inflation_factor(X_train_new.values, i) for i in
range(X_train_new.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
X_train_lm2 = sm.add_constant(X_train_new)
lr2 = sm.OLS(y_train, X_train_lm2).fit()
lr2.params
print(lr2.summary())
```

(Residual Analysis of Training data)

```
y_train_pred = lr8.predict(X_train_lm8)
res = y_train-y_train_pred
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((res), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)  # Plot heading
plt.xlabel('Errors', fontsize = 18)
```

```
wine_num=wine[[ 'volatile acidity', 'alcohol', 'quality']]
sns.pairplot(wine_num)
plt.show()
vif = pd.DataFrame()
from statsmodels.stats.outliers_influence import variance_inflation_factor
vif['Features'] = X_train_new.columns
vif['VIF'] = [variance_inflation_factor(X_train_new.values, i) for i in
range(X_train_new.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```

(Applying the scaling on the test sets)

```
df_test[:]=scaler.fit_transform(df_test[:])
df_test.head()
df_test.describe()
```

(Dividing into X_test and Y_test)

```
y_test = df_test.pop('quality')
X_test = df_test
X_test.info()
#Selecting the variables that were part of final model.
col1=X_train_new.columns
X_test=X_test[col1]
# Adding constant variable to test dataframe
X_test_lm8 = sm.add_constant(X_test)
X_test_lm8.info()
```

```
y_pred = lr8.predict(X_test_lm8)
fig = plt.figure()
plt.scatter(y_test, y_pred, alpha=.5)
fig.suptitle('y_test vs y_pred', fontsize = 20)   # Plot heading
plt.xlabel('y_test', fontsize = 18)  # X-label
plt.ylabel('y_pred', fontsize = 16)
plt.show()
df= pd.DataFrame({'Actual':y_test,'Predictions':y_pred})
df['Predictions']= round(df['Predictions'],2)
df.head()
```

APPENDIX-B SCREENSHOTS

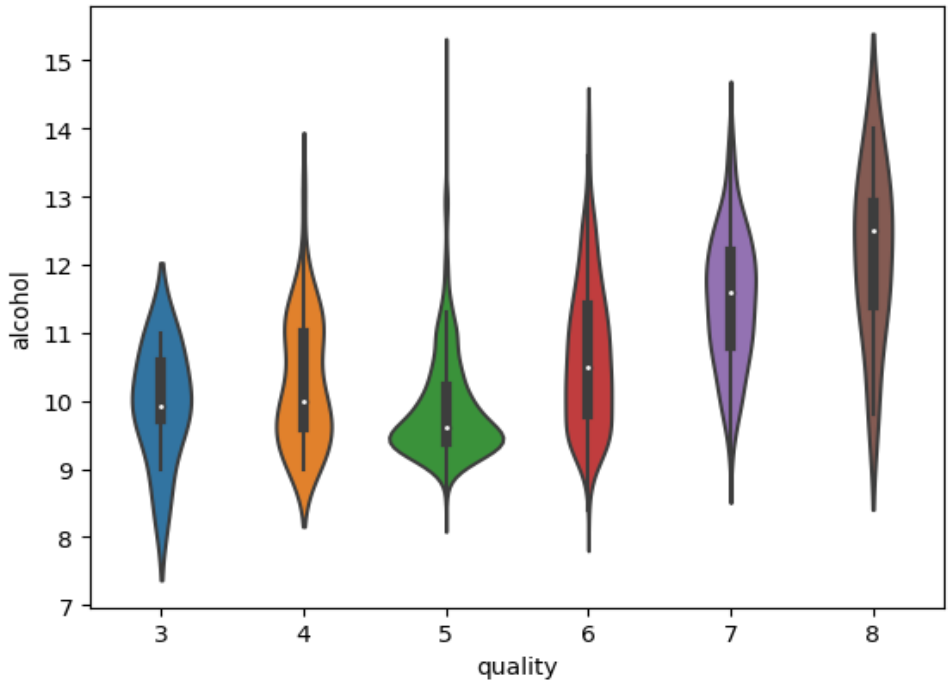

FIGURE B1: VIOLIN PLOT

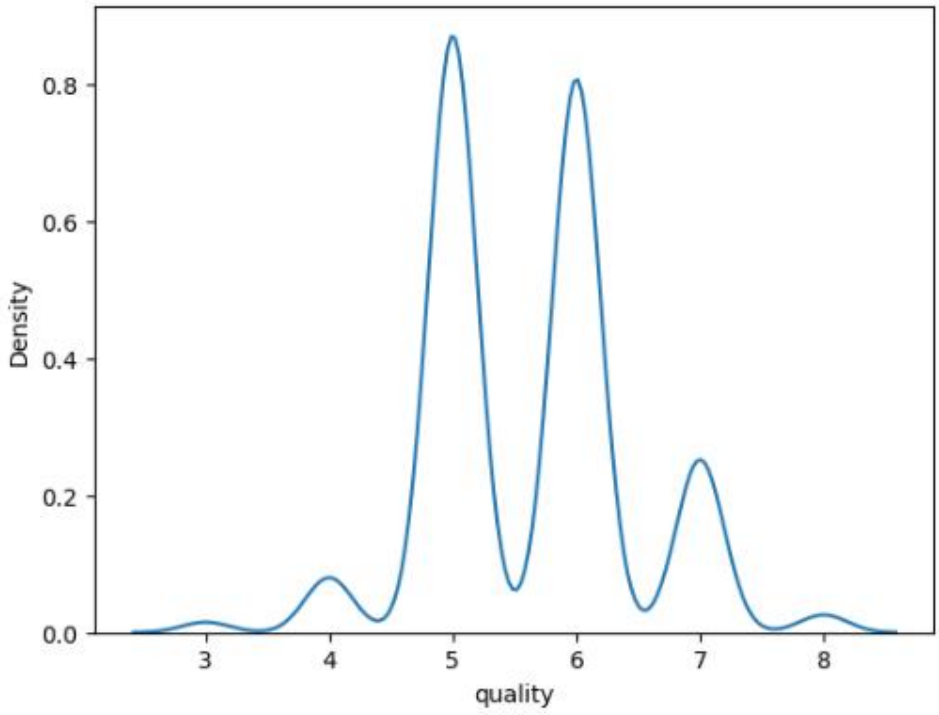

FIGURE B2: KERNEL DENSITY ESTIMATE(KDE)

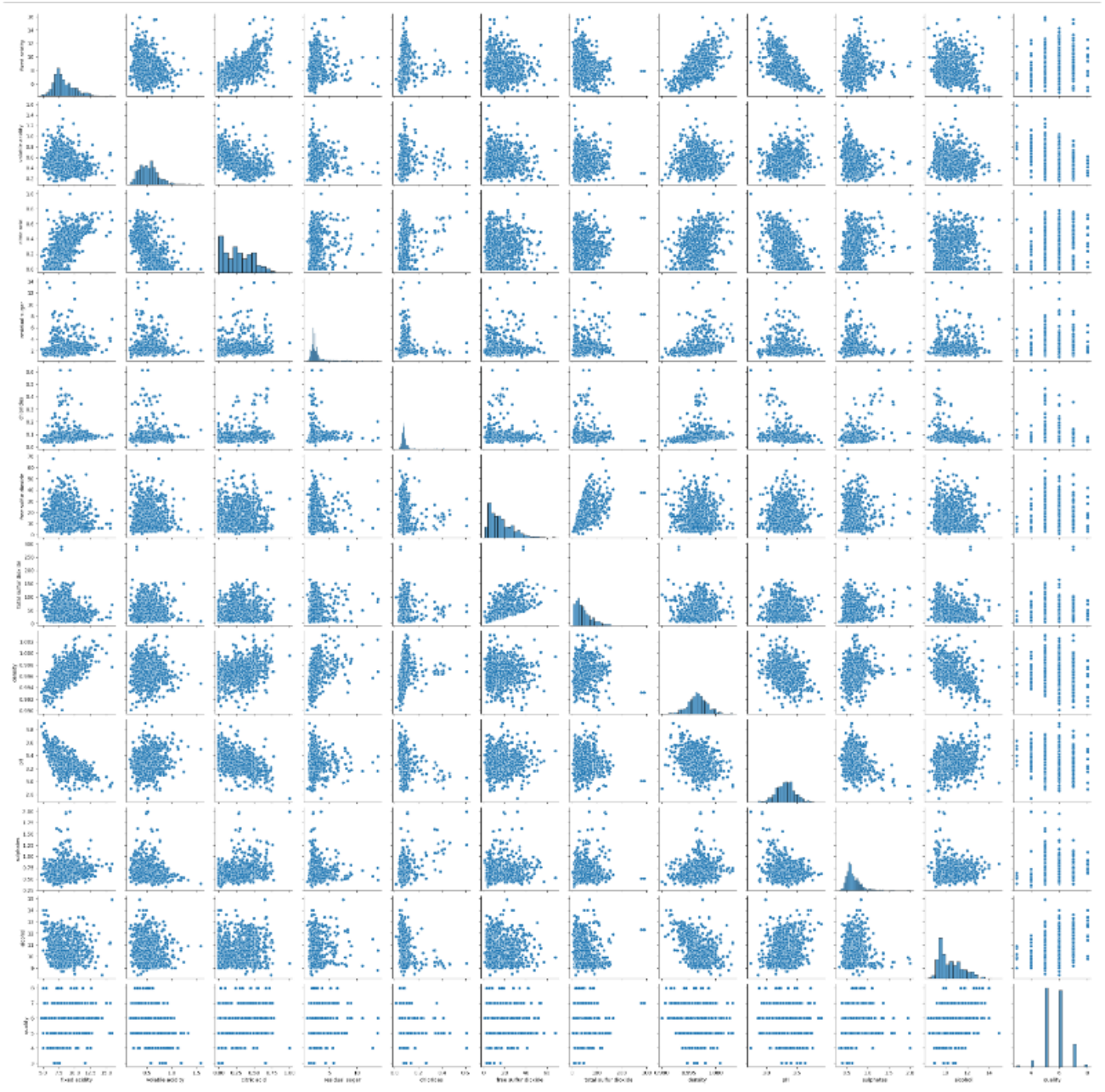

FIGURE B3: PAIRPLOT

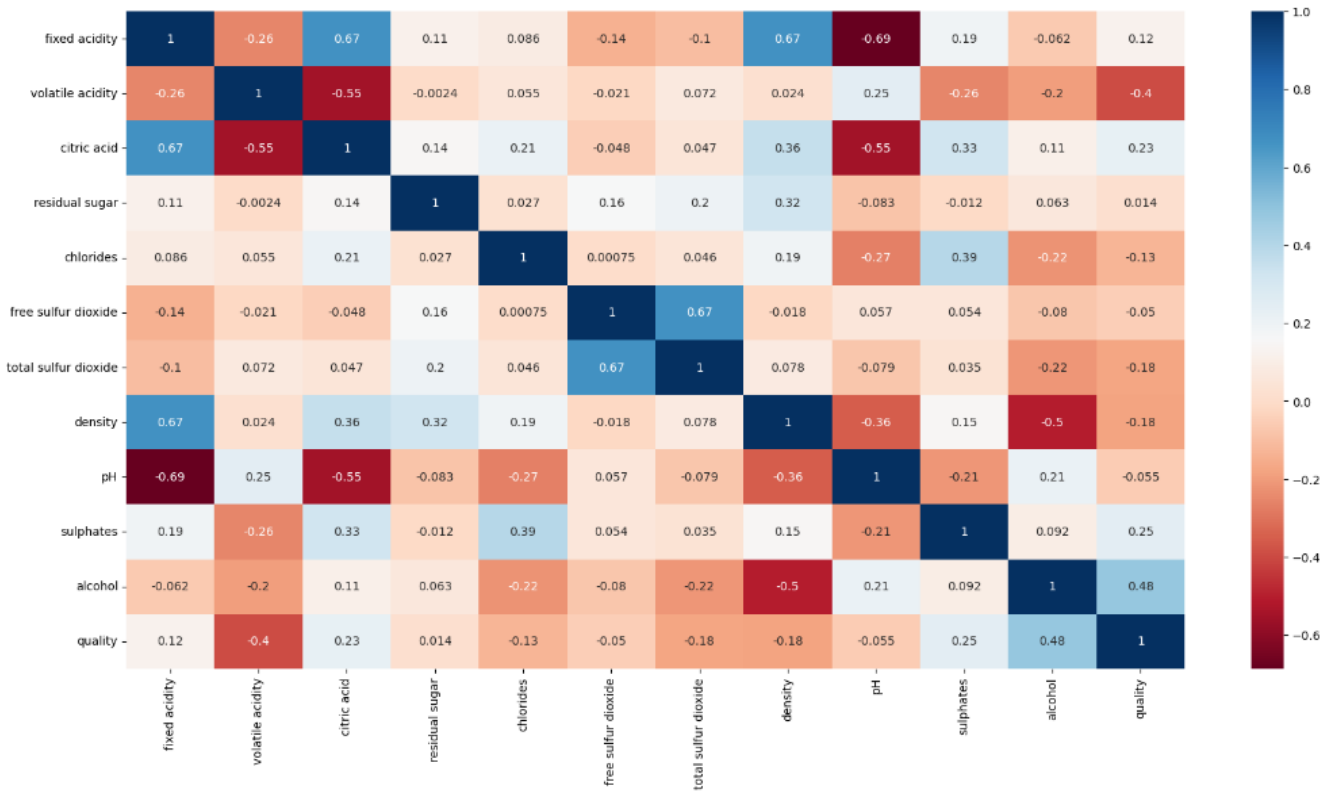

FIGURE B4: HEATMAP

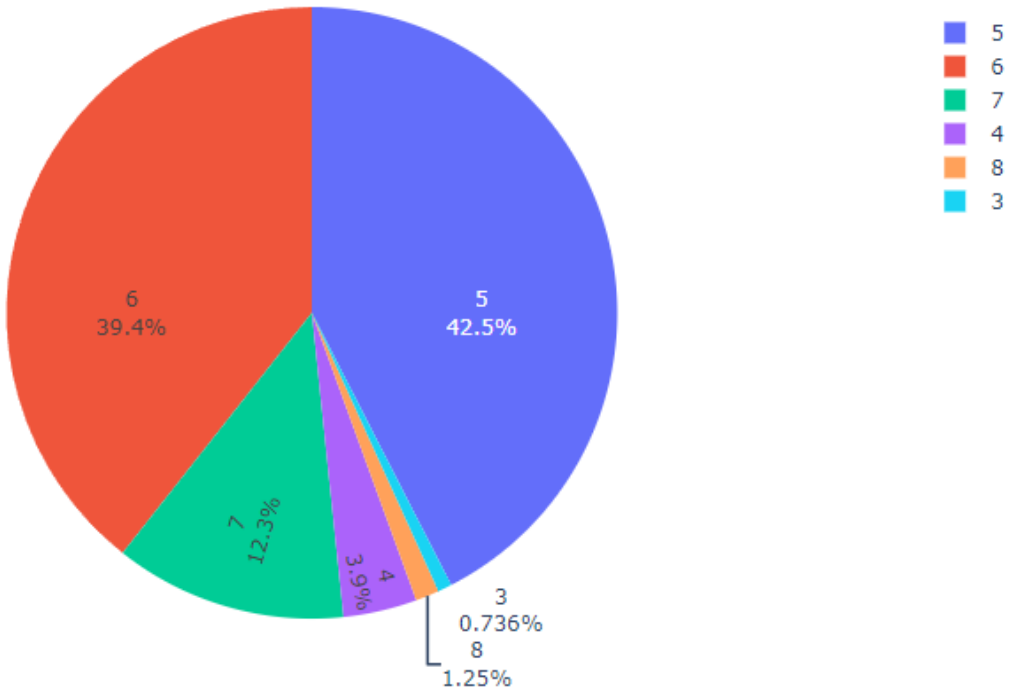

FIGURE B5: PIECHART

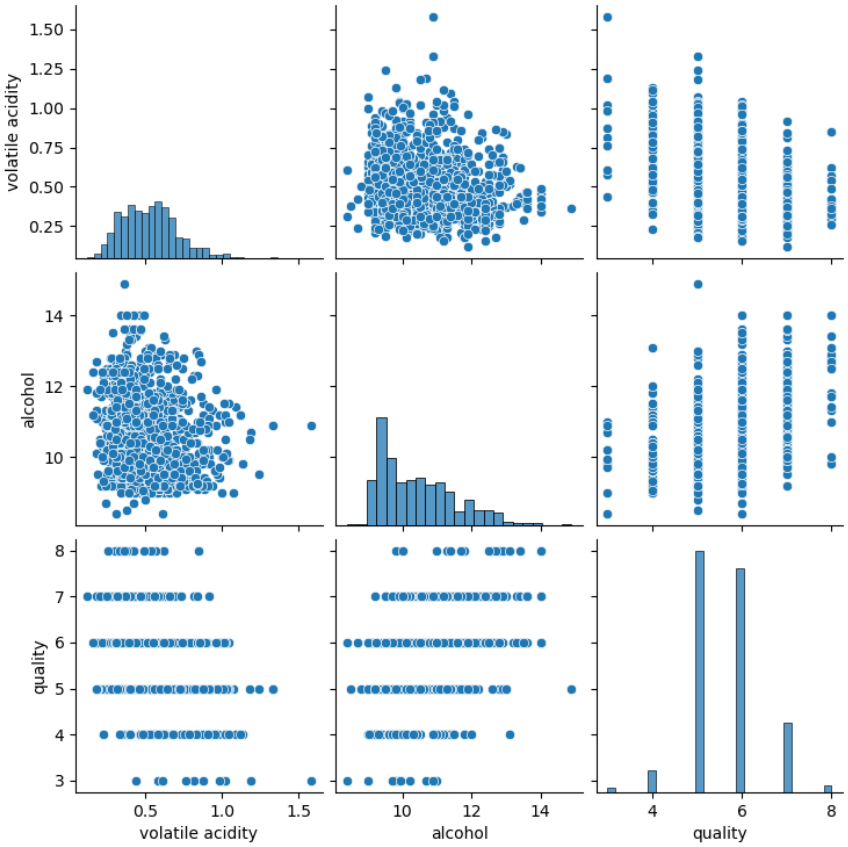

FIGURE B6: RELATIONSHIP BETWEEN X AND Y

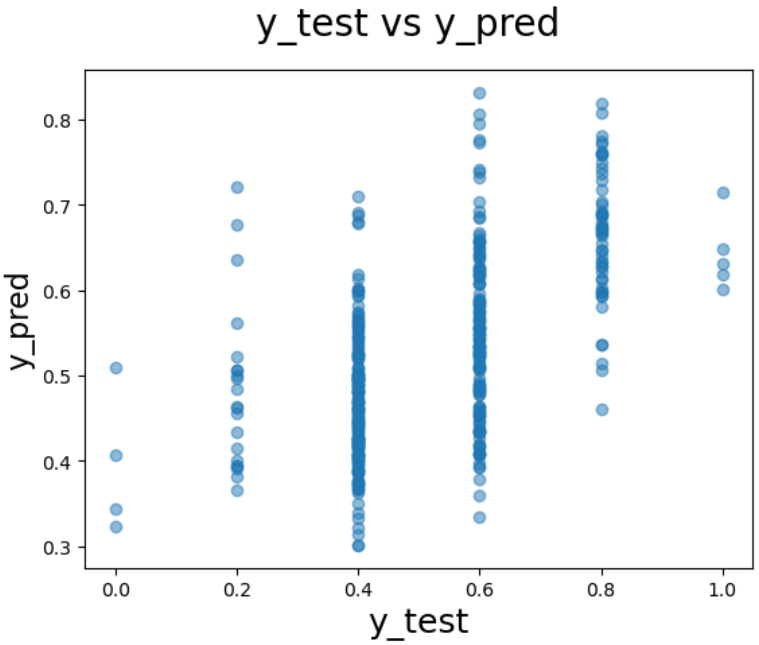

FIGURE B7: YTEST VS YPREDICTION