

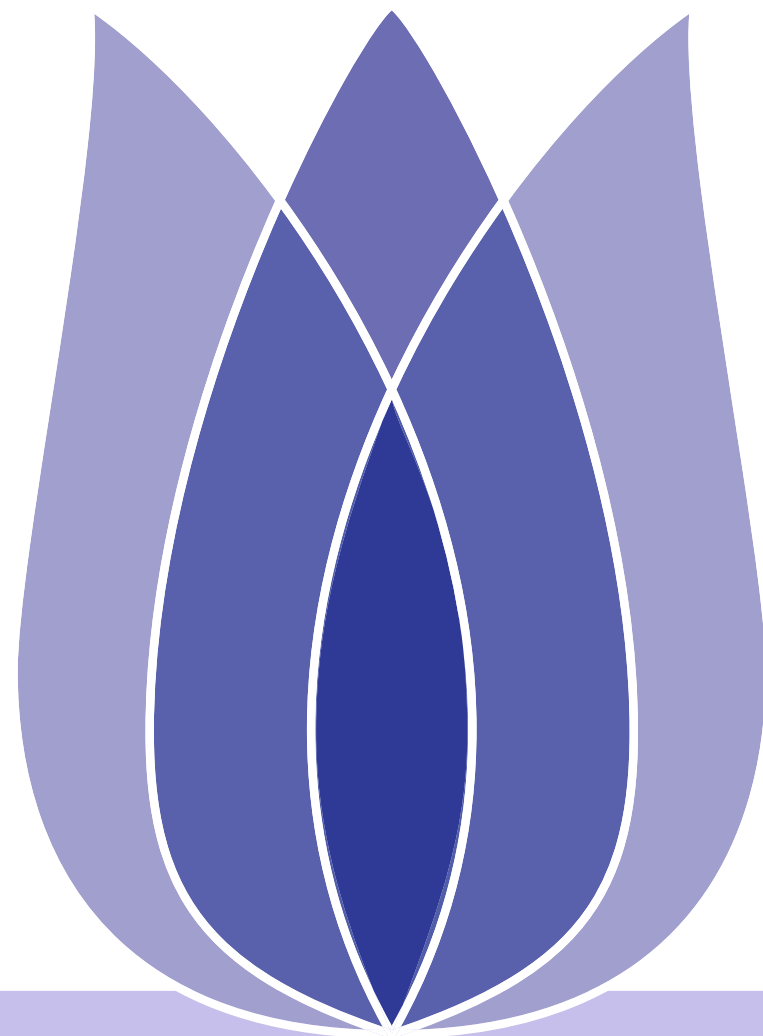
Ordinal Regression with a Tabular Wine Quality Dataset

Sen Han, Gang Li

Beijing Technology and Business University

Deakin University

(None)





Overview

- [Introduction](#)
- [Preliminaries](#)
- [Experiment and analysis](#)
- [Conclusion](#)
- [Submission](#)
- [Acknowledgement](#)

Introduction

Problem Description

Dataset Description

Preliminaries

Wine Features

SMOTE

Scoring Method

Experiment and analysis

Dataset Analysis

Analysis So Far

Classification Approach

Conclusion

Submission

Acknowledgement



Introduction

Problem Description

Dataset Description

Preliminaries

Experiment and analysis

Conclusion

Submission

Acknowledgement

Introduction



Problem Description

- Introduction
- Problem Description**
- Dataset Description
- Preliminaries
- Experiment and analysis
- Conclusion
- Submission
- Acknowledgement

Defn

Ordinal regression was conducted on a dataset derived from a deep learning model trained on the quality dataset of red variants of the "Vinho Verde" wine from Spain. This dataset characterizes the impact of various chemical substances present in wine on its quality. The quality grades are ordinal and imbalanced, with common wines being significantly more prevalent than either high-quality or low-quality wines.



Dataset Description

- [Introduction](#)
- [Problem Description](#)
- [Dataset Description](#)
- [Preliminaries](#)
- [Experiment and analysis](#)
- [Conclusion](#)
- [Submission](#)
- [Acknowledgement](#)

Wine Quality

This datasets is related to red variants of the Portuguese "Vinho Verde" wine.The dataset describes the amount of various chemicals present in wine and their effect on it’s quality. The datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).Your task is to predict the quality of wine using the given data.



- [Introduction](#)
- [Preliminaries](#)**
- [Wine Features](#)
- [SMOTE](#)
- [Scoring Method](#)
- [Experiment and analysis](#)
- [Conclusion](#)
- [Submission](#)
- [Acknowledgement](#)

Preliminaries



Wine Features

Introduction
Preliminaries
Wine Features
SMOTE
Scoring Method
Experiment and analysis
Conclusion
Submission
Acknowledgement

In order to facilitate an understanding of the meanings of various data points in the dataset, it is necessary to provide a brief introduction to several features of wine that are relevant to the dataset.

- **Fixed acidity:** Most acids involved with wine or fixed or nonvolatile (do not evaporate readily).
- **Volatile acidity:** The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste.
- **Citric acid:** Found in small quantities, citric acid can add 'freshness' and flavor to wines.
- **Residual sugar:** The amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet. SS



Wine Features

Introduction
Preliminaries
Wine Features
SMOTE
Scoring Method
Experiment and analysis
Conclusion
Submission
Acknowledgement

- **Chlorides:** The amount of salt in the wine.
- **Free sulfur dioxide:** The free form of SO_2 exists in equilibrium between molecular SO_2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine.
- **Total sulfur dioxide:** Amount of free and bound forms of SO_2 ; in low concentrations, SO_2 is mostly undetectable in wine, but at free SO_2 concentrations over 50 ppm, SO_2 becomes evident in the nose and taste of wine.
- **Density:** The density of water is close to that of water depending on the percent alcohol and sugar content.
- **pH:** Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale.
- **Sulphates:** A wine additive which can contribute to sulfur dioxide gas (SO_2) levels, which acts as an antimicrobial and antioxidant.
- **Alcohol:** The percent alcohol content of the wine.
- **Quality:** Wine quality rating.



- Introduction
- Preliminaries
- Wine Features
- SMOTE**
- Scoring Method
- Experiment and analysis
- Conclusion
- Submission
- Acknowledgement

- **Data Imbalance in Dataset:** In the provided dataset, the samples with quality grades 5 and 6 are substantially more numerous than those of other grades, necessitating a consideration of the potential impacts of employing the SMOTE technique.
- **What is SMOTE?** SMOTE (Synthetic Minority Over-sampling Technique) is a method employed in data science and machine learning to address the issue of class imbalance in classification problems. Class imbalance refers to the scenario where the instance count of one class (the minority class) is significantly lower than that of other classes (the majority classes).
- **Impact on Machine Learning:** This imbalance can lead to biased performance or suboptimal results in machine learning models, as they are often dominated by the majority class and tend to overlook the minority class.
- **Benefits of SMOTE:** By generating synthetic samples, SMOTE aids in reducing the overfitting issues that simple over-sampling might cause.



Scoring Method

[Introduction](#)

[Preliminaries](#)

[Wine Features](#)

[SMOTE](#)

[Scoring Method](#)

[Experiment and analysis](#)

[Conclusion](#)

[Submission](#)

[Acknowledgement](#)

- Submissions are scored on the **quadratic weighted kappa**, which measures agreement between two ratings.
- This metric varies from 0 (random agreement) to 1 (complete agreement). It may be negative if there is less agreement than chance.
- The calculation involves several steps:
 1. Construct an $N \times N$ histogram matrix O where $O_{i,j}$ is the number of agreements between raters.
 2. Calculate a weight matrix w as follows:

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

3. Calculate the expected outcome matrix E , assuming no correlation as the outer product of sums of O .
4. The **quadratic weighted kappa** is then:

$$\kappa = 1 - \frac{\sum w_{i,j} O_{i,j}}{\sum w_{i,j} E_{i,j}}$$





[Introduction](#)

[Preliminaries](#)

[Experiment and analysis](#)

[Dataset Analysis](#)

[Analysis So Far](#)

[Classification Approach](#)

[Conclusion](#)

[Submission](#)

[Acknowledgement](#)

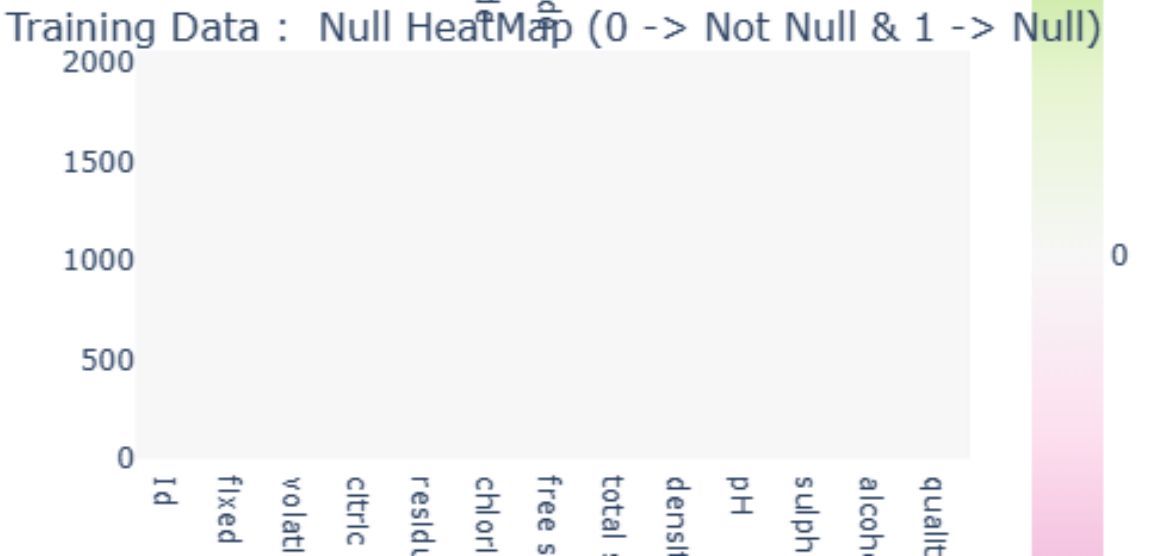
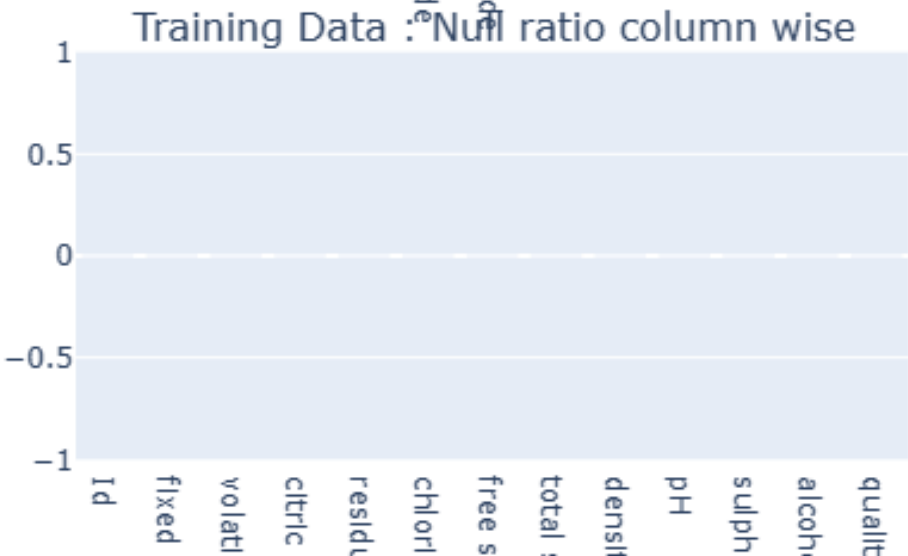
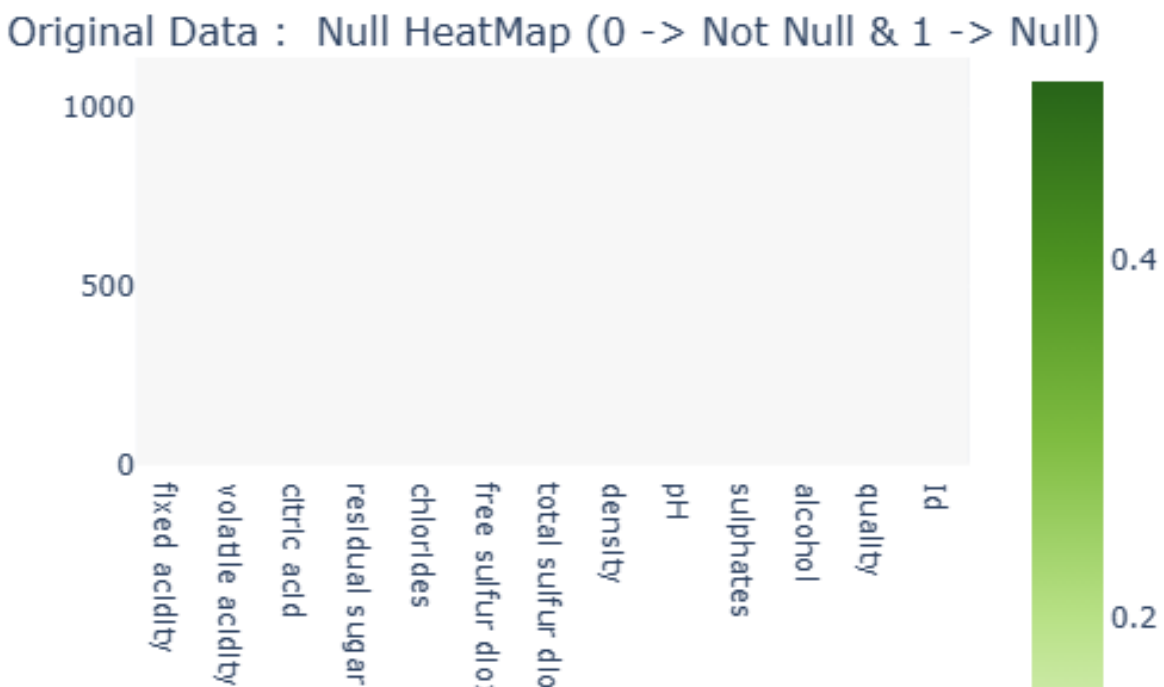
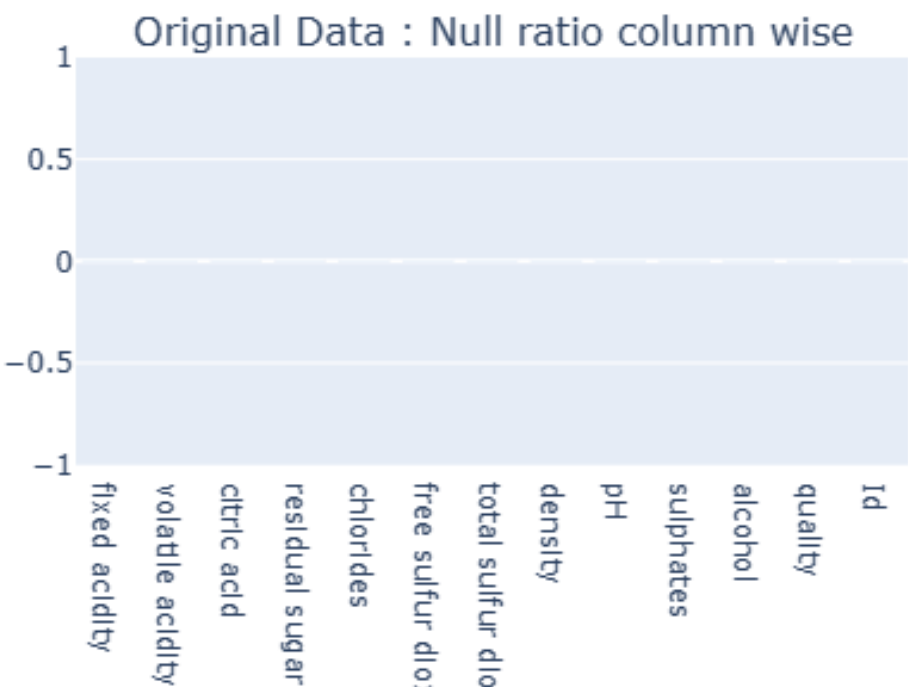
Experiment and analysis



Dataset Analysis

- Introduction
- Preliminaries
- Experiment and analysis
- Dataset Analysis**
- Analysis So Far
- Classification Approach
- Conclusion
- Submission
- Acknowledgement

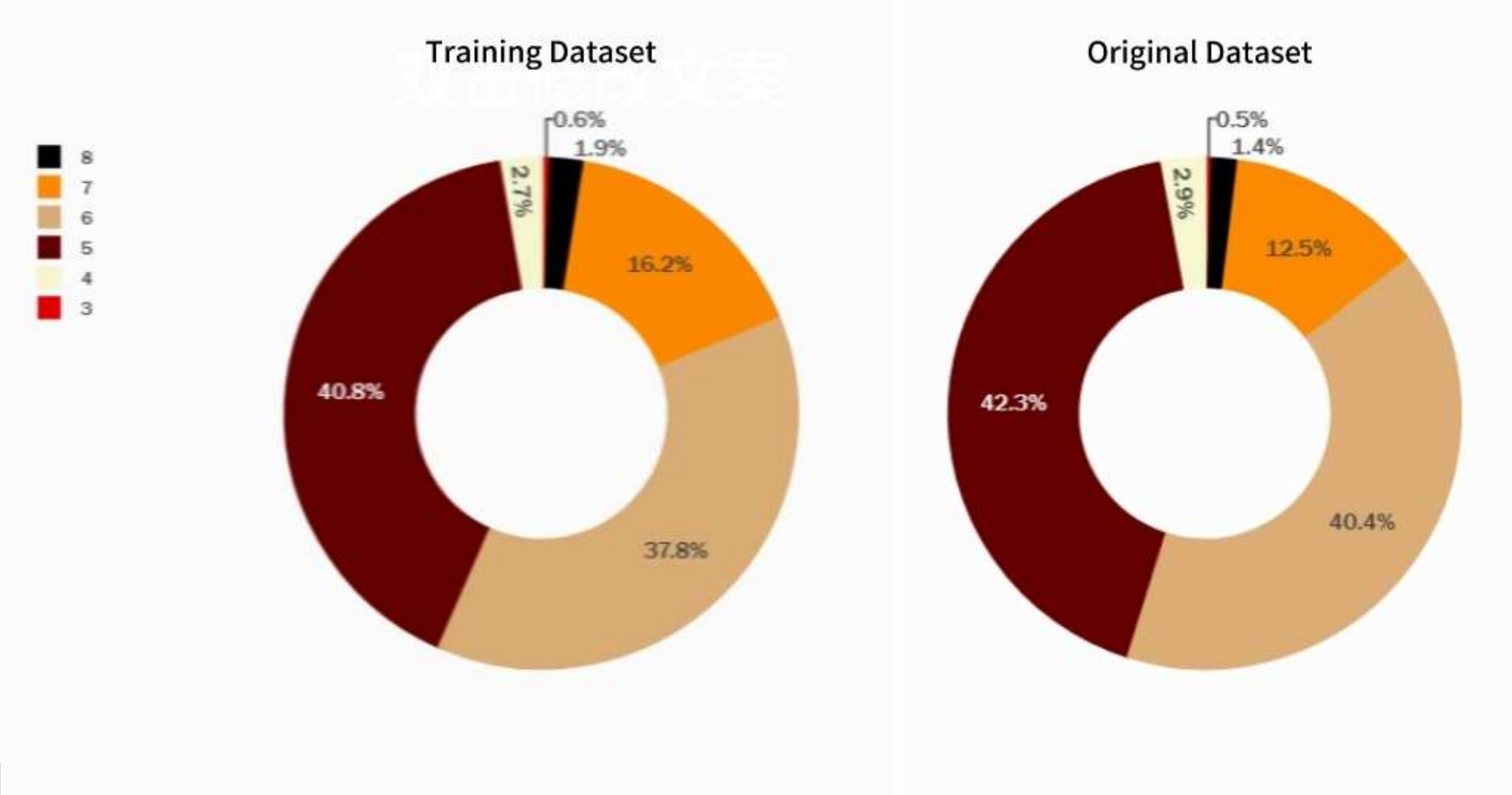
- Analysis of the Original Dataset, Training Dataset, and Testing Dataset
 - Null Value Detection
- The datasets do not contain any missing values, thus negating the necessity for missing value processing.





- [Introduction](#)
- [Preliminaries](#)
- [Experiment and analysis](#)
- [Dataset Analysis](#)**
- [Analysis So Far](#)
- [Classification Approach](#)
- [Conclusion](#)
- [Submission](#)
- [Acknowledgement](#)

■ Target Variable Analysis

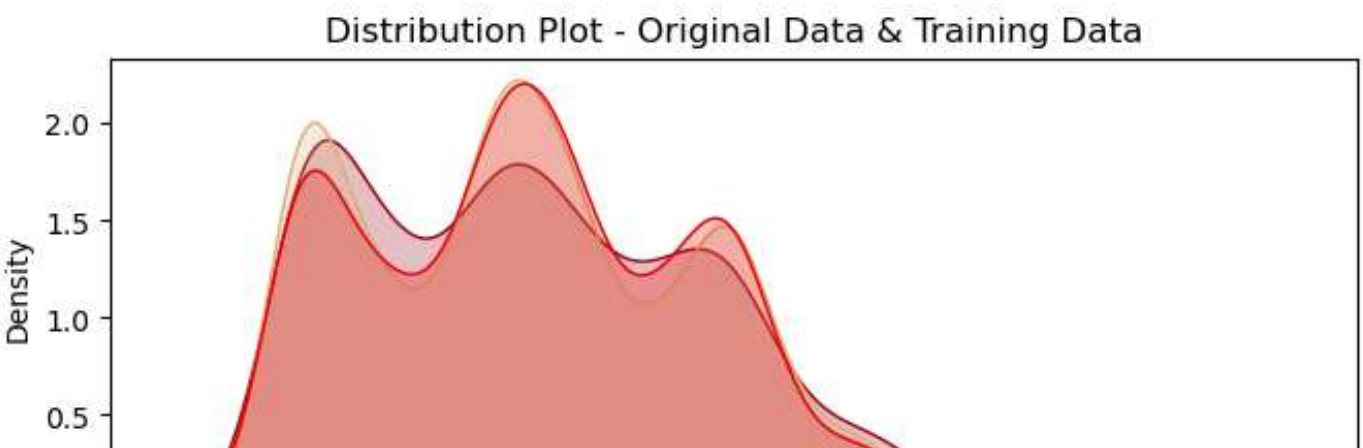
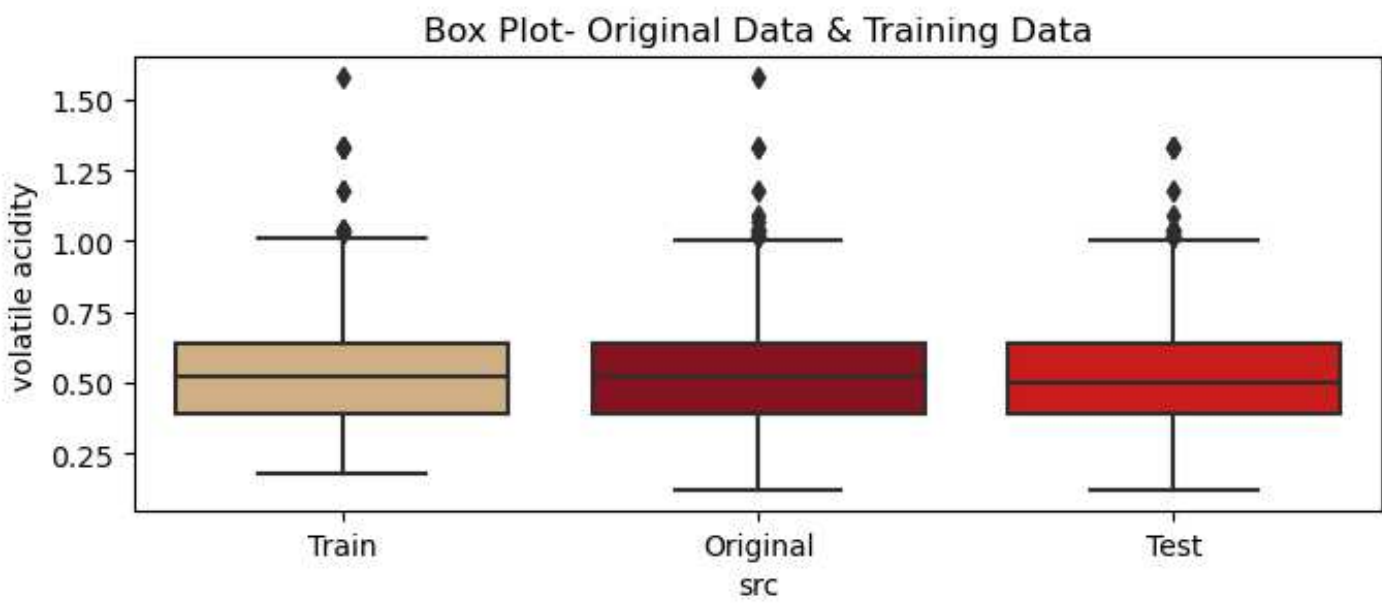
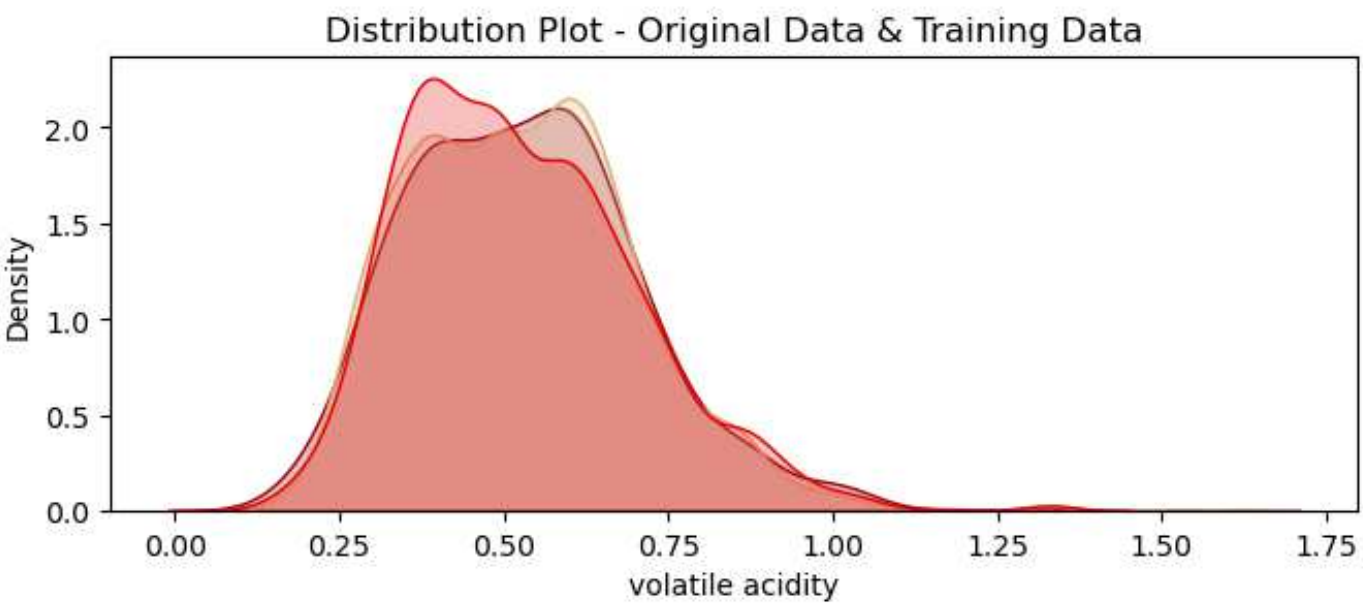
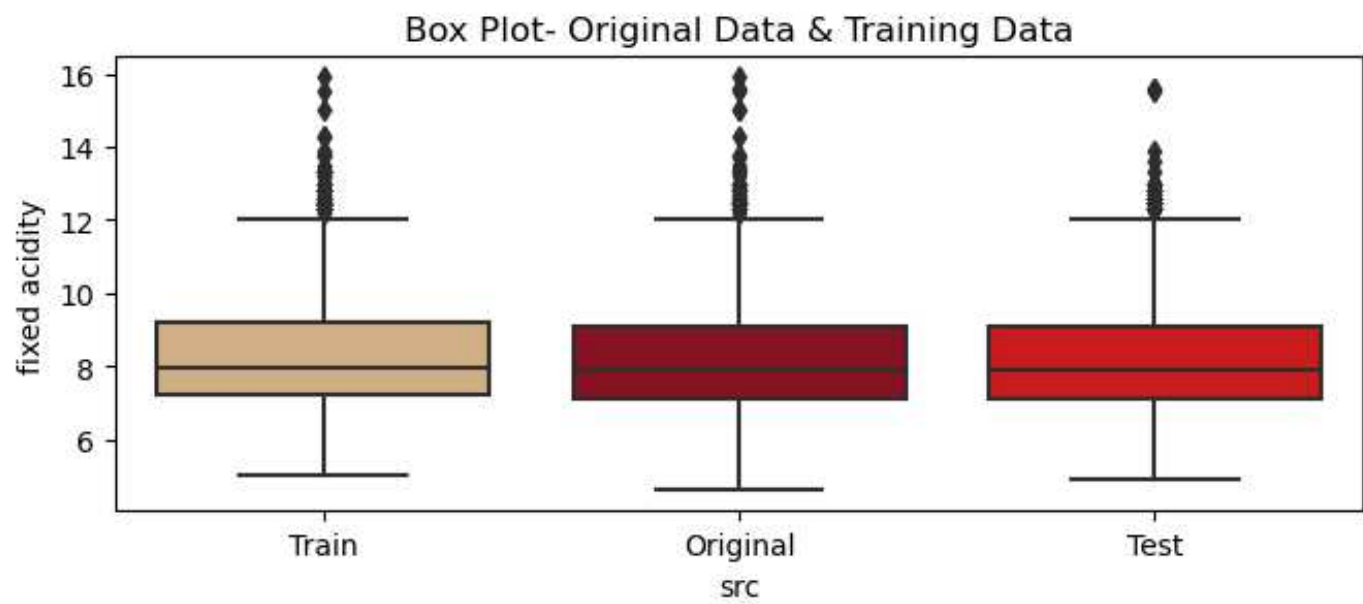
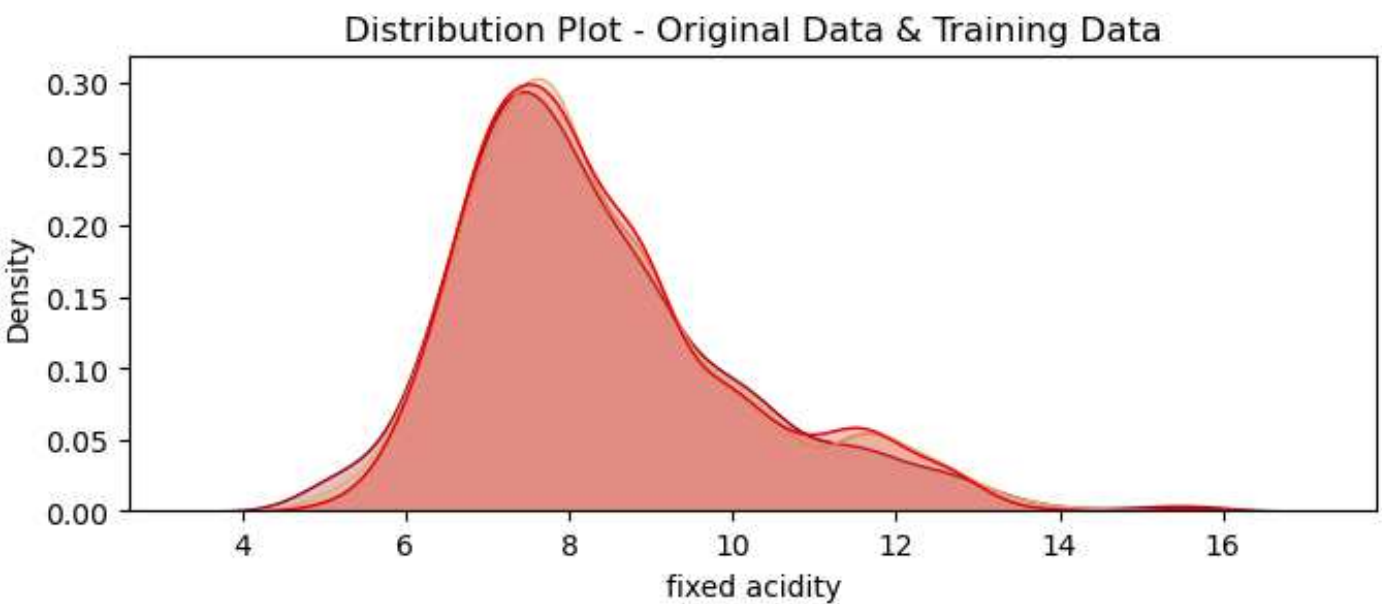




Dataset Analysis

- [Introduction](#)
- [Preliminaries](#)
- [Experiment and analysis](#)
- [Dataset Analysis](#)**
- [Analysis So Far](#)
- [Classification Approach](#)
- [Conclusion](#)
- [Submission](#)
- [Acknowledgement](#)

■ Univariate Analysis





Dataset Analysis

- [Introduction](#)
- [Preliminaries](#)
- [Experiment and analysis](#)
- [Dataset Analysis](#)**
- [Analysis So Far](#)
- [Classification Approach](#)
- [Conclusion](#)
- [Submission](#)
- [Acknowledgement](#)

For all three datasets, the distributions are almost identical, thus enabling the amalgamation of the original dataset with the training dataset for the purpose of training.

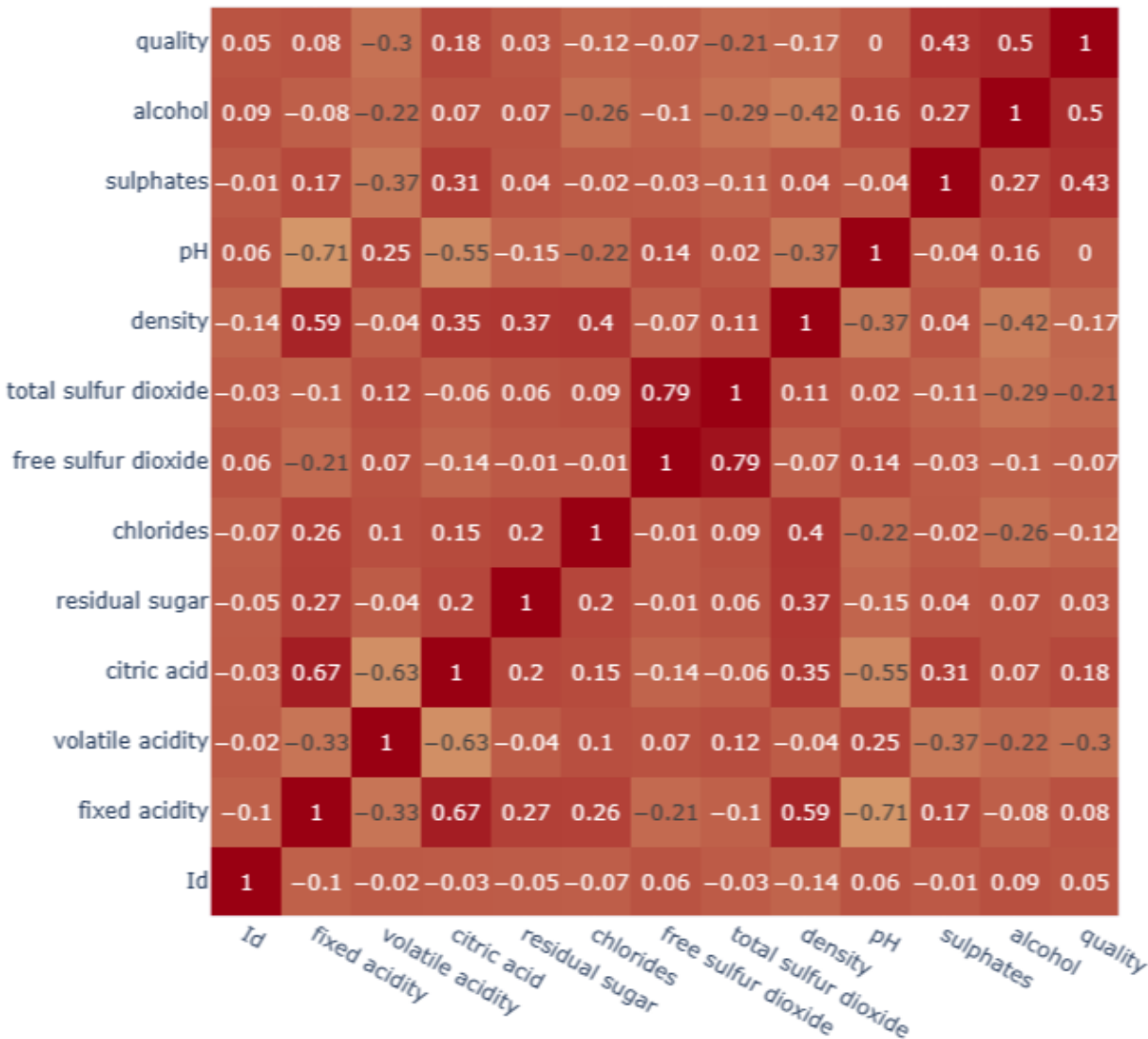


Combining Data and Relationship Matrix

- Introduction
- Preliminaries
- Experiment and analysis
- Dataset Analysis**
- Analysis So Far
- Classification Approach
- Conclusion
- Submission
- Acknowledgement

- Merged the original dataset with the training dataset to create a combined training dataset.
- Calculated the pairwise Spearman correlation coefficients for all numerical columns in the dataset and visualized them in a matrix form.

The heatmap below illustrates the monotonic relationship between pairs of variables, using a color scale to signify the strength of the correlation. Darker shades represent stronger correlations, with one color for positive and another for negative correlations. The numerical values in the heatmap’s cells indicate the Spearman correlation coefficients for the features they intersect.





Multivariate Analysis wrt Quality

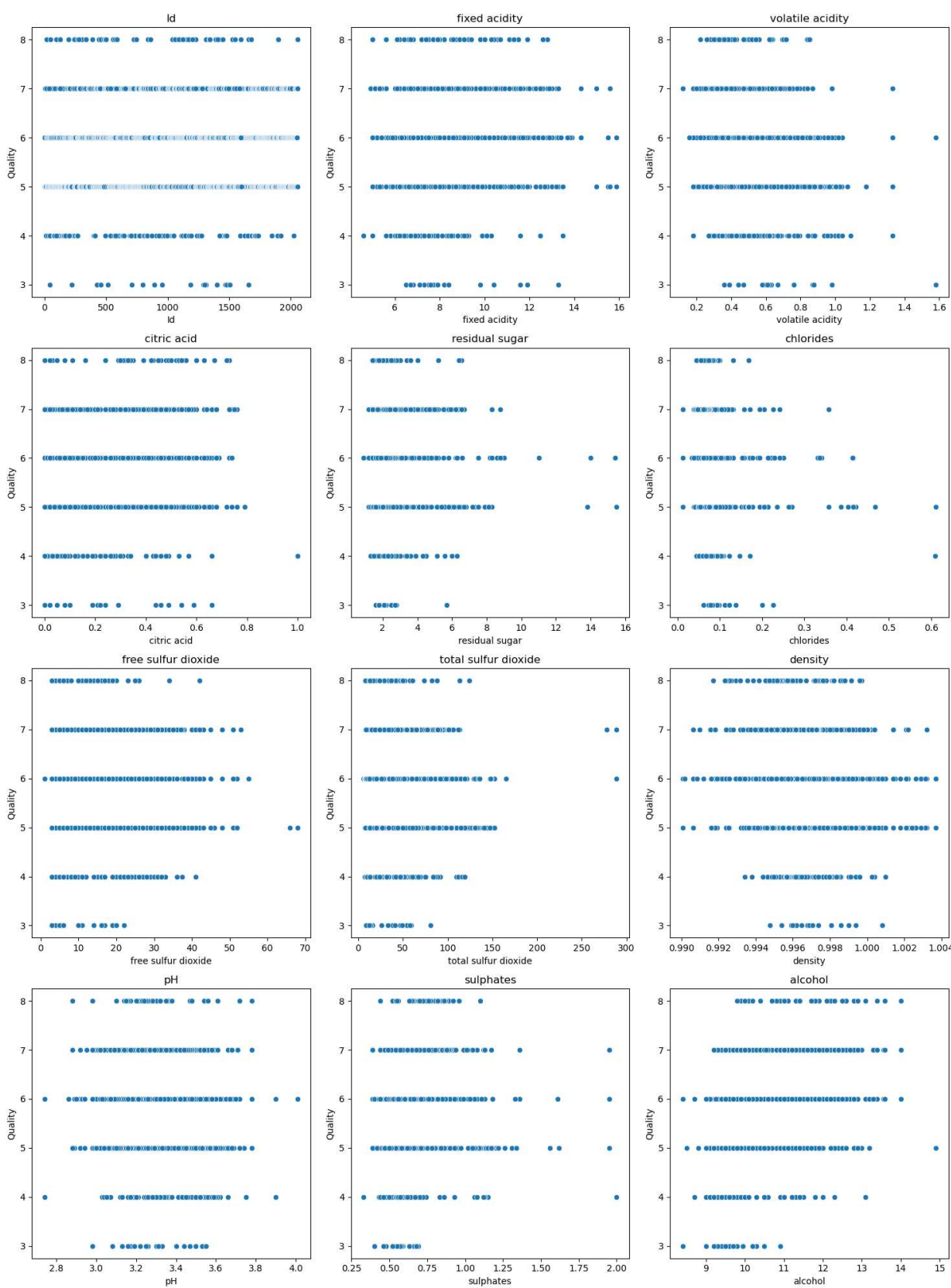
- Introduction
- Preliminaries
- Experiment and analysis
- Dataset Analysis**
- Analysis So Far
- Classification Approach
- Conclusion
- Submission
- Acknowledgement

Relation with Quality using pairplot (Combined Training and Original Data)

Analyzing the relationship between various features and the wine quality through visual methods provides insightful observations:

- Higher alcohol content tends to correspond with higher quality ratings.
- Lower volatile acidity is often associated with higher quality wines.
- A higher sulphate content might indicate a trend towards higher quality.

Such visual analyses are pivotal in understanding the influence of individual components on the overall quality of wine.





Relation with Quality using Boxplot

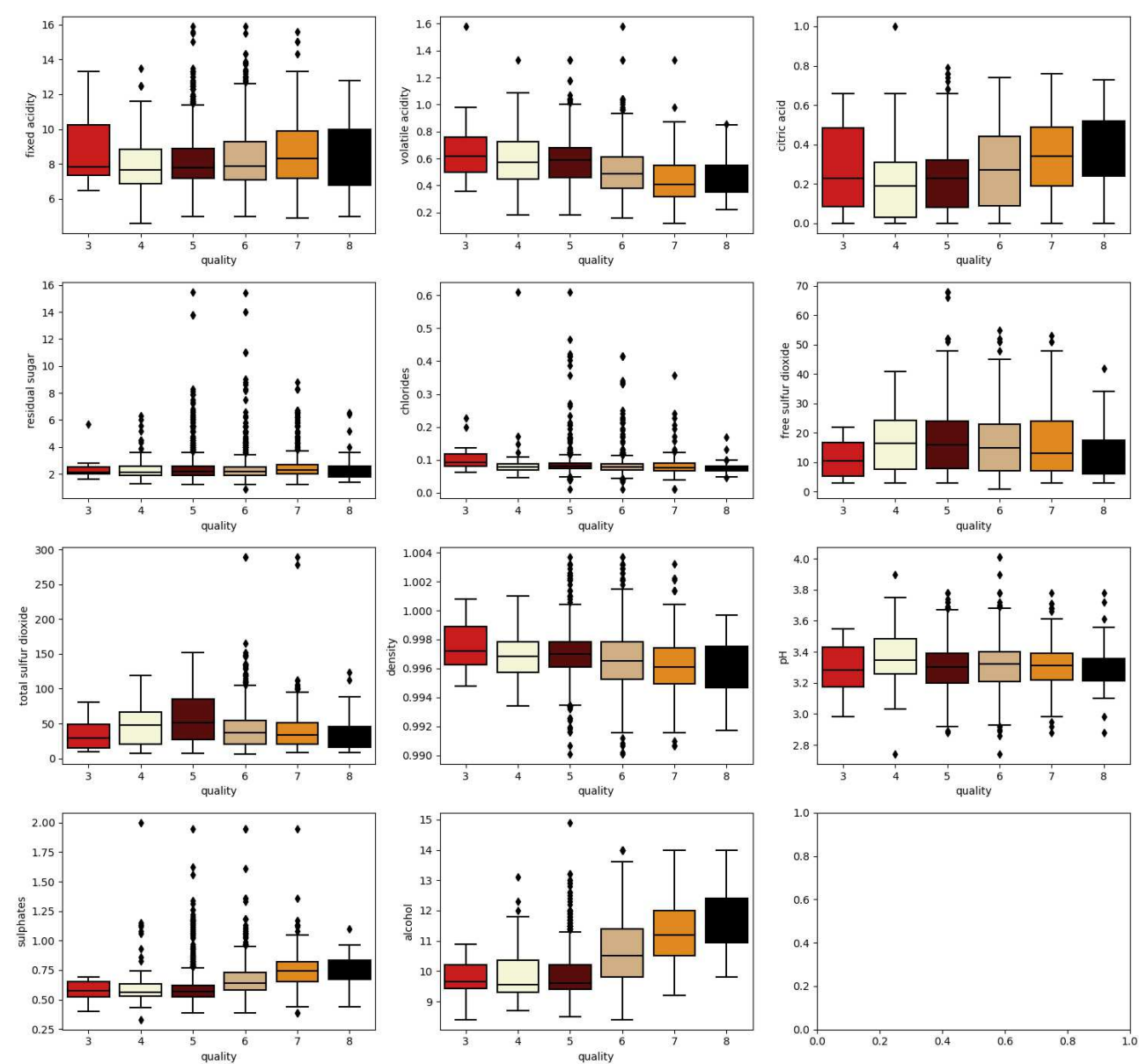
- Introduction
- Preliminaries
- Experiment and analysis
- Dataset Analysis**
- Analysis So Far
- Classification Approach
- Conclusion
- Submission
- Acknowledgement

Combined Training and Original Data Analysis:

Through boxplots, we can observe the distribution of features across different quality levels. Key insights include:

- Higher-quality wines have higher medians in sulphates, alcohol, and citric acid.
- Conversely, they have lower medians in chlorides, density, and residual sugar.

These observations can guide further analysis and predictive modeling.





Analysis So Far

- [Introduction](#)
- [Preliminaries](#)
- [Experiment and analysis](#)
- [Dataset Analysis](#)
- [Analysis So Far](#)
- [Classification Approach](#)
- [Conclusion](#)
- [Submission](#)
- [Acknowledgement](#)

- The distribution of data points across the training, testing, and original datasets is approximately the same, suggesting consistency in data collection or generation.
- The relationships between variables show similar patterns in both the test set and the combined dataset (which includes training and original data), indicating that the test set may be a representative subset of the overall data.



Data Preparation: Normal vs SMOTE

- Introduction
- Preliminaries
- Experiment and analysis
- Dataset Analysis
- Analysis So Far**
- Classification Approach
- Conclusion
- Submission
- Acknowledgement

Regular Training Data

- Class distribution in the combined training dataset shows imbalance:

Quality	Count
5	925
6	868
7	333
4	62
8	38
3	13

SMOTE Training Data

- After applying SMOTE, class distribution is balanced:

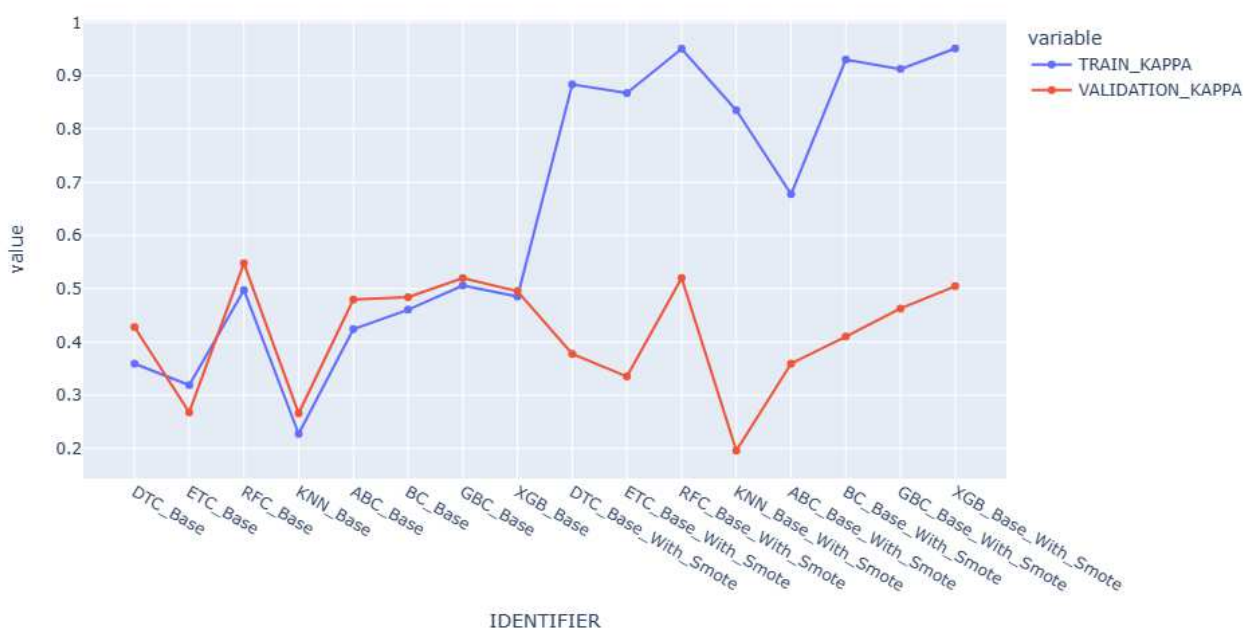
Quality	Count
3	925
4	925
5	925
6	925
7	925
8	925



Classification Approach

- Introduction
- Preliminaries
- Experiment and analysis
- Dataset Analysis
- Analysis So Far
- Classification Approach**
- Conclusion
- Submission
- Acknowledgement

Train various models separately on 70% of the data randomly drawn from both the combined training dataset and the SMOTE-enhanced training dataset. Then test these models on the remaining 30% of the data to calculate their respective Kappa scores.





[Introduction](#)

[Preliminaries](#)

[Experiment and analysis](#)

[Conclusion](#)

[Submission](#)

[Acknowledgement](#)

Conclusion



Conclusion

- [Introduction](#)
- [Preliminaries](#)
- [Experiment and analysis](#)
- [Conclusion](#)
- [Submission](#)
- [Acknowledgement](#)

In scenarios where models are trained using both the combined training dataset and the SMOTE-enhanced training dataset, it appears that there is no significant difference in performance on the validation set. However, the use of the SMOTE-enhanced dataset seems to improve model performance on the training set. This indicates that the application of the SMOTE technique may aid in better fitting the model to the training data, particularly when dealing with class imbalances in datasets.



[Introduction](#)

[Preliminaries](#)

[Experiment and analysis](#)

[Conclusion](#)

[Submission](#)

[Acknowledgement](#)

Submission



Submission

Introduction
Preliminaries
Experiment and analysis
Conclusion
Submission
Acknowledgement

- After an extensive evaluation of different models under various conditions,
- the Random Forest, Light Gradient Boosting Machine, and Gradient Boosting Classifier have been chosen.
- These models will be applied to the test dataset for ordinal regression using a voting mechanism (mode),
- and the predictions will be exported to a file named `xyz.csv`.



- [Introduction](#)
- [Preliminaries](#)
- [Experiment and analysis](#)
- [Conclusion](#)
- [Submission](#)
- [Acknowledgement](#)

Acknowledgement



Acknowledgement

- [Introduction](#)
- [Preliminaries](#)
- [Experiment and analysis](#)
- [Conclusion](#)
- [Submission](#)
- [Acknowledgement](#)

I would like to express my sincere gratitude to Zhang Baojie, Wang Tianjiao, and Li Leyan for their invaluable support during my learning journey. Their assistance has been instrumental in overcoming the challenges faced along the way.



Contact Information

Associate Professor Gang Li
School of Information Technology
Deakin University, Australia



GANGLI@TULIP.ORG.AU



TEAM FOR UNIVERSAL LEARNING AND INTELLIGENT PROCESSING

