HM 1 Writeup

Xiang Li

Introduction

In this brief study, our aim is to predict the quality of wines using machine learning models. We address the question of how to predict the quality rating of wines based on various chemical features. To achieve this goal, we employ two different types of machine learning models: the K Nearest Neighbors (kNN) classifier and the Gaussian Naïve Bayes classifier. We assess the performance of these models through cross-validation and accuracy on a test set.

Data

The dataset used in this analysis is the "winequality.csv" dataset. It consists of 12 rows and 1600 columns, include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality. These features, represented as numerical values, serve as the input for training the machine learning models. Our objective is to use these features to predict the quality of wines, with quality ratings being discrete levels, typically ranging from 3 to 8.

Methods

To gain an initial understanding of the current state of wine quality in the dataset, we first began with a basic statistical examination and visual representation. Then we employed two distinct machine learning models: the K Nearest Neighbors classifier and the Gaussian Naïve Bayes classifier.

The K Nearest Neighbors (kNN) is a method that classifies wines based on the similarity to their neighbors in the chemical feature space. If a wine's features are close to those of high-quality wines, it's likely to be classified as high quality. For this method, we assessed its performance on the training set through cross-validation and finalized the evaluation using a test set.

Naïve Bayes, which is a classification algorithm based on Bayes' theorem, is like a method that learns from historical data to classify wines into quality categories. It considers the probability of specific chemical features given a certain wine quality, allowing it to make predictions about the quality of new wines. For Naïve Bayes, we also used cross-validation and test set accuracy to evaluate its performance.

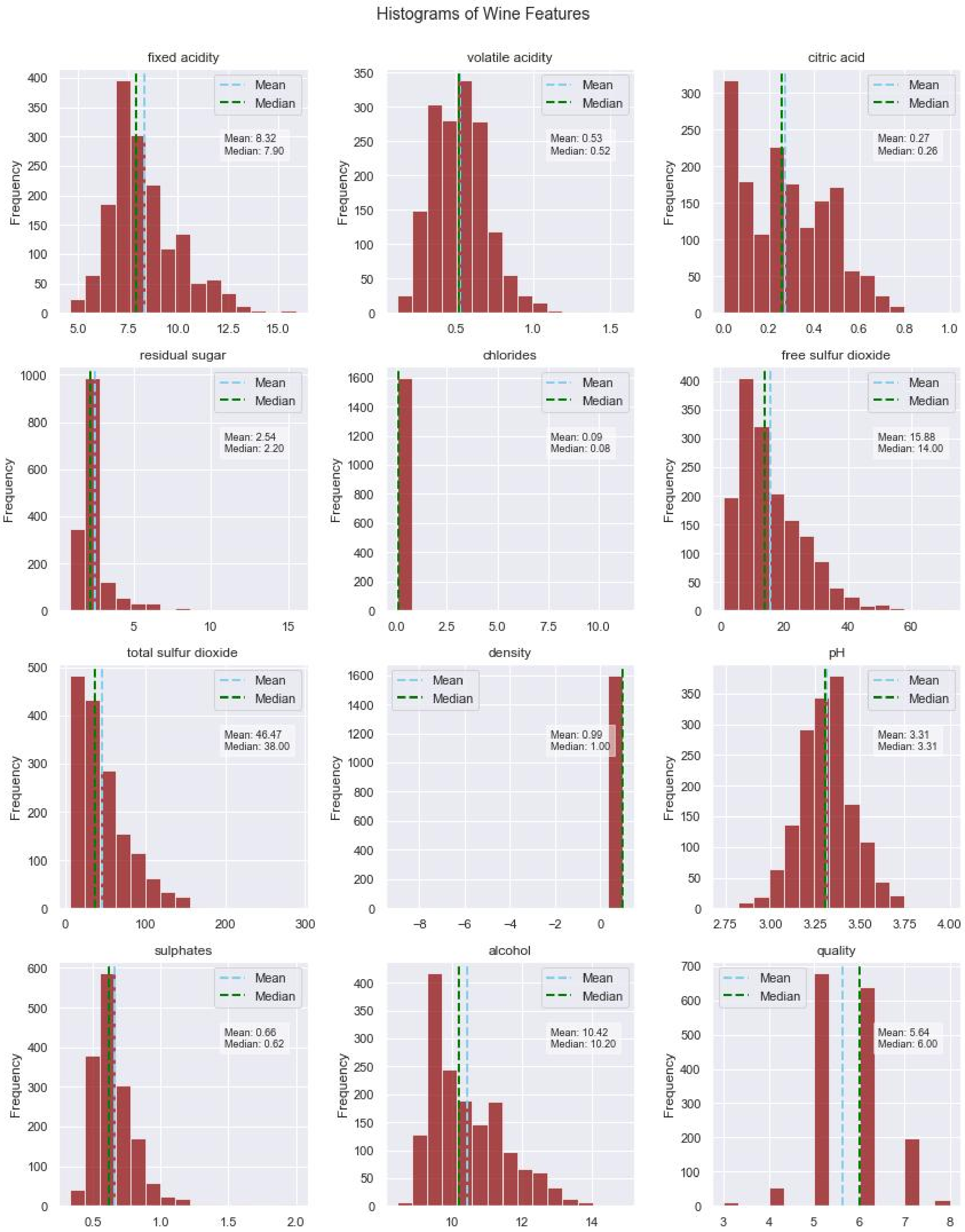
Prior to training, we first began with a basic statistical examination and visual representation (see Appendix A and B for details), and we performed data preprocessing, including standardization and handling missing values.

Results

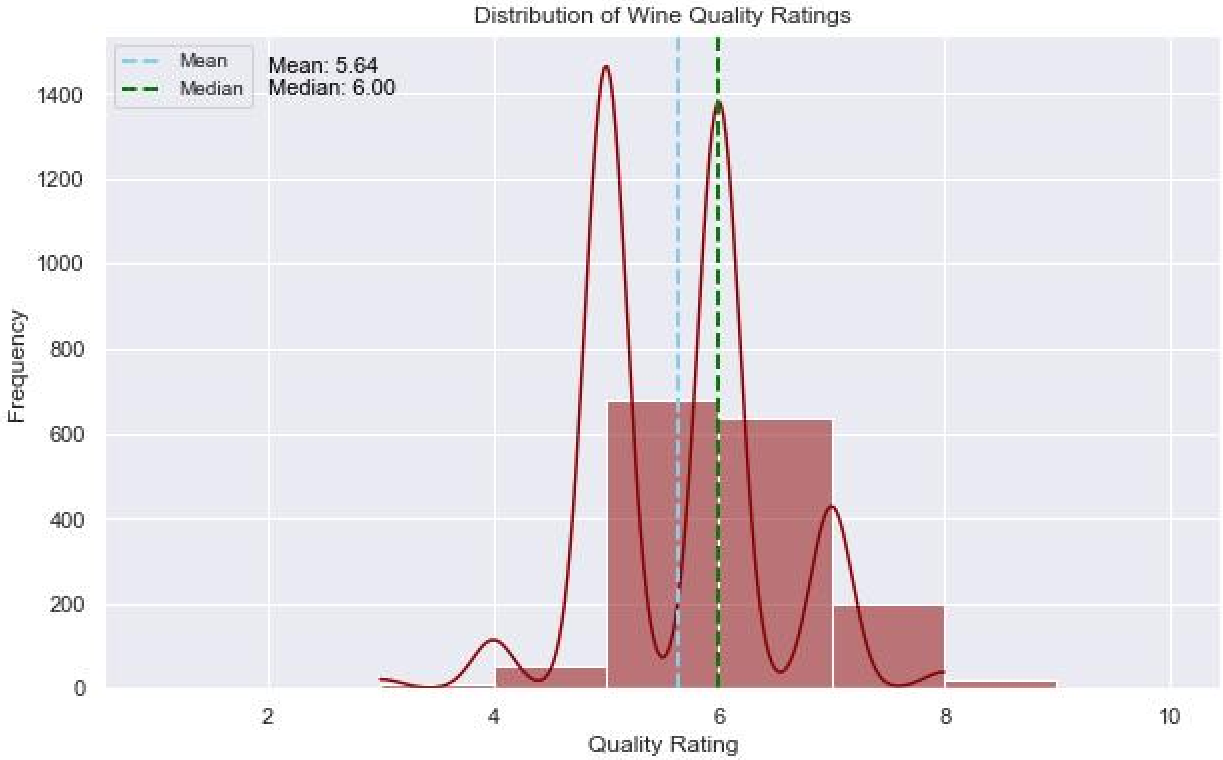
Preliminary results indicate that alcohol content and volatile acidity may play crucial roles in determining wine quality. For K Nearest Neighbors Classifier, the Cross-validation mean accuracy is approximately 55.36% and the Test set accuracy is approximately 45.63%. For Gaussian Naïve Bayes Classifier, the Cross-validation mean accuracy is approximately 40.42%, and the Test set accuracy is approximately 40.31%.

The results suggest that the K Nearest Neighbors model demonstrated better performance in both average accuracy and test set accuracy. However, accuracy alone is not the sole evaluation criterion, and consideration of other performance metrics, especially in dealing with imbalanced classes, is recommended.

Appendix A


Histograms of Wine Features
fixed acidity
Frequency
Mean
Median
Mean: 8.32
Median: 7.90
0
50
100
150
200
250
300
350
400
5.0
7.5
10.0
12.5
15.0
volatile acidity
Mean: 0.53
Median: 0.52
0
50
100
150
200
250
300
350
0.5
1.0
1.5
citric acid
Mean: 0.27
Median: 0.26
0
50
100
150
200
250
300
0.0
0.2
0.4
0.6
0.8
1.0
residual sugar
Mean: 2.54
Median: 2.20
0
200
400
600
800
1000
5
10
15
chlorides
Mean: 0.09
Median: 0.08
0
200
400
600
800
1000
1200
1400
1600
0.0
2.5
5.0
7.5
10.0
free sulfur dioxide
Mean: 15.88
Median: 14.00
0
50
100
150
200
250
300
350
400
0
20
40
60
total sulfur dioxide
Mean: 46.47
Median: 38.00
0
100
200
300
400
500
0
100
200
300
density
Mean: 0.99
Median: 1.00
0
200
400
600
800
1000
1200
1400
1600
−8
−6
−4
−2
0
pH
Mean: 3.31
Median: 3.31
0
50
100
150
200
250
300
350
2.75
3.00
3.25
3.50
3.75
4.00
sulphates
Mean: 0.66
Median: 0.62
0
100
200
300
400
500
600
0.5
1.0
1.5
2.0
alcohol
Mean: 10.42
Median: 10.20
0
50
100
150
200
250
300
350
400
10
12
14
quality
Mean: 5.64
Median: 6.00
0
100
200
300
400
500
600
700
3
4
5
6
7
8

Appendix B


Distribution of Wine Quality Ratings
Mean
Median
Mean: 5.64
Median: 6.00
Frequency
0
200
400
600
800
1000
1200
1400
Quality Rating
2
4
6
8
10