Red Wine Quality Analysis

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

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.

Description of repository structure

```
cmse802_project/
├── Data/
│   ├── winequality-red.csv
├── Notebook/
│   ├── cmse802_project.ipynb
├── Result/
│   ├── Figure/
│   │   ├──Alcohol Content by Wine Quality.jpg
│   │   ├──Box Plots for Each Feature in the Dataset.jpg
│   │   ├──Chloride Levels by Wine Quality.jpg
│   │   ├──Citric Acid Levels by Wine Quality.jpg
│   │   ├──Confusion Matrix for Decision Tree.jpg
│   │   ├──Confusion Matrix for Logistic Regression.jpg
│   │   ├──Correlation Heatmap of Wine Features.jpg
│   │   ├──Count of Wines by Quality.jpg
│   │   ├──Density Plots for Each Feature in the Wine Dataset.jpg
│   │   ├──Free Sulfur Dioxide Levels by Wine Quality.jpg
```

```
│    │    ├───Histograms for Each Feature in the Dataset.jpg
│    │    ├───Residual Sugar Levels by Wine Quality.jpg
│    │    ├───Sulphate Levels by Wine Quality.jpg
│    │    ├───Variation of Fixed Acidity in Different Qualities of Wine.jpg
│    │    ├───pairplot.jpg
│    │    ├───violinplot Alcohol Content by Wine Quality.jpg

├─── Report/
│    ├─── Wine_Quality_Project_Report.pdf
├───. ignore
├─── README.md
└─── requirements.txt
```

Explanation of key files and directories

- **Data**: Contains the wine quality dataset (winequality-red.csv)

- **Notebook**: Contains Python code for data analysis and model building

- **Results/Figures**: Stores visualizations, charts, and model performance metrics

- **Report**: Contains .pdf report file.

- **README.md** contains the repository name, brief description and some basic instructions for setting up and running your code with requirements.

- **requirements.txt** contains required libraries

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

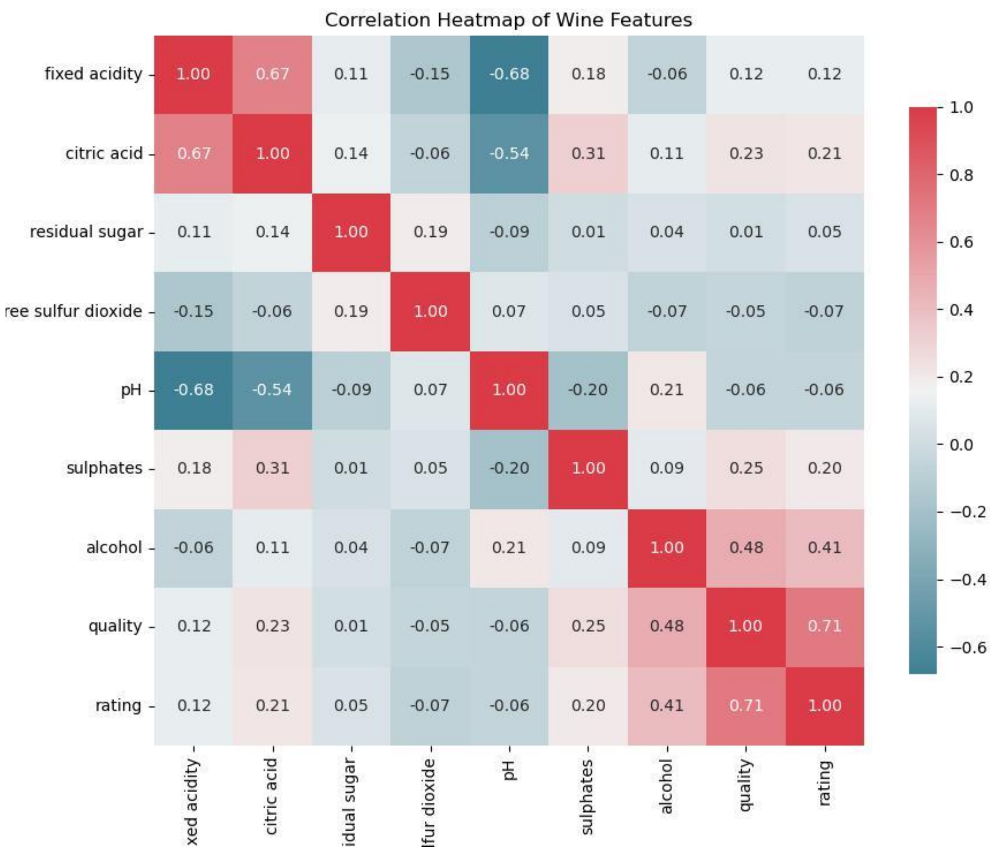
These tasks help in understanding the dataset, identifying key relationships between features, and building a model to predict wine ratings.

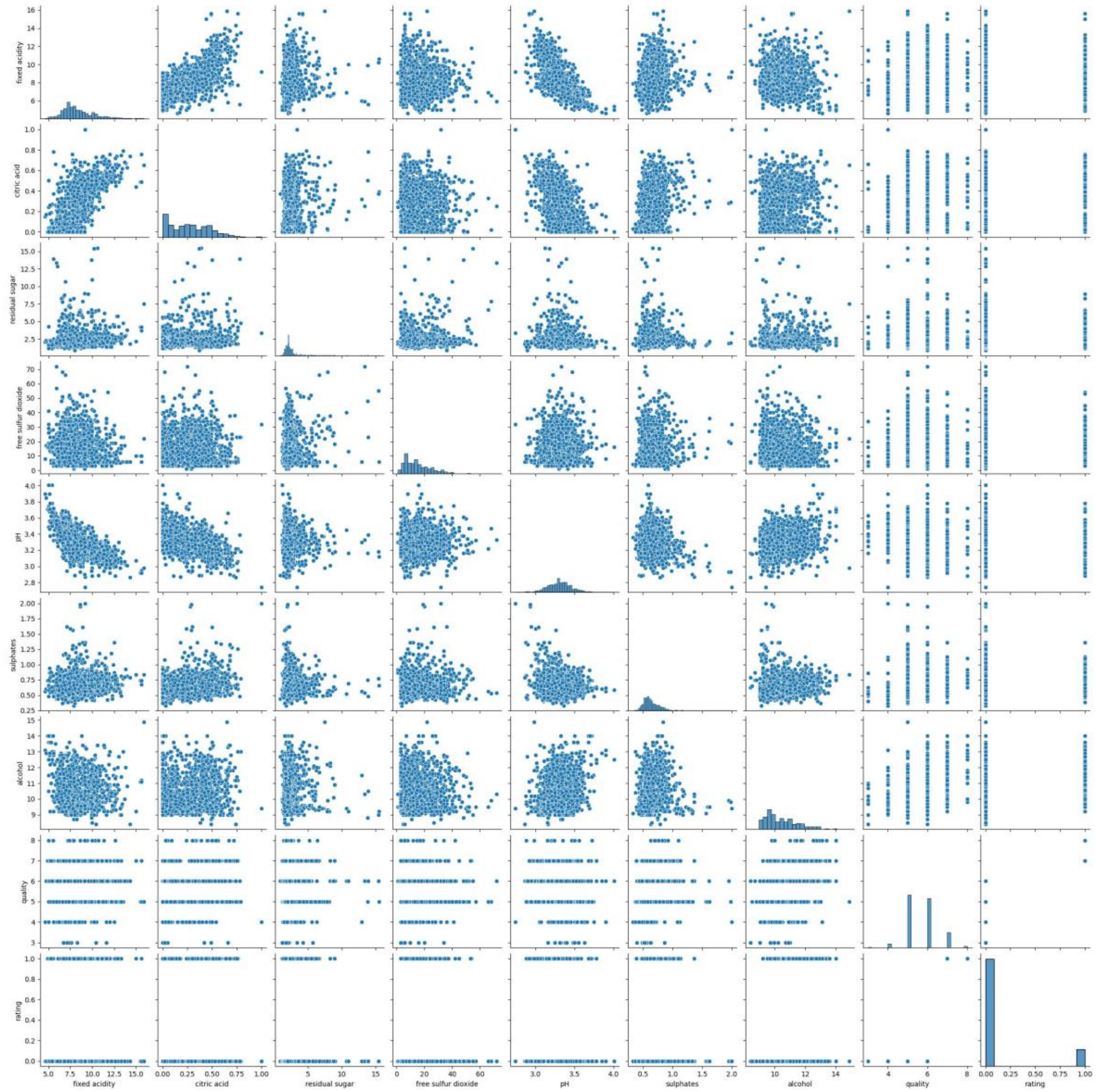
Brief description of the process for each task

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

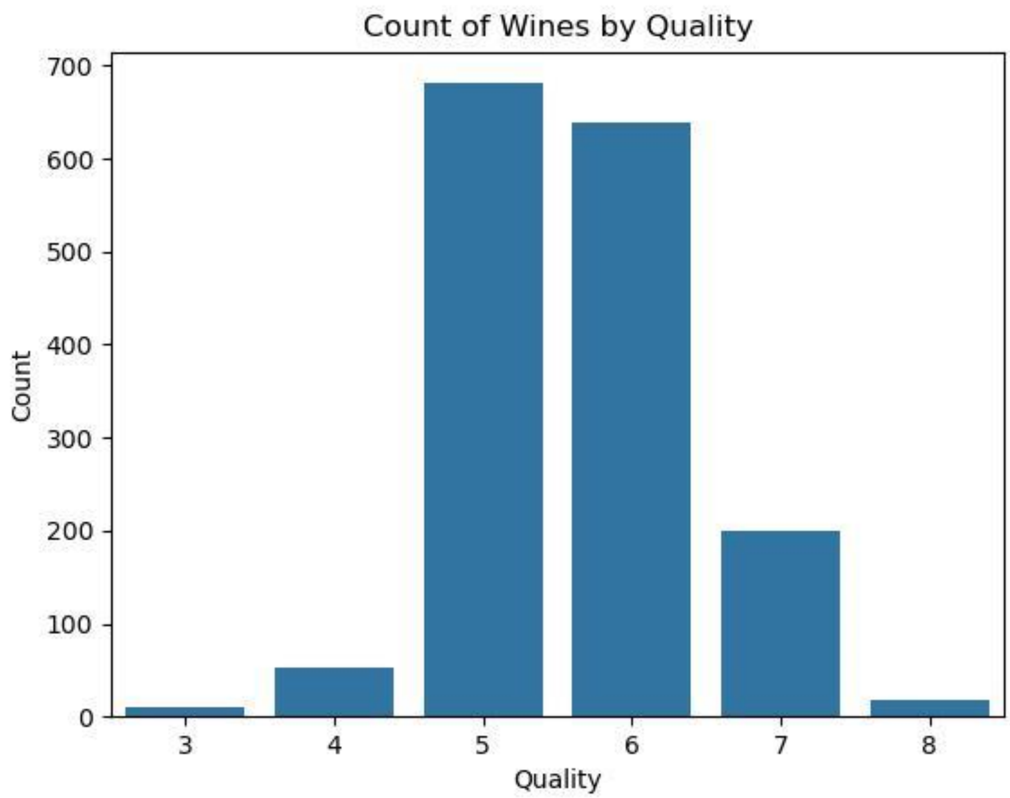




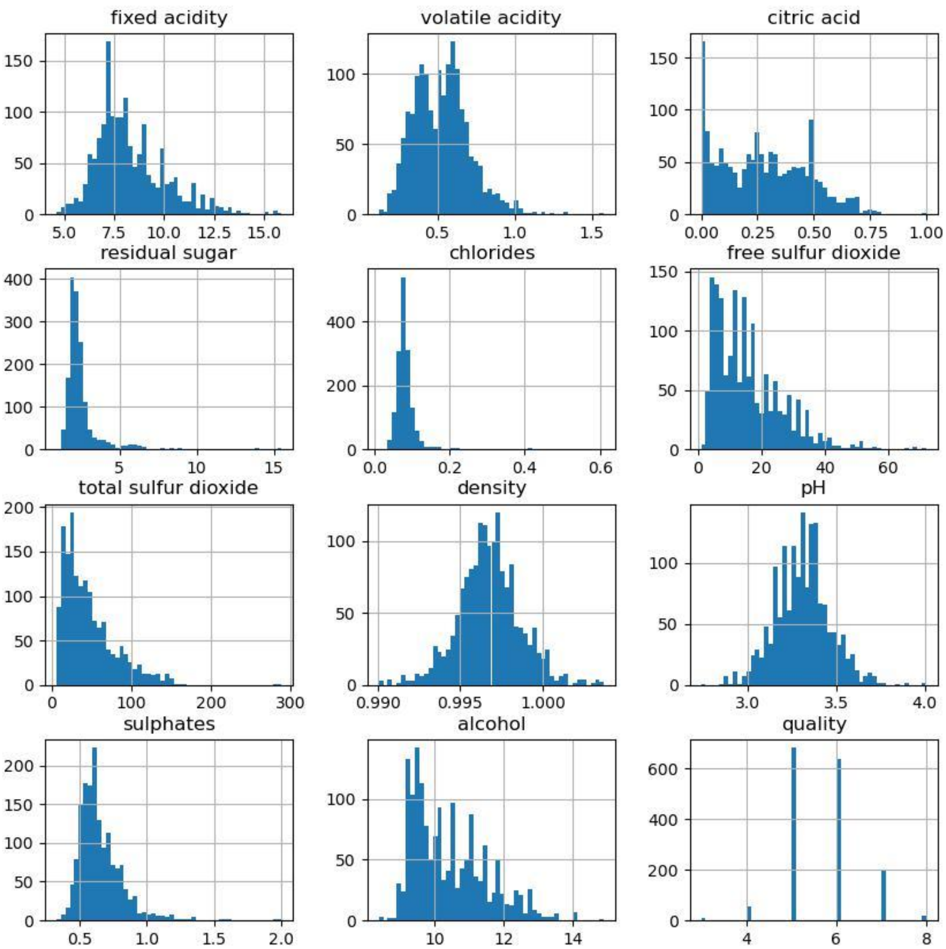
- Alcohol, sulphates, and citric acid levels show significant variations across wine ratings. These features are likely to be important for predicting wine quality.
- Strongly correlated items are:
    1. fixed acidity and citric acid.
    2. free sulphur dioxide and total sulphor dioxide.
    3. fixed 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

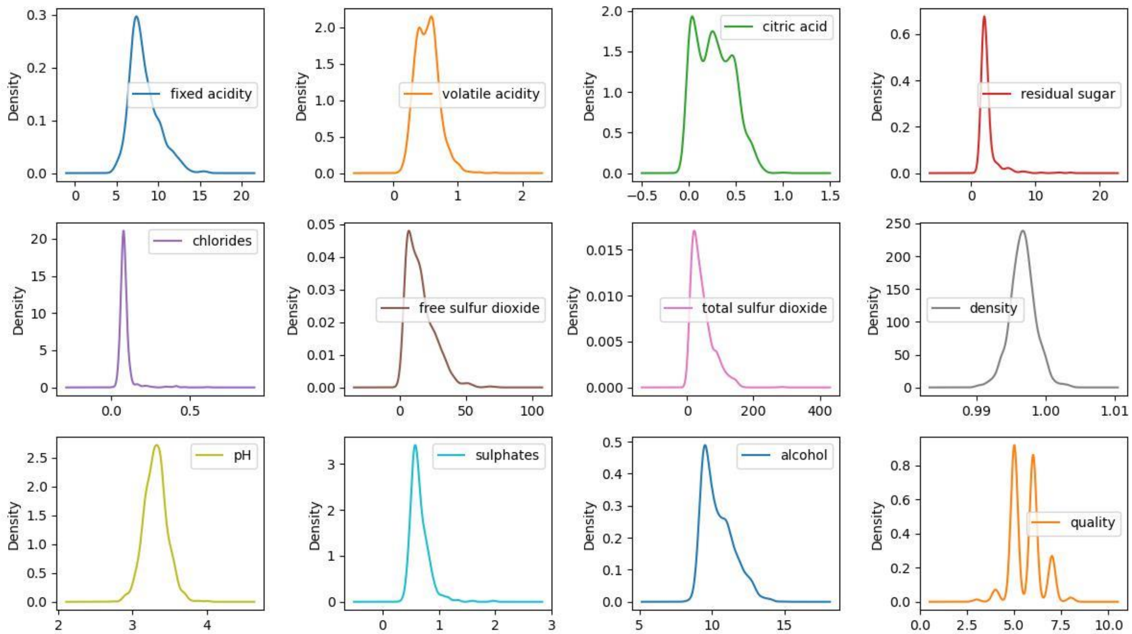
Various plots such as box plots, violin plots, and heatmaps were created to explore data patterns.



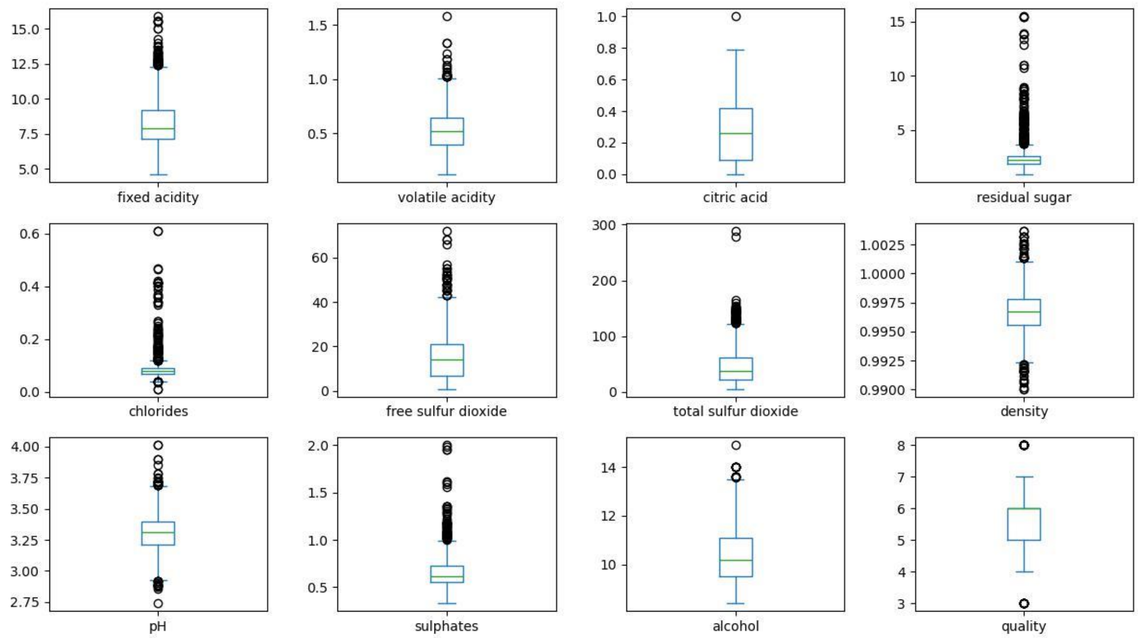
Histograms for Each Feature in the Dataset

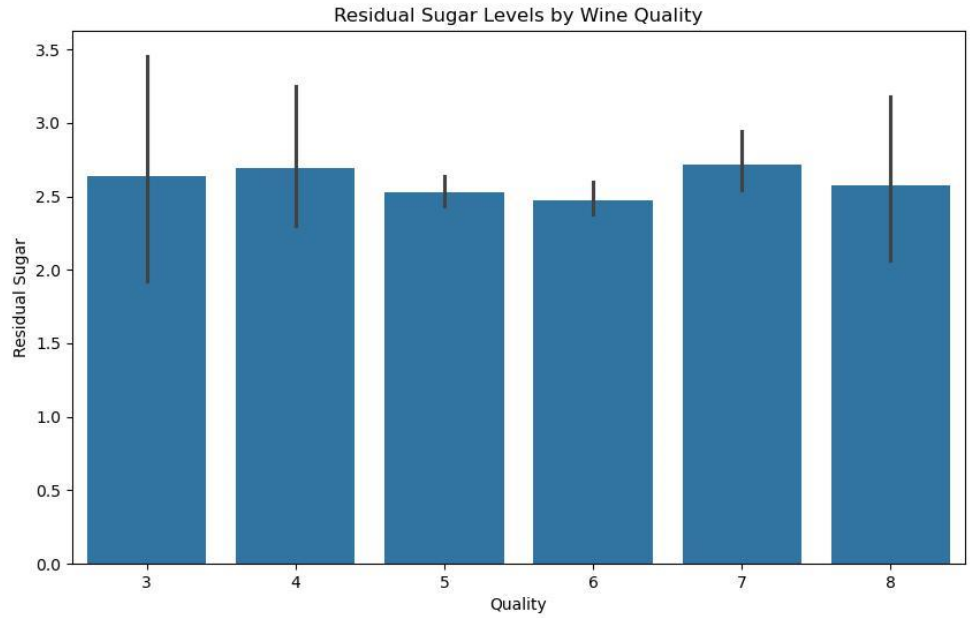


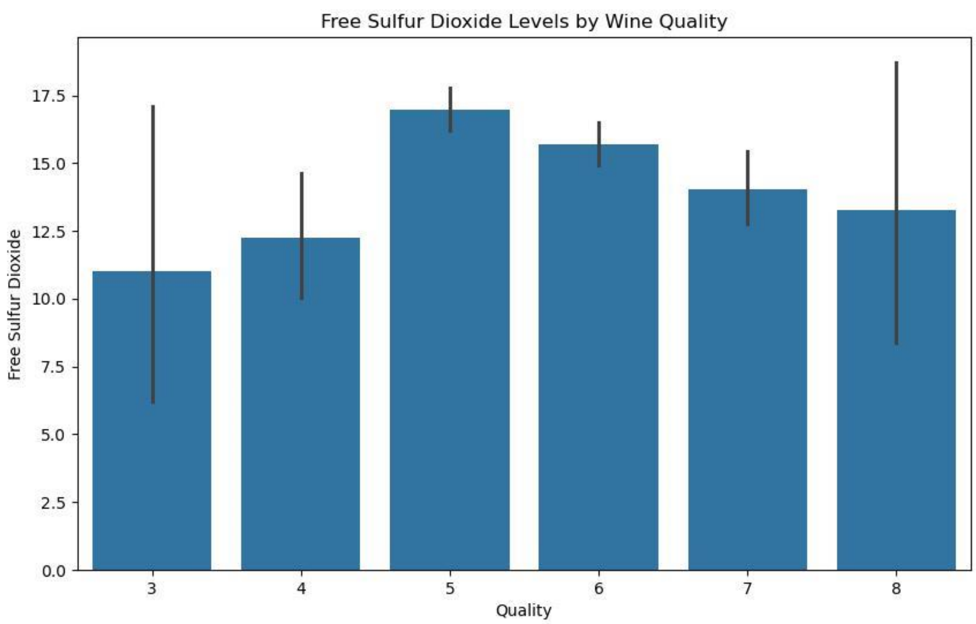
Density Plots for Each Feature in the Wine Dataset

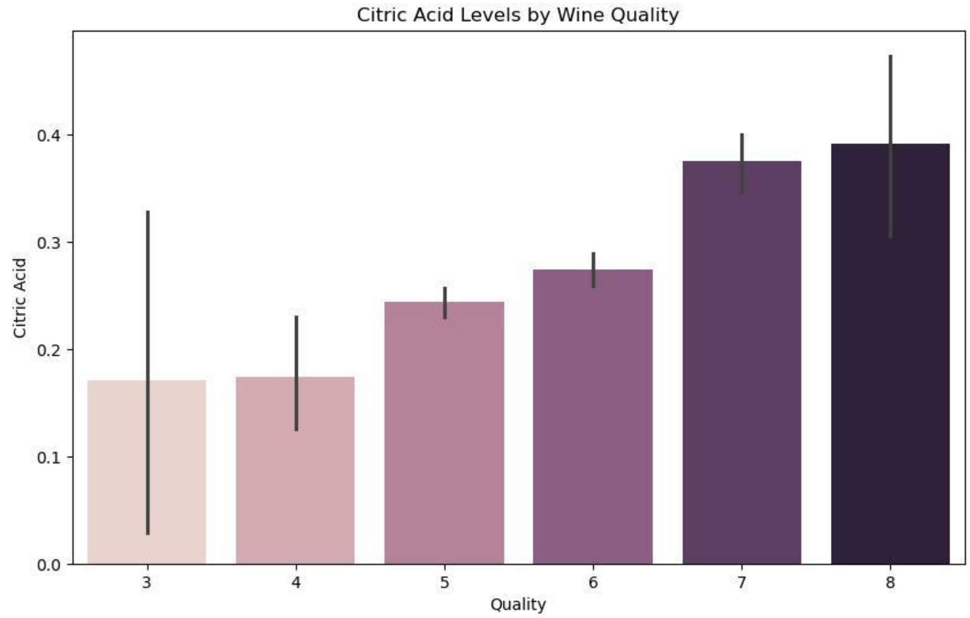


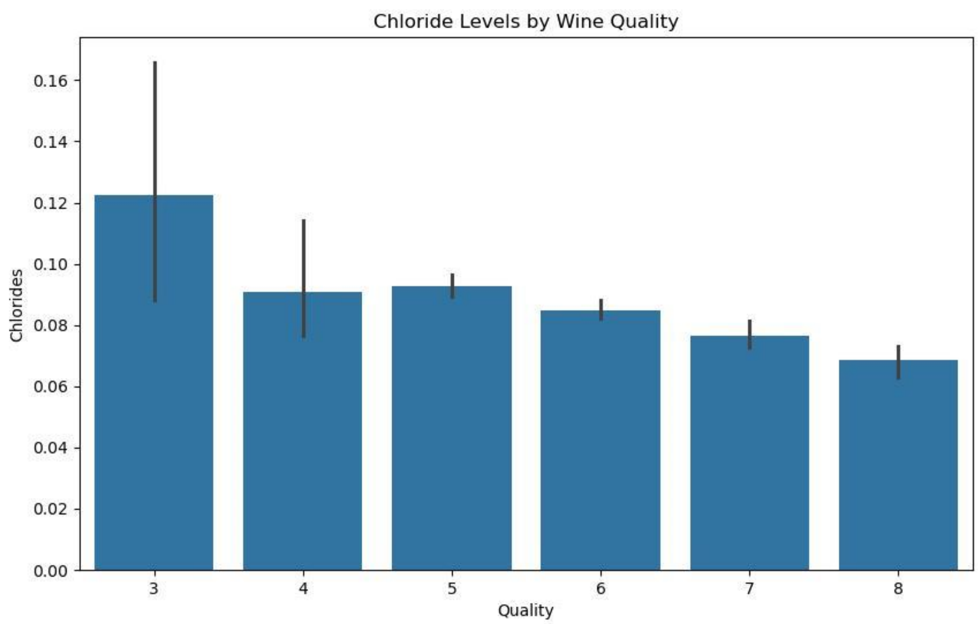
Box Plots for Each Feature in the Dataset

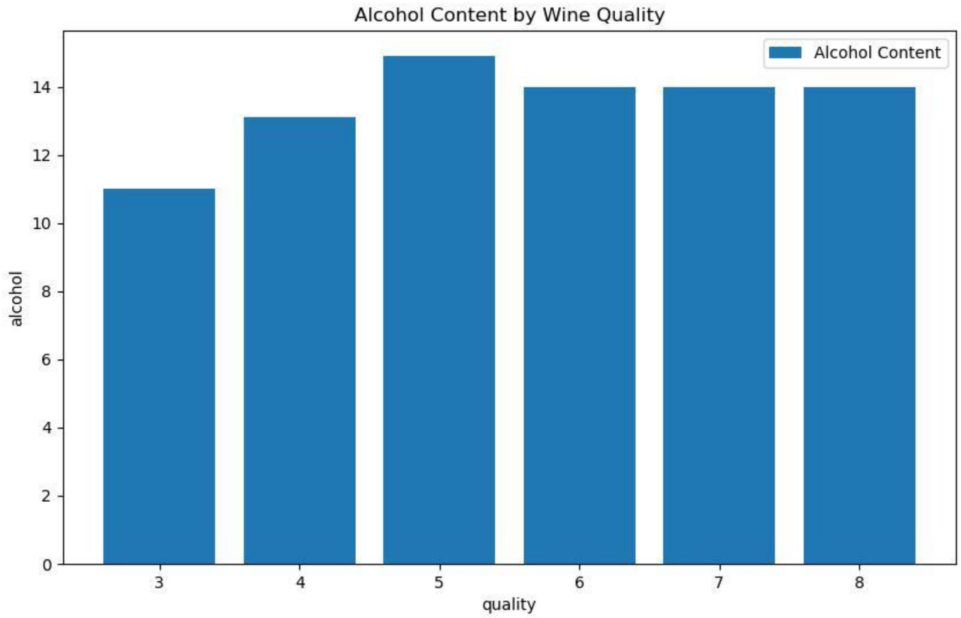


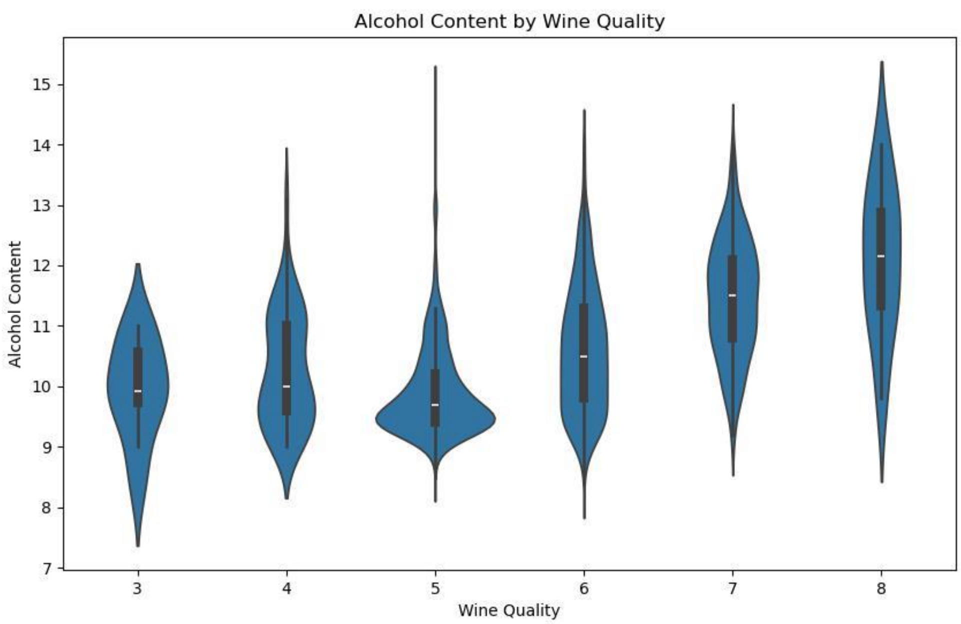

Residual Sugar Levels by Wine Quality
Residual Sugar
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
3
4
5
6
7
8
Quality

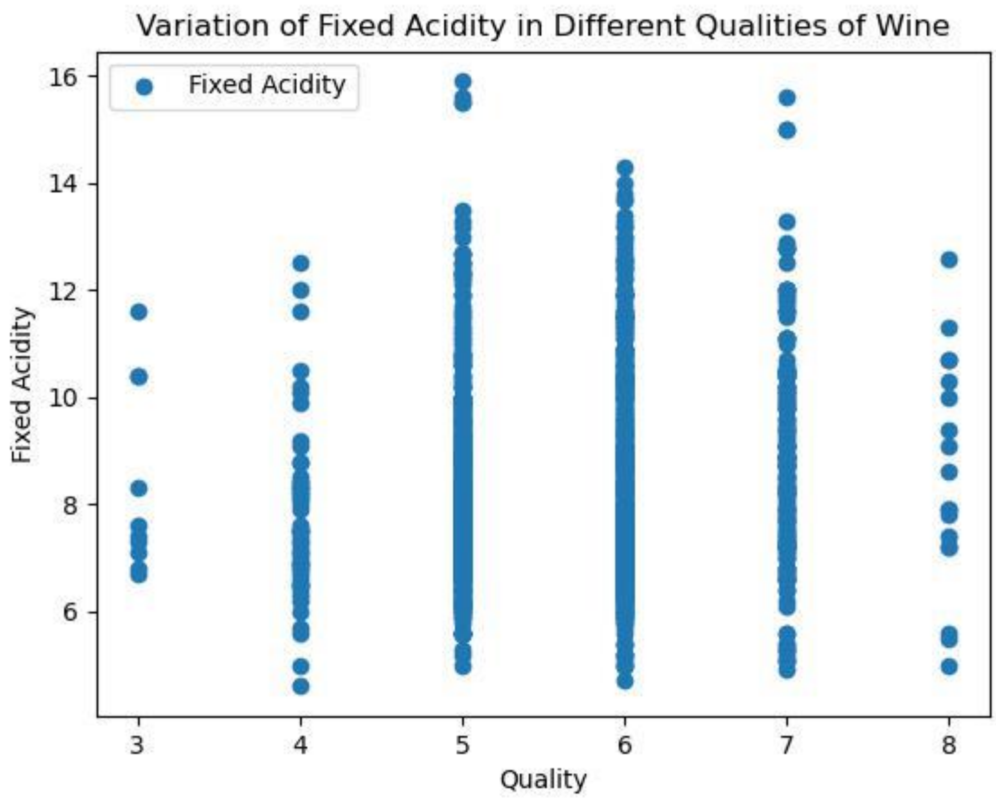

Free Sulfur Dioxide Levels by Wine Quality
Free Sulfur Dioxide
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
3
4
5
6
7
8
Quality

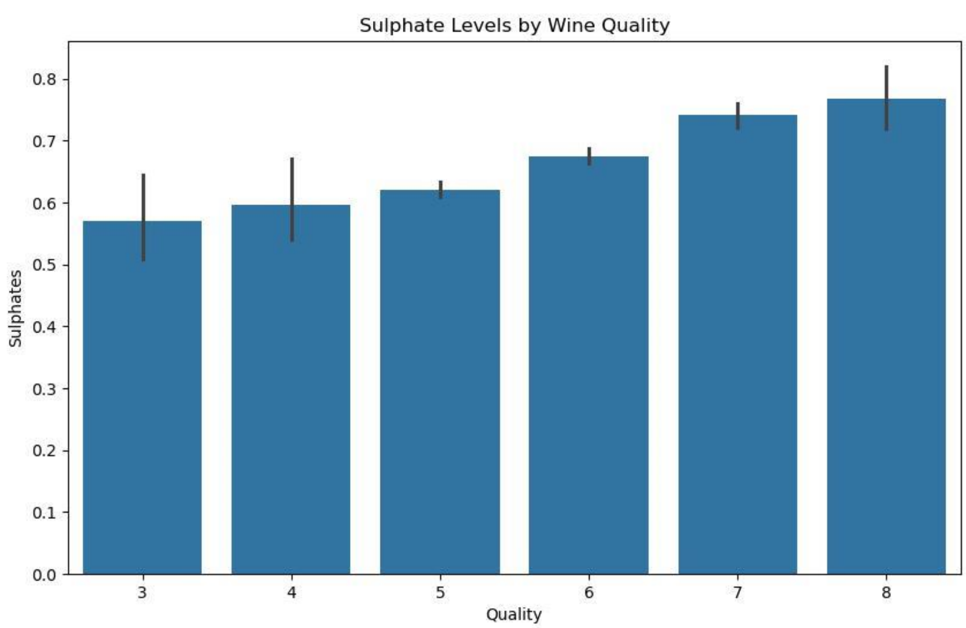

Citric Acid Levels by Wine Quality
Citric Acid
0.0
0.1
0.2
0.3
0.4
3
4
5
6
7
8
Quality


Chloride Levels by Wine Quality
Chlorides
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
3
4
5
6
7
8
Quality

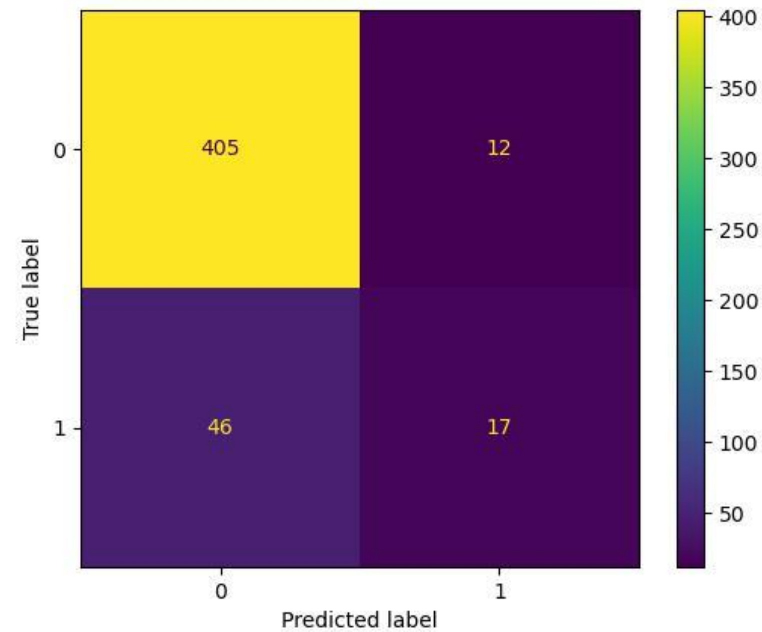

Alcohol Content by Wine Quality
Alcohol Content
alcohol
quality
0
2
4
6
8
10
12
14
3
4
5
6
7
8


Alcohol Content by Wine Quality
Alcohol Content
Wine Quality
7
8
9
10
11
12
13
14
15
3
4
5
6
7
8

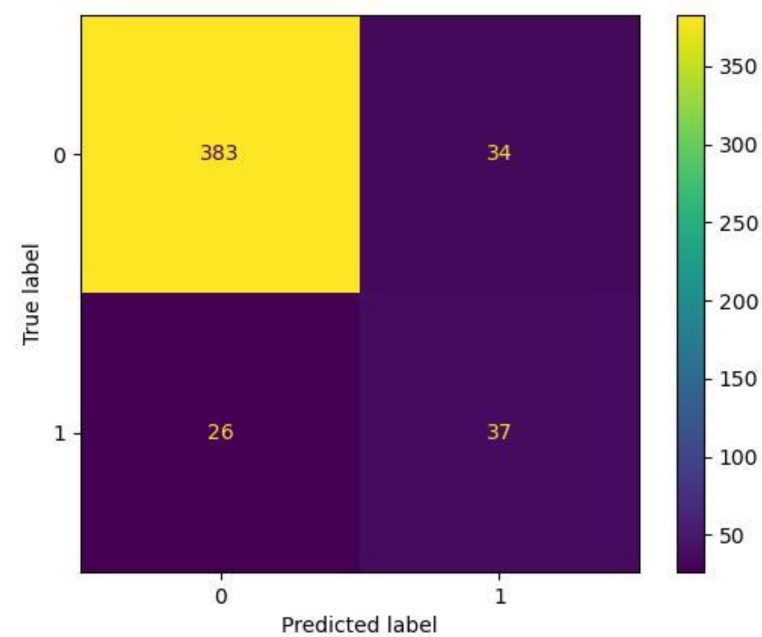

Variation of Fixed Acidity in Different Qualities of Wine
Fixed Acidity
Fixed Acidity
Quality
16
14
12
10
8
6
3
4
5
6
7
8


Sulphate Levels by Wine Quality
Sulphates
Quality
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
3
4
5
6
7
8

Confusion Matrix for Logistic Regression:



Confusion Matrix for Decision Tree:



Proposed Approach

Description of 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 87.9% 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

Description of any initial experiments or model testing

Rating:

Bad (quality > 7) 1382

Good (quality < 7) 217

fixed acidity	**citric acid**	**residual sugar**	**free sulfur dioxide**	**pH**	**sulphates**	**alcohol**	**quality**	
rating								
0	8.236831	0.254407	2.512120	16.172214	3.314616	0.644754	10.251037	5.408828
1	8.847005	0.376498	2.708756	13.981567	3.288802	0.743456	11.518049	7.082949

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 model.

Explanation of the solutions implemented

Balanced the dataset by adjusting class weights in the Logistic Regression model.

Next Steps

Outline of upcoming tasks and milestones

1. Hyperparameter tuning for improved model performance.
2. 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

3. 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