

Project Overview

Project Title

Red Wine Quality Analysis

GitHub Repository Link

https://github.com/dhakaadi/cmse802_project.git

Brief project description

This project involves analyzing the quality of red wines using a variety of chemical features. Data visualization, correlation analysis, and machine learning models are used to explore relationships between wine quality and its chemical composition, followed by predicting wine ratings using a classifier.

Project Setup

Description of repository structure

The repository is organized into directories for data, Jupyter notebook and results. The main file for analysis is the Jupyter notebook, which is used to run the entire analysis.

Explanation of key files and directories

- Data: Contains the wine quality dataset (winequality-red.csv)
- Jupyter notebook: Contains Python code for data analysis and model building
- Results: Stores visualizations, charts, and model performance metrics

List of dependencies and setup instructions

The following dependencies are required:

- numpy
- pandas
- matplotlib
- seaborn
- scikit-learn

To set up the environment, use the following commands:

```
`pip install -r requirements.txt`
```

Completed Tasks

List of tasks completed from the homework

1. Data exploration and visualization
2. Bivariate Analysis and Correlation analysis
3. Feature importance extraction
4. Data pre-processing
5. Started model building and evaluation

Justification for each task's relevance to your project

These tasks help in understanding the dataset, identifying key relationships between features, and building a model to predict wine ratings.

Brief description of the process for each task

Data exploration: Initial understanding of distributions and counts using histograms and box plots.

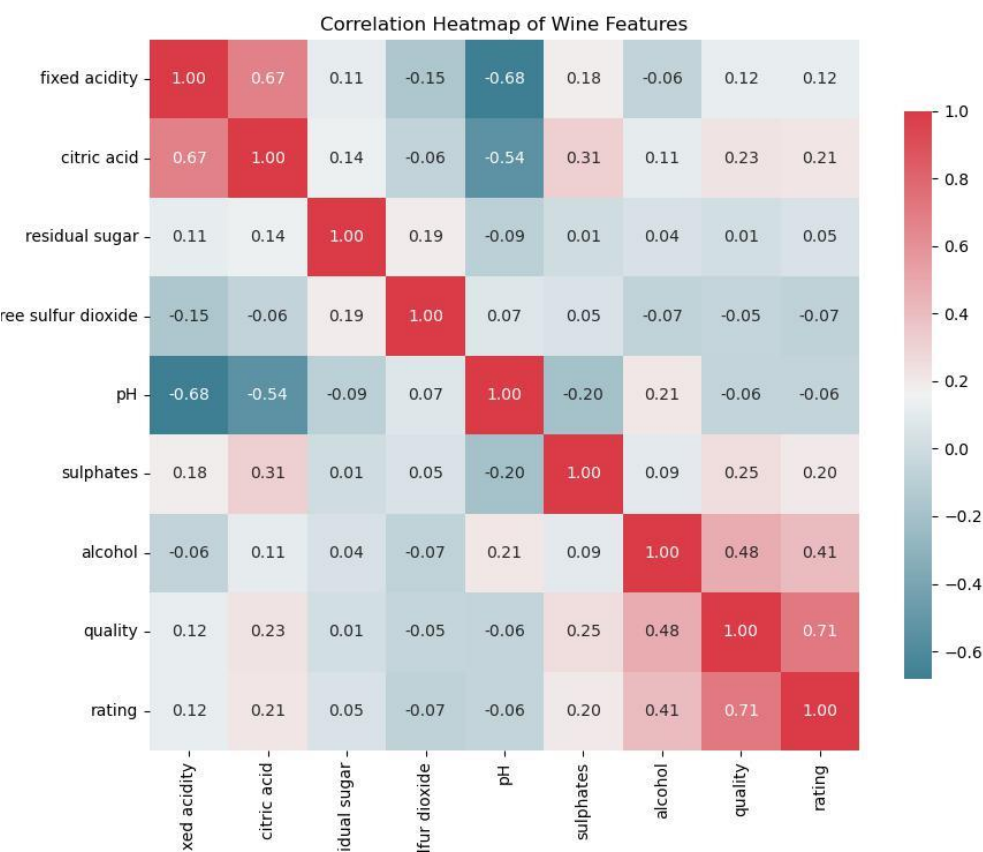
Correlation analysis: Generated a heatmap to see how features are correlated with wine quality.

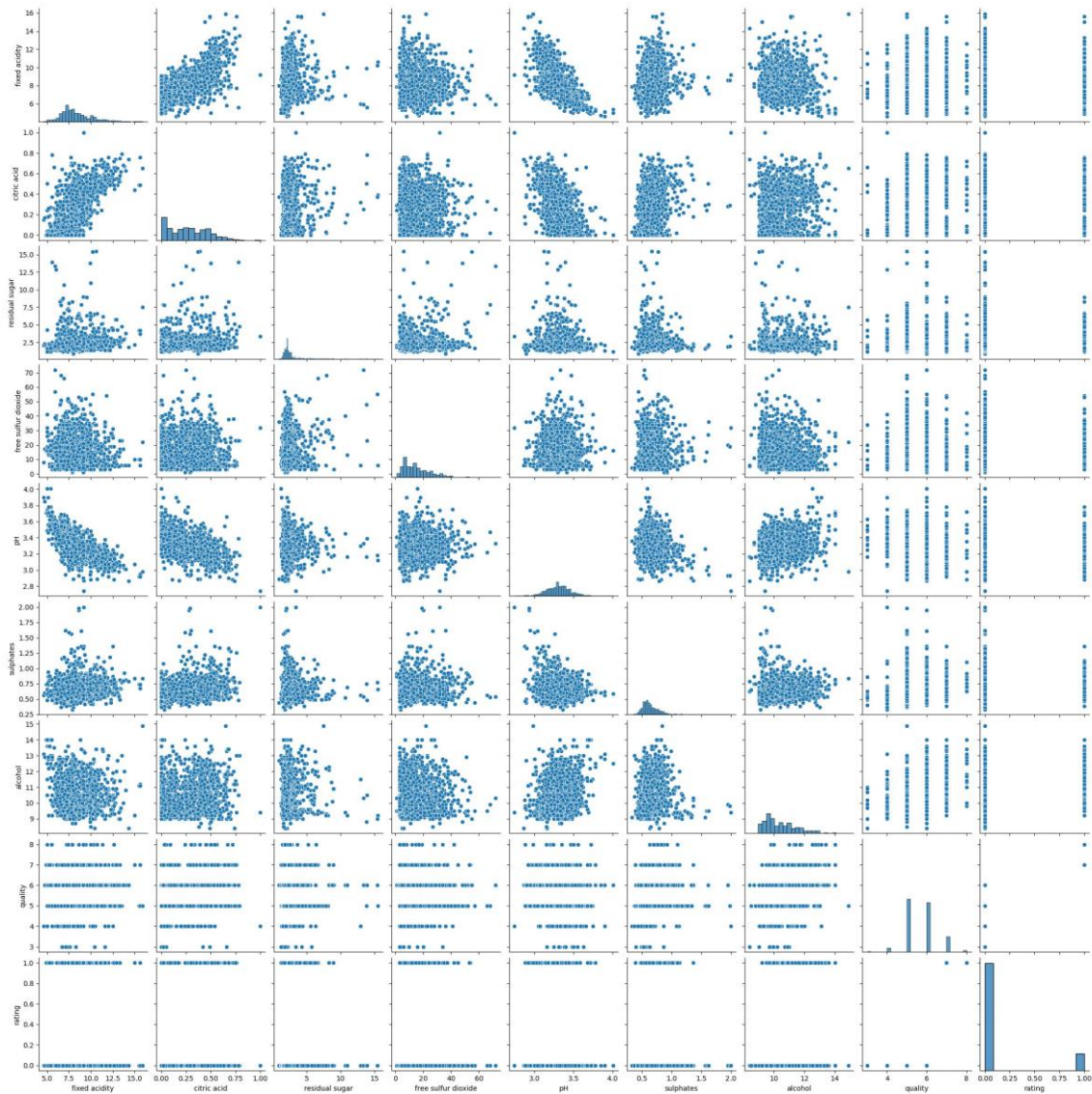
Feature importance: Used ExtraTreesClassifier to identify which features contribute most to wine ratings.

Model building: Logistic regression and Decision Tree model built to predict wine quality ratings.

Initial Analysis and Findings

Summary of key findings from your initial analysis

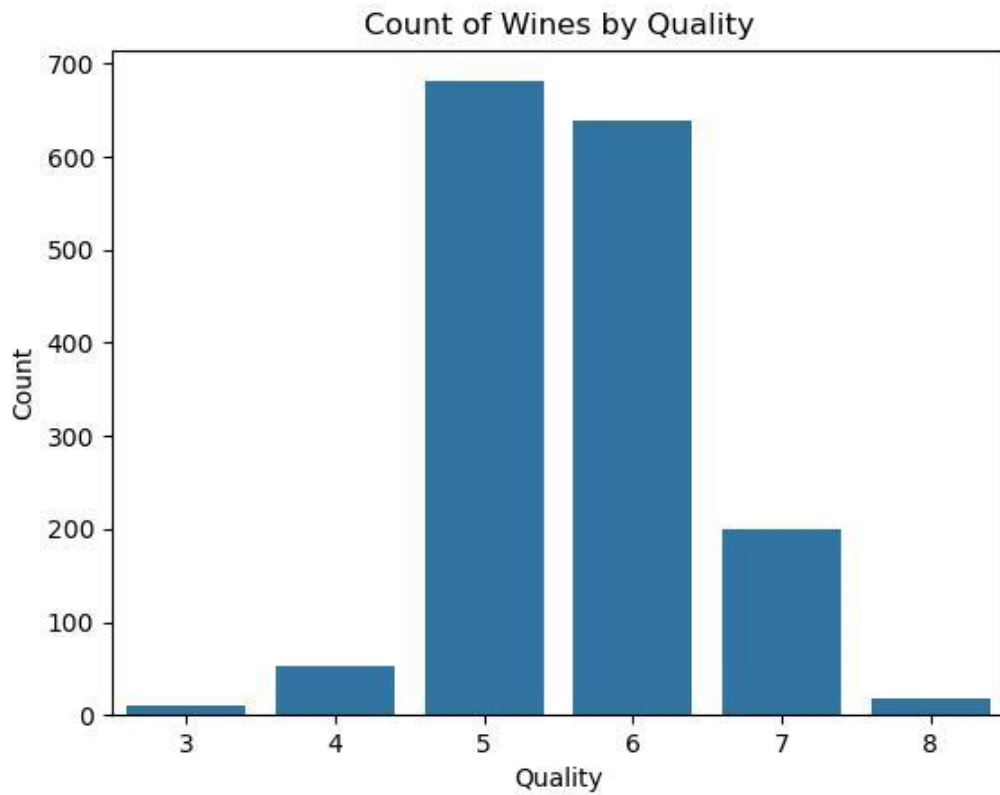




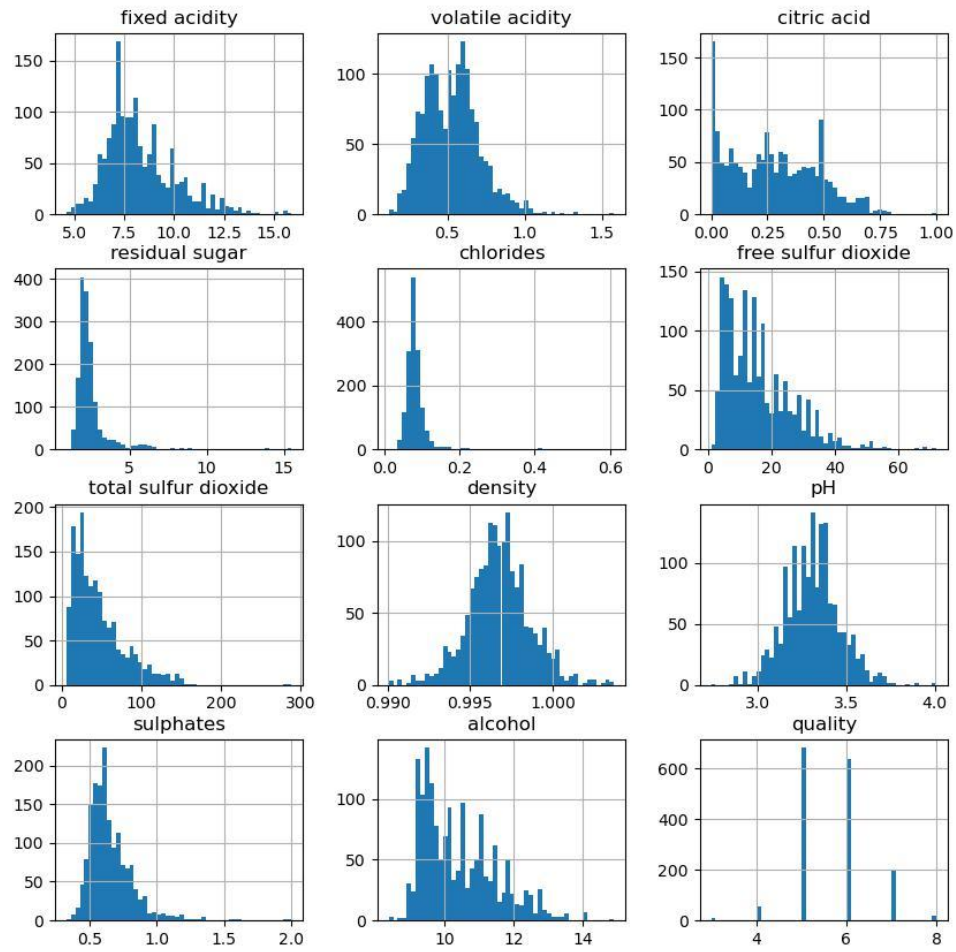
- Alcohol, sulphates, and citric acid levels show significant variations across wine ratings. These features are likely to be important for predicting wine quality.
- Strongly correlated items are:
 1. fixed acidity and citric acid.
 2. free sulphur dioxide and total sulphur dioxide.
 3. fixed acidity and density.
 4. alcohol and quality
- Volatile acidity, total sulfur dioxide, chlorides, density are very less related to the dependent variable thus dropped

Relevant data visualizations

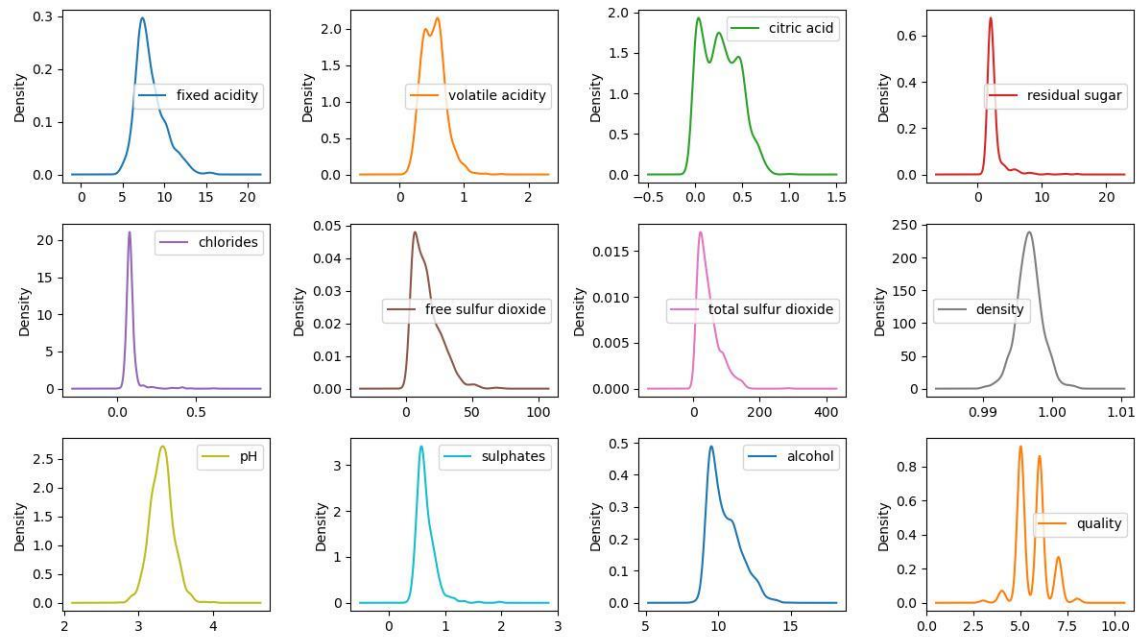
Various plots such as box plots, violin plots, and heatmaps were created to explore data patterns.



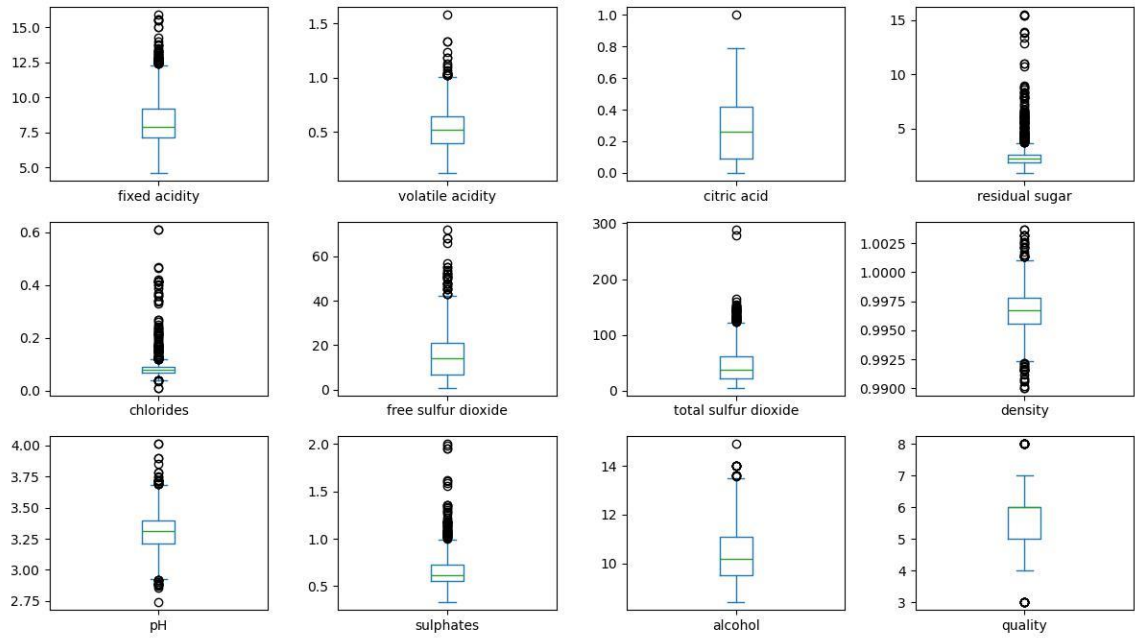
Histograms for Each Feature in the Dataset

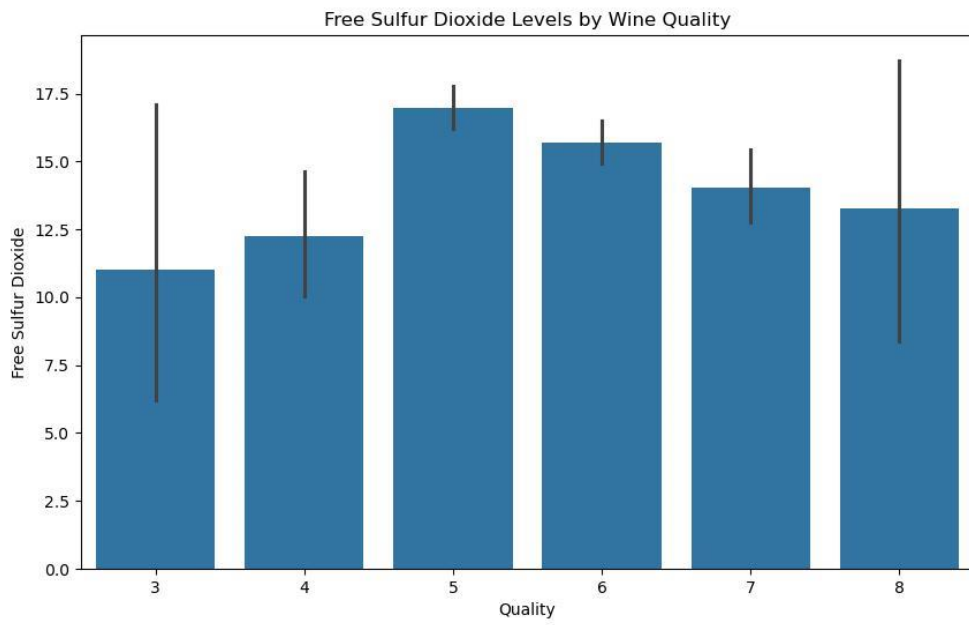
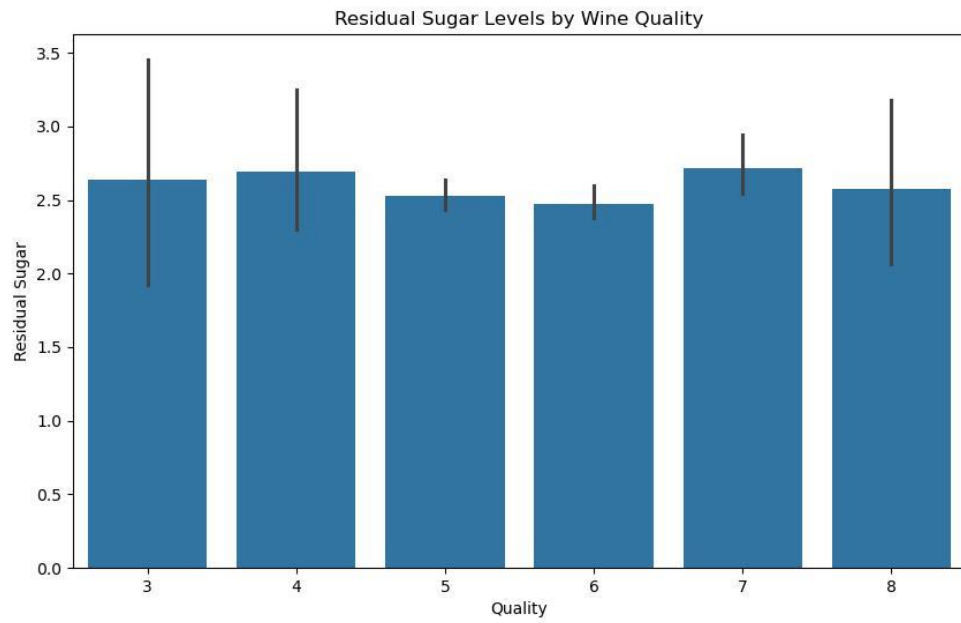


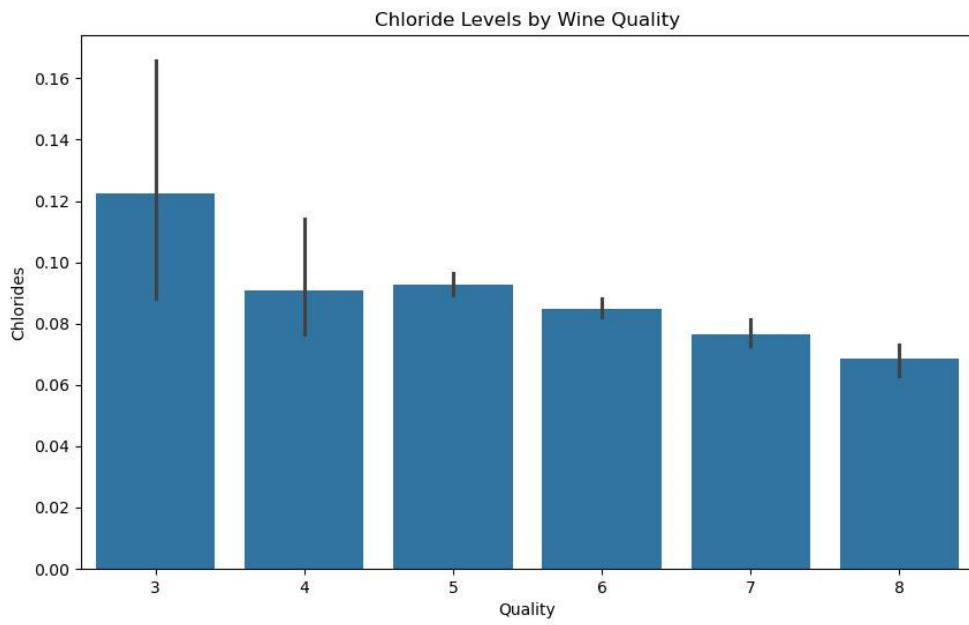
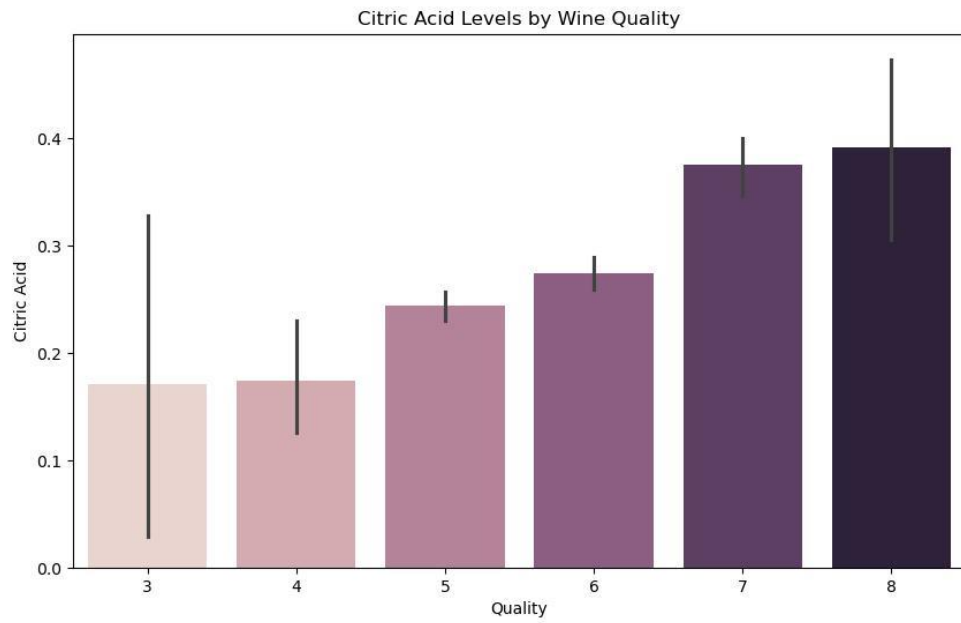
Density Plots for Each Feature in the Wine Dataset

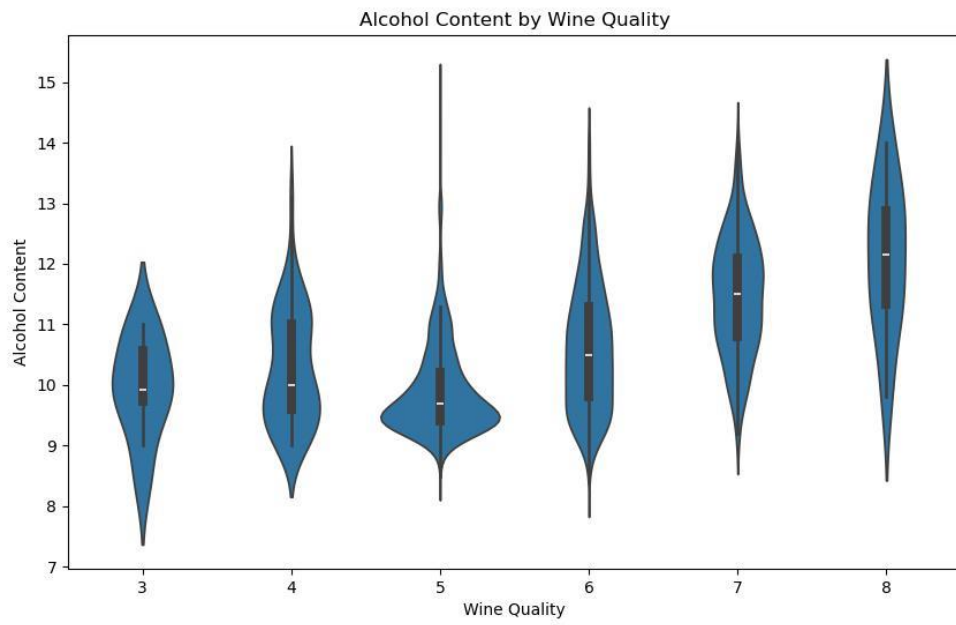
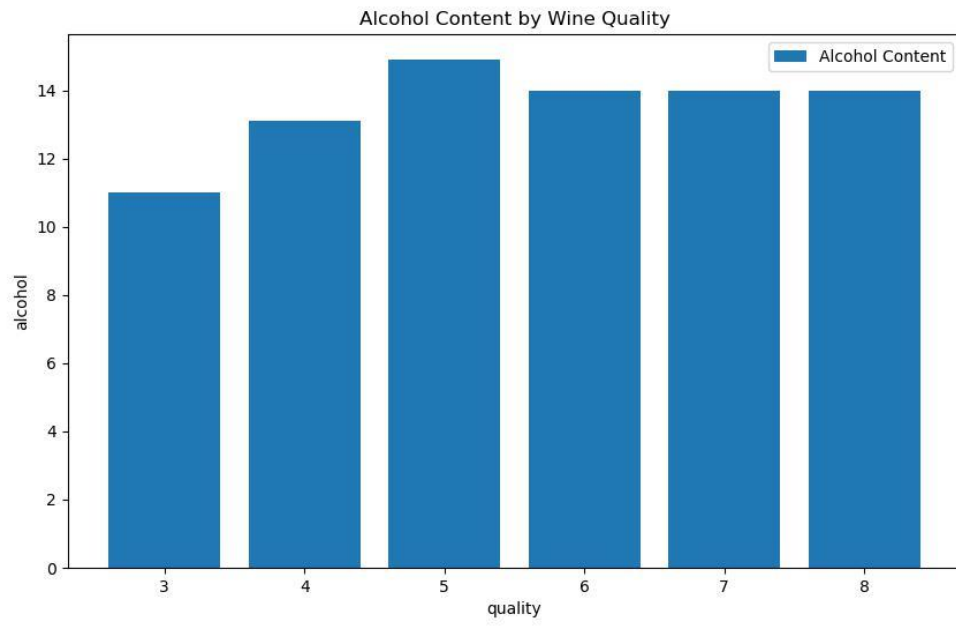


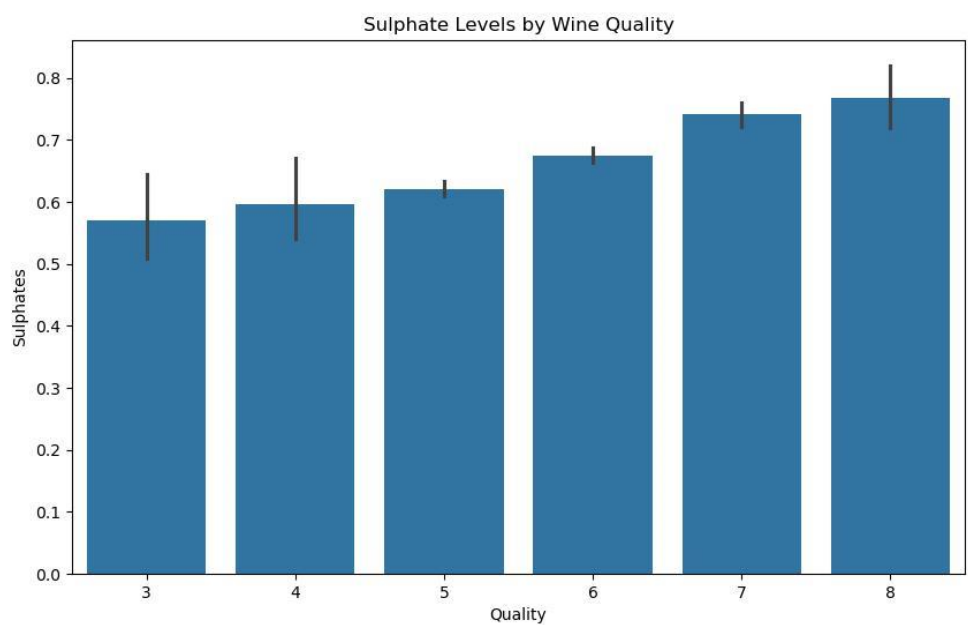
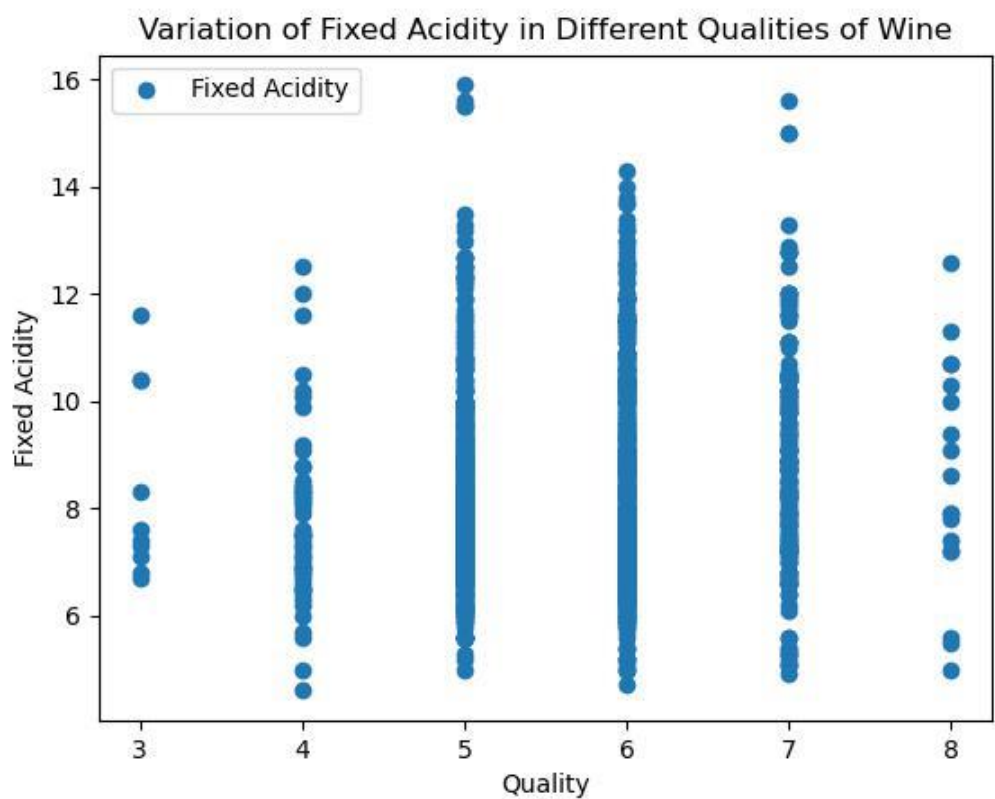
Box Plots for Each Feature in the Dataset



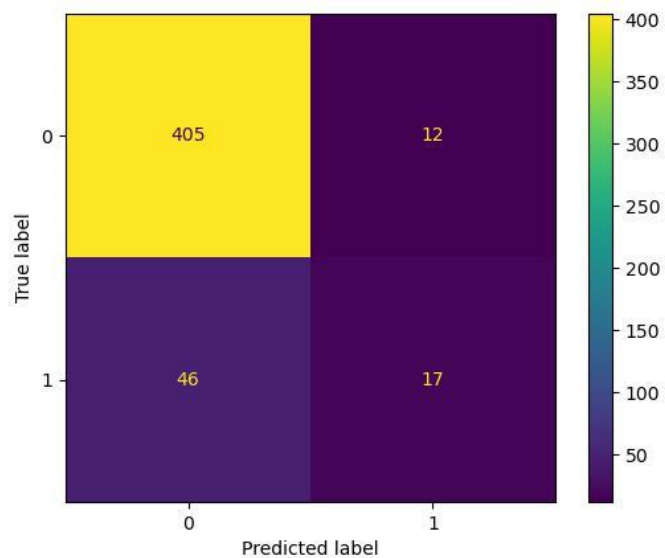




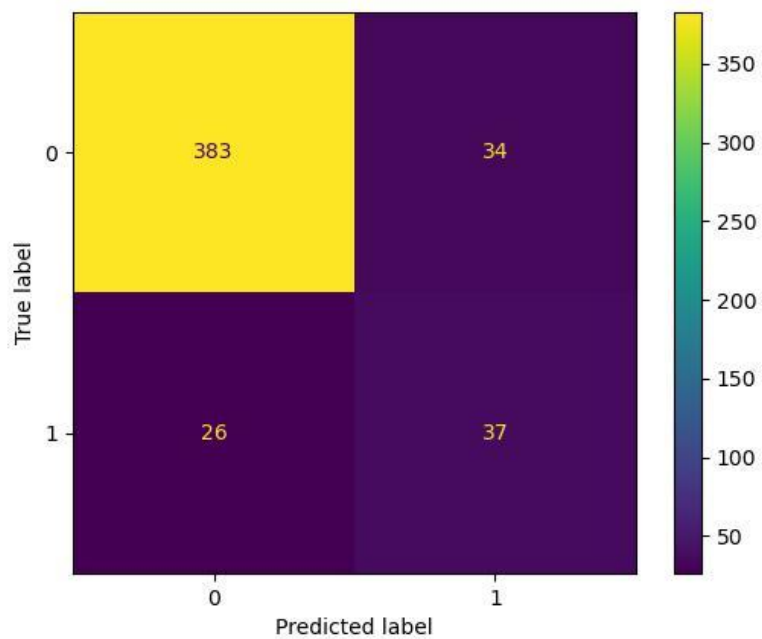




Confusion Matrix for Logistic Regression:



Confusion Matrix for Decision Tree:



Proposed Approach

Description of your proposed machine learning approach

The project employs **supervised learning techniques**, primarily

- Logistic Regression (Done)
- Decision Tree (Done)
- Random Forest
- k-nearest neighbors (k-NN)
- Support Vector Machine
- GaussianNB
- Xgboost
- Multi Layer Perceptron
- Artificial Neural Networks

to classify wines as Good or Bad based on their chemical features.

Justification for chosen methods

Logistic Regression (Completed):

Justification: Logistic Regression is simple, interpretable, and works well for binary classification problems. It helps understand the impact of each feature on wine quality prediction.

Process: The Logistic Regression model was trained using the wine dataset, and an accuracy of around 70% was achieved. The model was evaluated using confusion matrices and classification reports.

Decision Tree (Completed):

Justification: Decision Trees provide a visual, easily interpretable way to understand decision-making processes for classification tasks. They are useful for capturing non-linear relationships in the data.

Process: The Decision Tree classifier was trained on the dataset, splitting features to determine the most important ones for predicting wine quality. Feature importance scores were also extracted.

Random Forest:

Justification: Random Forest is an ensemble learning method that builds multiple decision trees and aggregates their results, providing higher accuracy and robustness to overfitting compared to a single decision tree.

Process: The Random Forest classifier will be trained to predict wine quality. This model will also output feature importance scores, enhancing interpretability.

k-nearest neighbors (k-NN):

Justification: k-NN is a simple, instance-based algorithm that does not make assumptions about data distribution. It can work well in cases where the decision boundary is complex.

Process: The k-NN model will use a specified number of nearest neighbors to classify wines based on their chemical properties. Accuracy and error rates will be computed for evaluation.

Support Vector Machine (SVM):

Justification: SVM is powerful for high-dimensional spaces and cases where a clear margin of separation between classes is required. It is particularly effective for small-to-medium-sized datasets.

Process: The SVM classifier will attempt to find the optimal hyperplane to classify wines as good or bad based on the features. Tuning of kernel functions will be part of the process.

GaussianNB:

Justification: Gaussian Naive Bayes is efficient and effective for data that follows a normal distribution. It can work well for simple, fast classification when feature independence is assumed.

Process: The GaussianNB classifier will be trained using the assumption of Gaussian (normal) distribution for the feature data. Performance metrics will be analyzed to check its effectiveness.

Xgboost:

Justification: Xgboost is a powerful, gradient boosting-based ensemble method that can handle missing data and is known for its performance in machine learning competitions.

Process: Xgboost will be applied to the wine dataset, with hyperparameter tuning to maximize predictive accuracy. Feature importance will be visualized to see which features contribute most to quality prediction.

Multi-Layer Perceptron (MLP):

Justification: MLP is a type of feedforward neural network that can model complex relationships between features, capturing non-linear interactions in the data.

Process: An MLP model will be trained on the dataset to predict wine ratings.
Hyperparameter tuning (number of layers, neurons, etc.) will be done to optimize performance.

Artificial Neural Networks (ANN):

Justification: ANNs are highly flexible and can capture complex, non-linear relationships. They are useful for larger datasets and complex prediction tasks like wine quality classification.

Process: A deep learning model will be designed using a neural network with multiple hidden layers. Various activation functions and optimizers will be tested to improve the model's accuracy and minimize error

Preliminary Results (if any)

Description of any initial experiments or model testing

Rating:

Bad (quality > 7) 1382

Good (quality < 7) 217

fixed acidi ty	citric acid	residu al sugar	free sulfur dioxid e	pH	sulphat es	alcoho l	quality	
0	8.2368	0.2544	2.5121	16.1722	3.31461	0.6447	10.2510	5.4088
	31	07	20	14	6	54	37	28
1	8.8470	0.3764	2.7087	13.9815	3.28880	0.7434	11.5180	7.0829
	05	98	56	67	2	56	49	49

Initial tests with Logistic Regression achieved an accuracy of approximately 87.9%.

Accuracy Score: 0.8791666666666667

Confusion Matrix:

[[405 12]

[46 17]]

Classification Report:

	precision	recall	f1-score	support
0	0.90	0.97	0.93	417
1	0.59	0.27	0.37	63
accuracy		0.88		480
macro avg	0.74	0.62	0.65	480
weighted avg	0.86	0.88	0.86	480

Initial tests with Decision Tree achieved an accuracy of approximately 87.5%.

Accuracy Score: 0.875

Confusion Matrix:

[[383 34]

[26 37]]

Classification Report:

	precision	recall	f1-score	support
0	0.94	0.92	0.93	417
1	0.52	0.59	0.55	63

accuracy		0.88	480	
macro avg	0.73	0.75	0.74	480
weighted avg	0.88	0.88	0.88	480

Challenges and Solutions

Description of significant challenges encountered

Handling imbalanced data, tuning hyperparameters for the Logistic Regression and Decision Tree models.

Explanation of the solutions implemented

Balanced the dataset by adjusting class weights in the Logistic Regression and Decision Tree models.

Next Steps

Outline of upcoming tasks and milestones

- Hyperparameter tuning for improved model performance.
- Test different classification algorithms:
 - Logistic Regression (Done)
 - Decision Tree (Done)
 - Random Forest
 - k-nearest neighbors (k-NN)
 - Support Vector Machine
 - GaussianNB
 - Xgboost
 - Multi Layer Perceptron
 - Artificial Neural Networks
- Perform cross-validation to ensure robustness of the model.

Conclusion

The project is progressing well, with initial models showing decent accuracy. Moving forward, the focus will be on improving model performance and exploring different algorithms.

References

<https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv>