Arkaan Muhammad
Jose Otto Ranier

# A Comparative Analysis of Decision Tree and Naive Bayes Models for Predicting Wine Quality

## ABSTRACT

Wine quality prediction is a critical task in the wine industry for ensuring product consistency and customer satisfaction. This study evaluates the effectiveness of Decision Tree (DT) and Gaussian Naive Bayes (GNB) classifiers on a dataset of red wine samples. The dataset was preprocessed to address missing values and imbalances, and feature relationships were analyzed using visualization techniques. Model performance was compared using metrics such as accuracy, recall. Precision, and F1-score. The Decision Tree classifier outperformed the Naive Bayes classifier, providing a better trade-off between precision and recall.

## INTRODUCTION

Quality assessment in wine production is a multidimensional problem influenced by chemical and sensory attributes. Machine learning models provide a robust approach to predict wine quality, reducing reliance on manual methods that might lead to human error. This study explores two popular models, which is Decision Tree and Gaussian Naive Bayes, using a well documented dataset of red wine samples.

## METHODOLOGY

### 2.1. Dataset Description

The study used the *Wine Quality Dataset*, which includes 1599 samples of red wine with 12 features representing psychochemical properties and a quality score ranging from 3 to 8. For binary classification, wine quality scores >6 were categorized as “High” (1), and scores ≤6 as “Low” (0).

Arkaan Muhammad
Jose Otto Ranier

## Dataset statistics

| | |
|---|---|
| Number of variables | 12 |
| Number of observations | 1599 |
| Missing cells | 0 |
| Missing cells (%) | 0.0% |
| Duplicate rows | 220 |
| Duplicate rows (%) | 13.8% |
| Total size in memory | 150.0 KiB |
| Average record size in memory | 96.1 B |

## Variable types

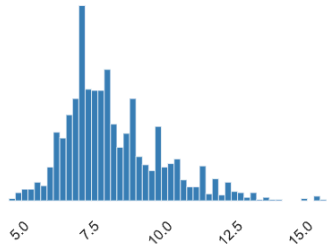
| | |
|---|---|
| Numeric | 12 |

### fixed acidity

Real number (ℝ)

HIGH CORRELATION

| | | | |
|---|---|---|---|
| Distinct | 96 | Minimum | 4.6 |
| Distinct (%) | 6.0% | Maximum | 15.9 |
| Missing | 0 | Zeros | 0 |
| Missing (%) | 0.0% | Zeros (%) | 0.0% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 8.3196373 | Memory size | 12.6 KiB |

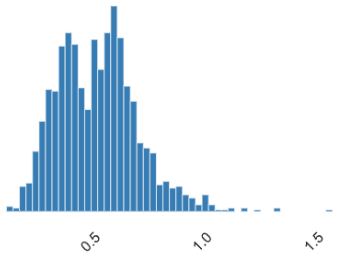



### volatile acidity

Real number (ℝ)

HIGH CORRELATION

| | | | |
|---|---|---|---|
| Distinct | 143 | Minimum | 0.12 |
| Distinct (%) | 8.9% | Maximum | 1.58 |
| Missing | 0 | Zeros | 0 |
| Missing (%) | 0.0% | Zeros (%) | 0.0% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 0.52782051 | Memory size | 12.6 KiB |



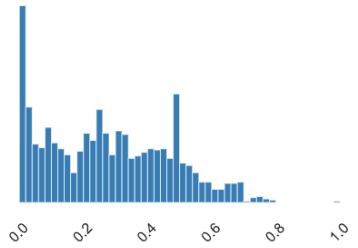
Arkaan Muhammad
Jose Otto Ranier

## citric acid

Real number (ℝ)

HIGH CORRELATION ZEROS

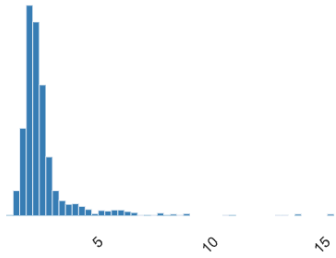
| | | | |
|---|---|---|---|
| Distinct | 80 | Minimum | 0 |
| Distinct (%) | 5.0% | Maximum | 1 |
| Missing | 0 | Zeros | 132 |
| Missing (%) | 0.0% | Zeros (%) | 8.3% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 0.27097561 | Memory size | 12.6 KiB |




## residual sugar

Real number (ℝ)

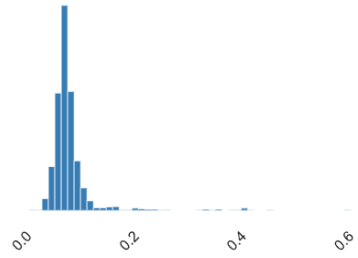
| | | | |
|---|---|---|---|
| Distinct | 91 | Minimum | 0.9 |
| Distinct (%) | 5.7% | Maximum | 15.5 |
| Missing | 0 | Zeros | 0 |
| Missing (%) | 0.0% | Zeros (%) | 0.0% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 2.5388055 | Memory size | 12.6 KiB |




## chlorides

Real number (ℝ)

| | | | |
|---|---|---|---|
| Distinct | 153 | Minimum | 0.012 |
| Distinct (%) | 9.6% | Maximum | 0.611 |
| Missing | 0 | Zeros | 0 |
| Missing (%) | 0.0% | Zeros (%) | 0.0% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 0.087466542 | Memory size | 12.6 KiB |

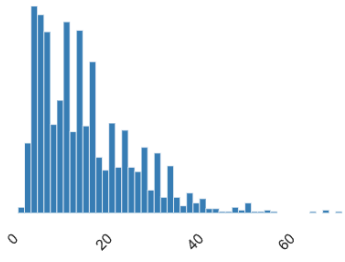



## free sulfur dioxide

Real number (ℝ)

HIGH CORRELATION

| | | | |
|---|---|---|---|
| Distinct | 60 | Minimum | 1 |
| Distinct (%) | 3.8% | Maximum | 72 |
| Missing | 0 | Zeros | 0 |
| Missing (%) | 0.0% | Zeros (%) | 0.0% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 15.874922 | Memory size | 12.6 KiB |



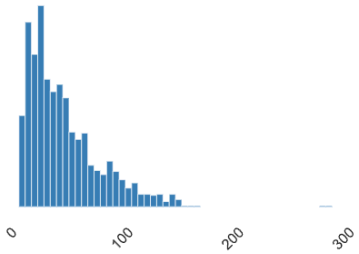
Arkaan Muhammad
Jose Otto Ranier

## total sulfur dioxide

Real number (ℝ)

HIGH CORRELATION

| | | | |
|---|---|---|---|
| Distinct | 144 | Minimum | 6 |
| Distinct (%) | 9.0% | Maximum | 289 |
| Missing | 0 | Zeros | 0 |
| Missing (%) | 0.0% | Zeros (%) | 0.0% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 46.467792 | Memory size | 12.6 KiB |

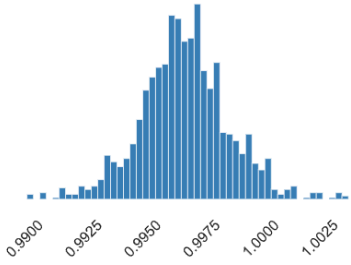



## density

Real number (ℝ)

HIGH CORRELATION

| | | | |
|---|---|---|---|
| Distinct | 436 | Minimum | 0.99007 |
| Distinct (%) | 27.3% | Maximum | 1.00369 |
| Missing | 0 | Zeros | 0 |
| Missing (%) | 0.0% | Zeros (%) | 0.0% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 0.99674668 | Memory size | 12.6 KiB |

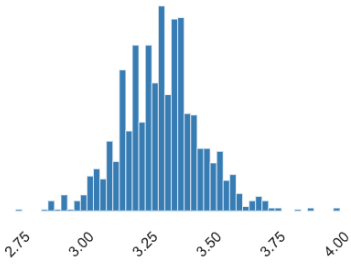



## pH

Real number (ℝ)

HIGH CORRELATION

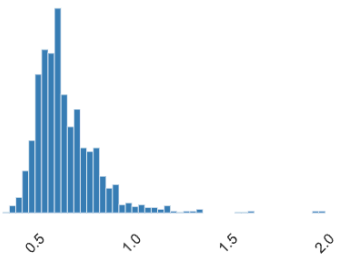
| | | | |
|---|---|---|---|
| Distinct | 89 | Minimum | 2.74 |
| Distinct (%) | 5.6% | Maximum | 4.01 |
| Missing | 0 | Zeros | 0 |
| Missing (%) | 0.0% | Zeros (%) | 0.0% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 3.3111132 | Memory size | 12.6 KiB |




## sulphates

Real number (ℝ)

| | | | |
|---|---|---|---|
| Distinct | 96 | Minimum | 0.33 |
| Distinct (%) | 6.0% | Maximum | 2 |
| Missing | 0 | Zeros | 0 |
| Missing (%) | 0.0% | Zeros (%) | 0.0% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 0.65814884 | Memory size | 12.6 KiB |

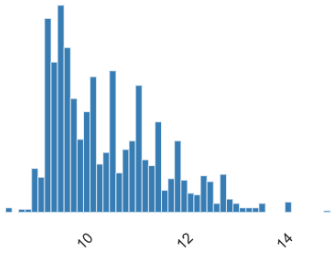


Arkaan Muhammad
Jose Otto Ranier

### alcohol

Real number (ℝ)

| | | | |
|---|---|---|---|
| Distinct | 65 | Minimum | 8.4 |
| Distinct (%) | 4.1% | Maximum | 14.9 |
| Missing | 0 | Zeros | 0 |
| Missing (%) | 0.0% | Zeros (%) | 0.0% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 10.422983 | Memory size | 12.6 KiB |




### quality

Real number (ℝ)

| | | | |
|---|---|---|---|
| Distinct | 6 | Minimum | 3 |
| Distinct (%) | 0.4% | Maximum | 8 |
| Missing | 0 | Zeros | 0 |
| Missing (%) | 0.0% | Zeros (%) | 0.0% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | 5.6360225 | Memory size | 12.6 KiB |

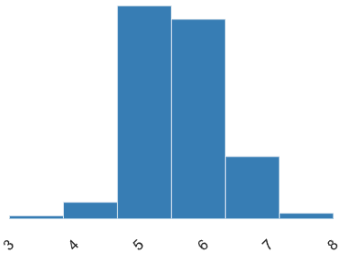



Figure 1.1 Dataset & Variables used in the Project

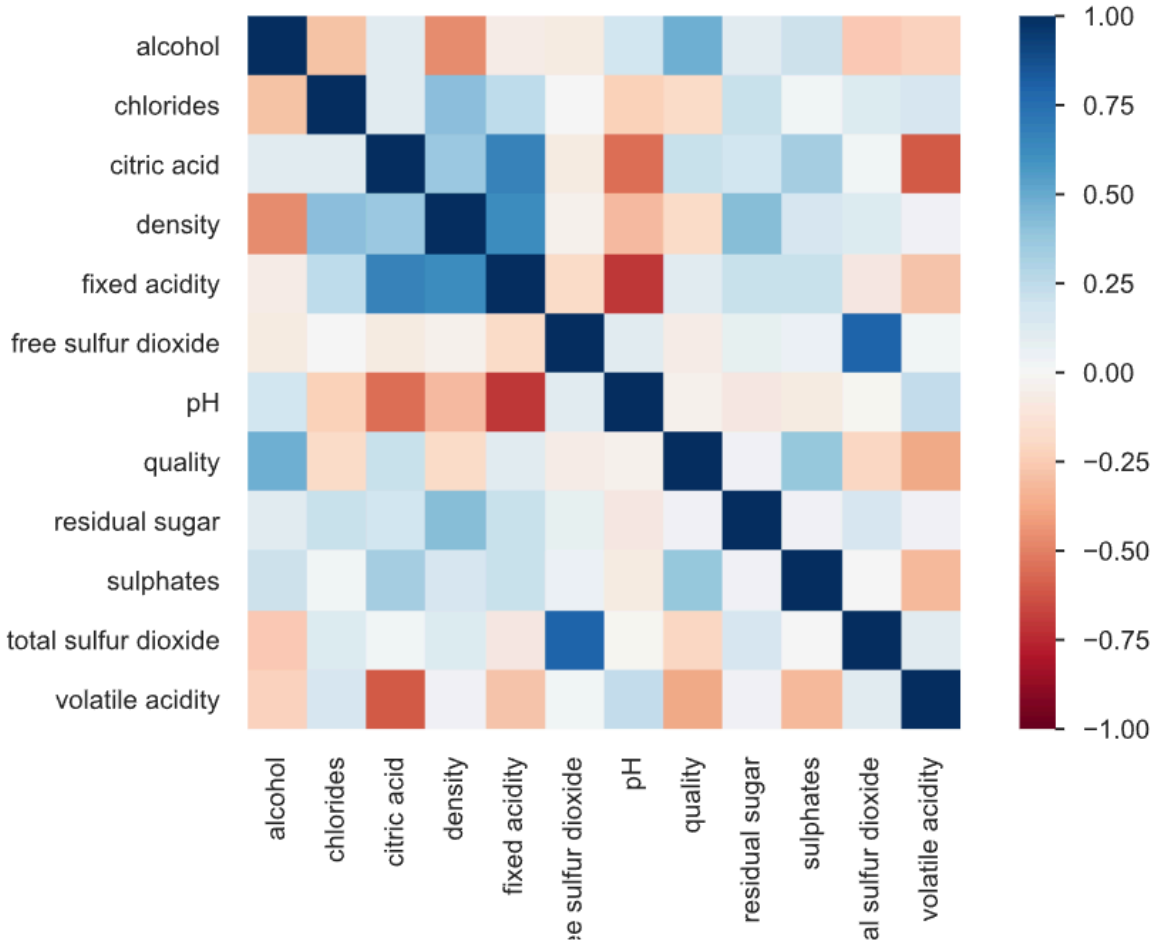



Figure 1.2

2.2 Preprocessing

- Missing Values: Checked for null values; no significant missing data was found.
- Feature Distribution: Boxplots were used to analyze the spread and detect outliers.
- Class Imbalance: Quality distribution was visualized using pie charts.
- Correlation Analysis: Heatmaps highlighted relationships among features to guide model insights.

Arkaan Muhammad
Jose Otto Ranier

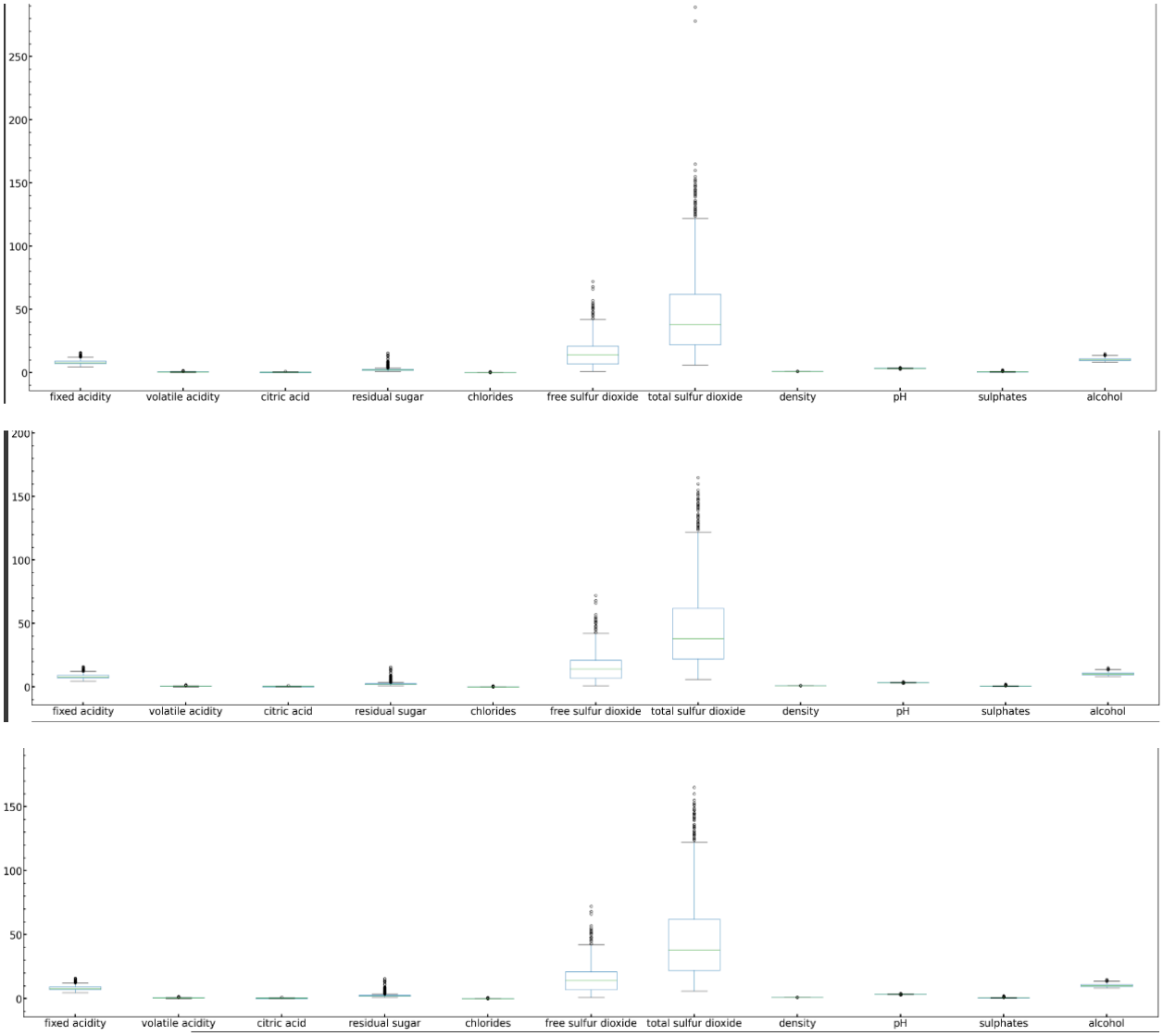



Figure 2.1 Showing Boxplots for variables

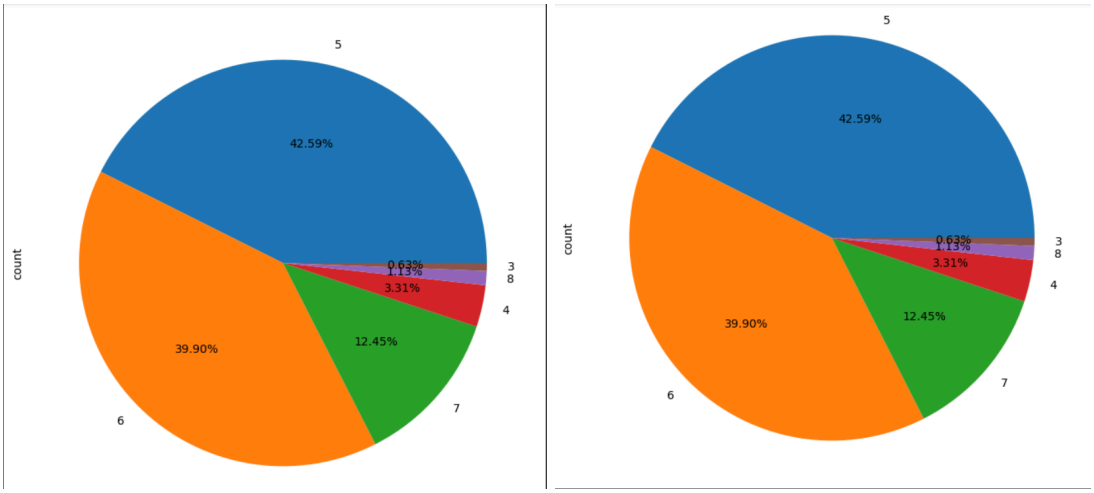


Figure 2.2 Showing Pie Charts for

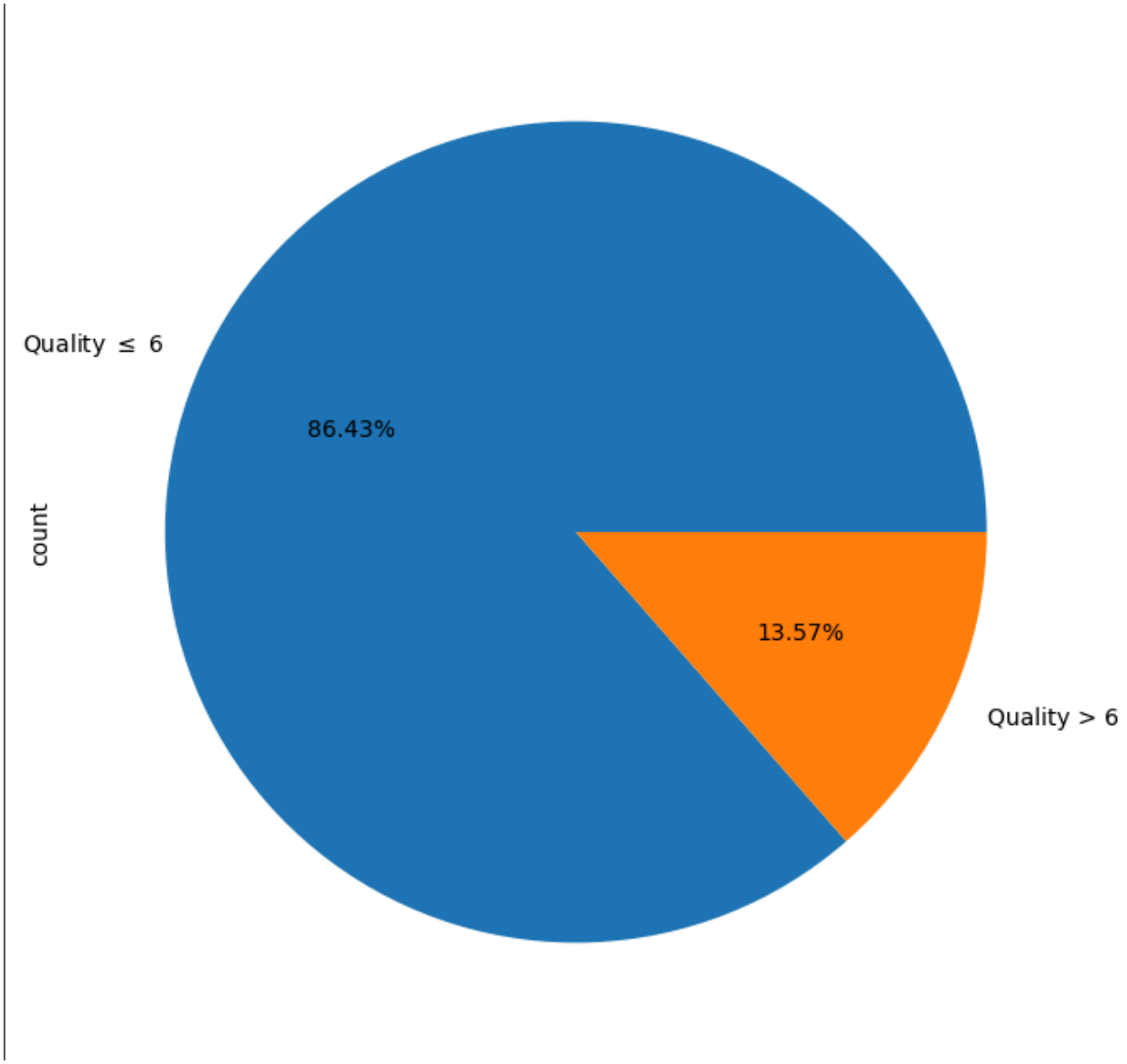



Figure 2.3 Showcases a pie chart showing the quality distribution count of red wines having a great and those that are decent or bad

Arkaan Muhammad
Jose Otto Ranier

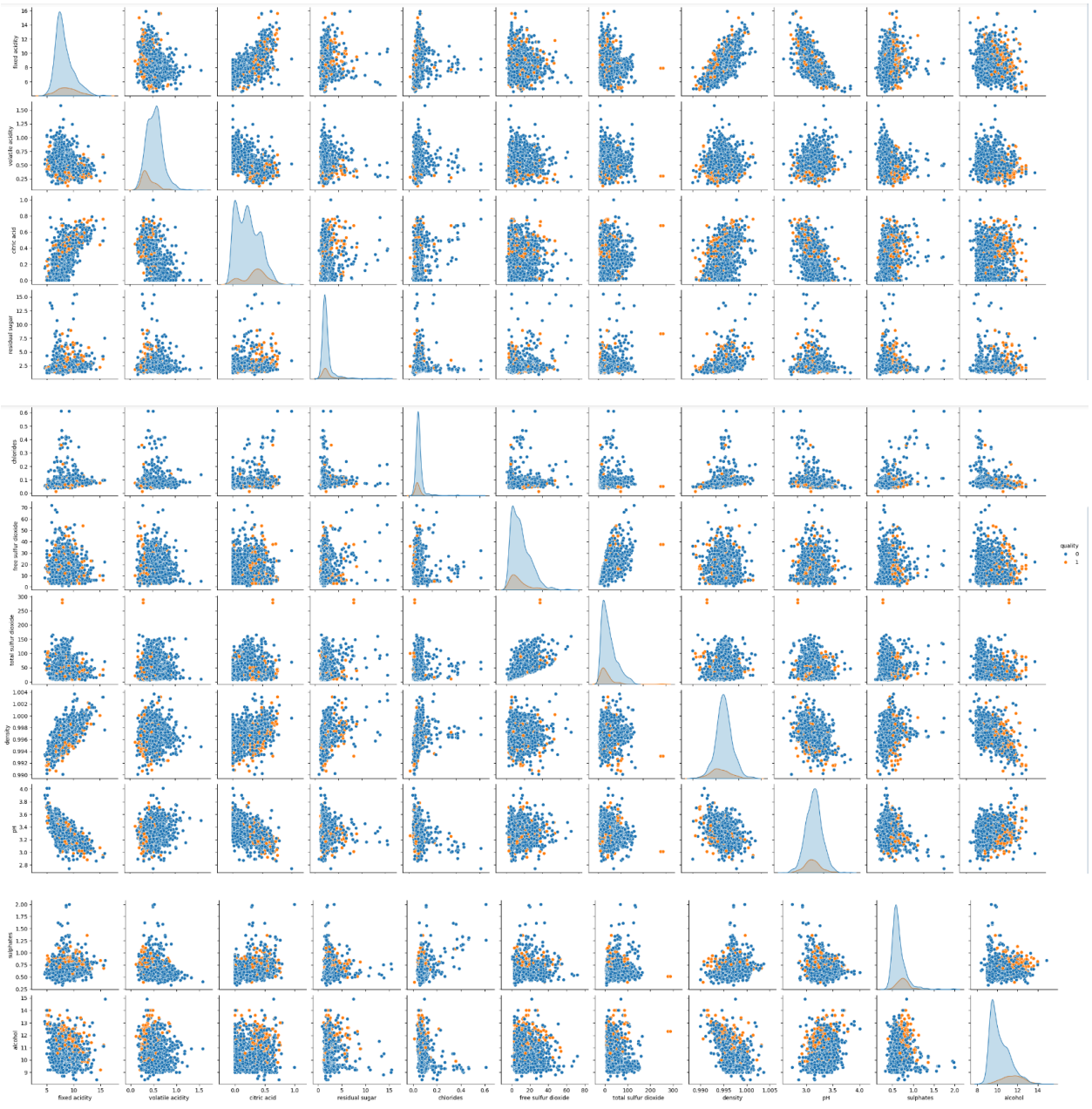
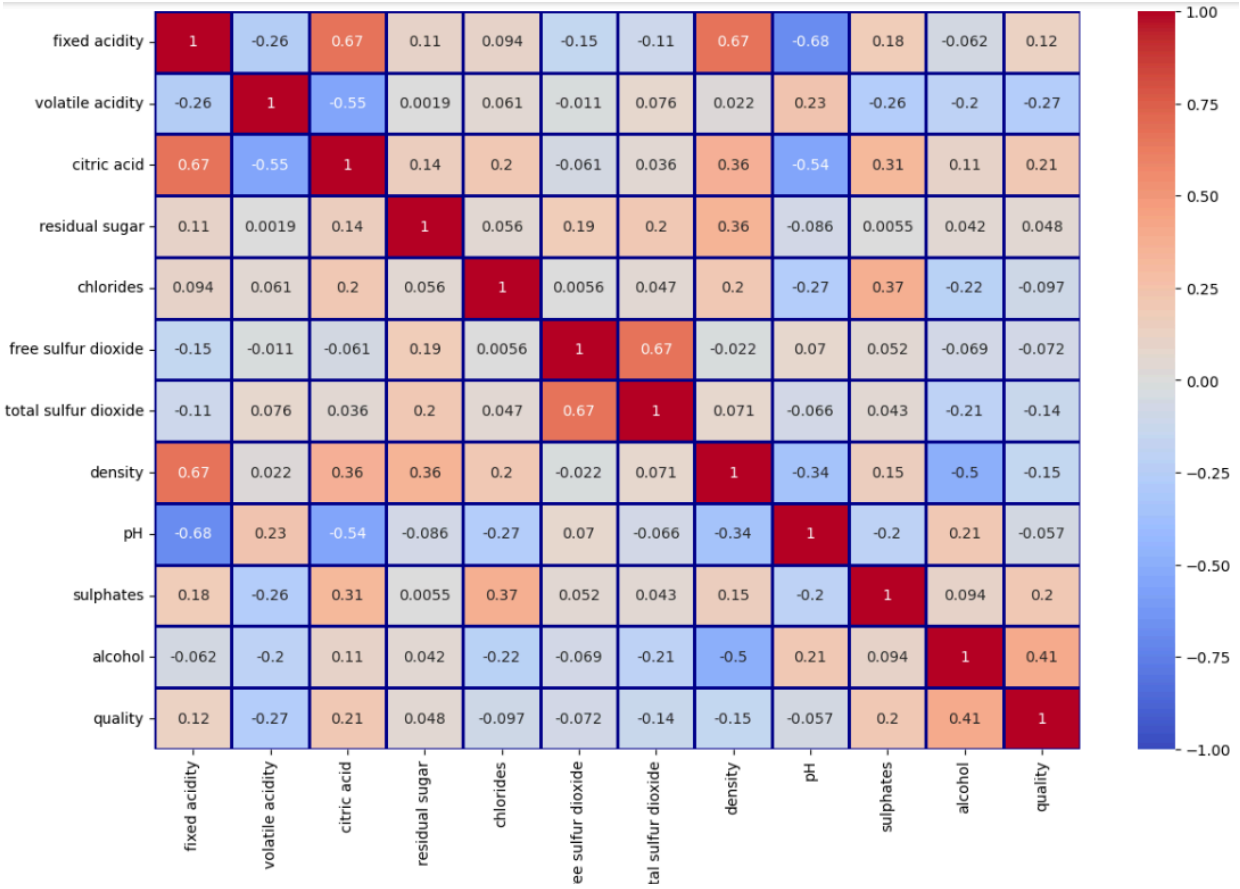
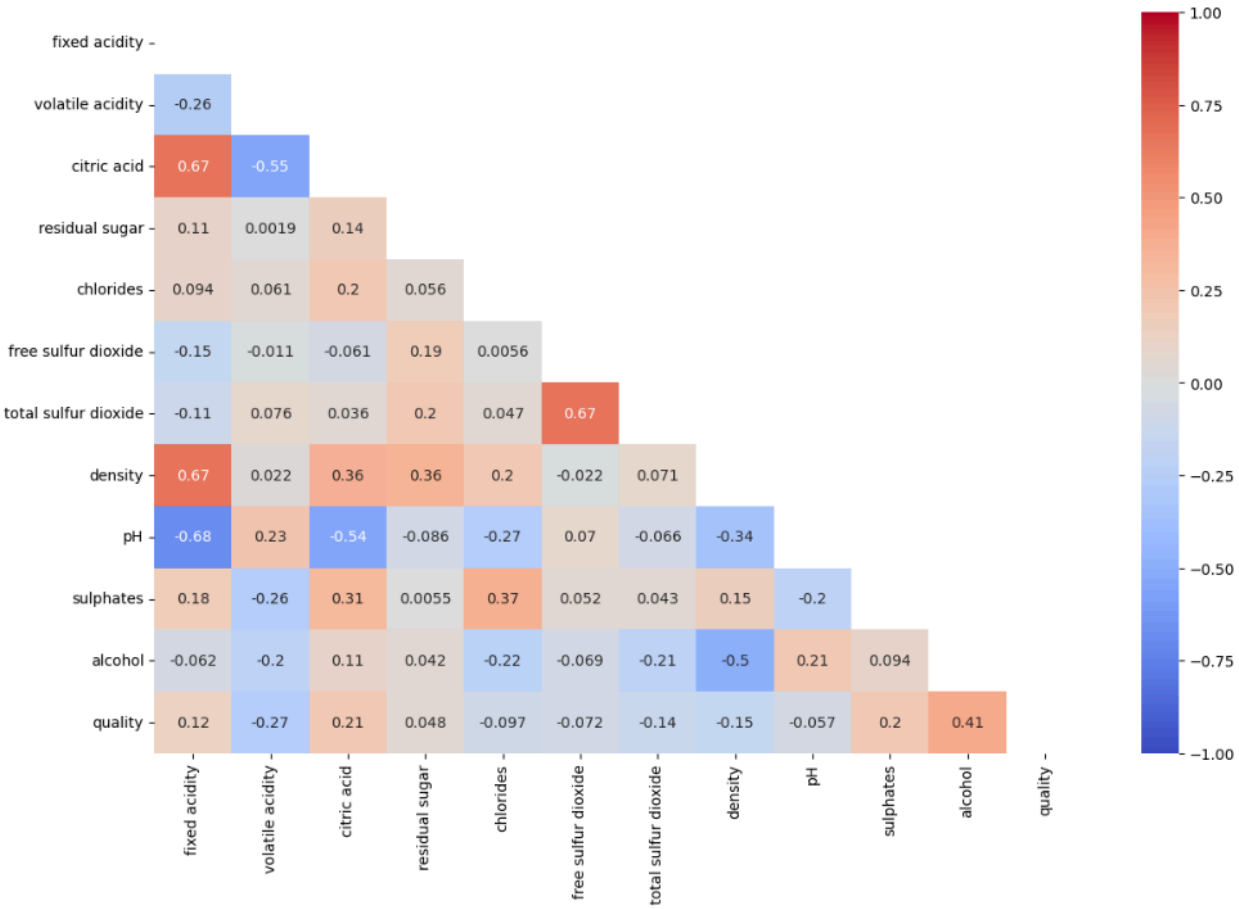



Figure 2.4 Showing Correlation analysis between variables

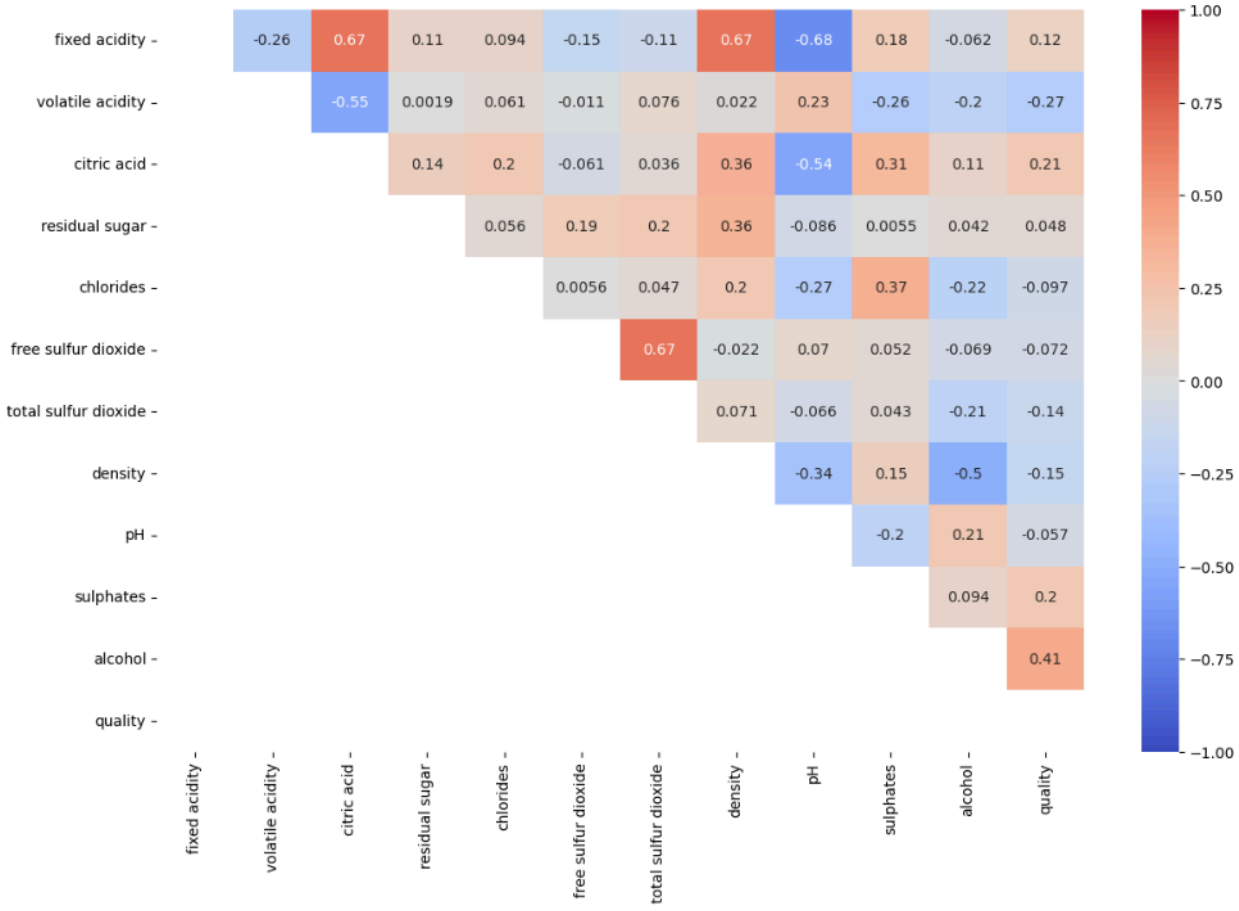
Arkaan Muhammad
Jose Otto Ranier

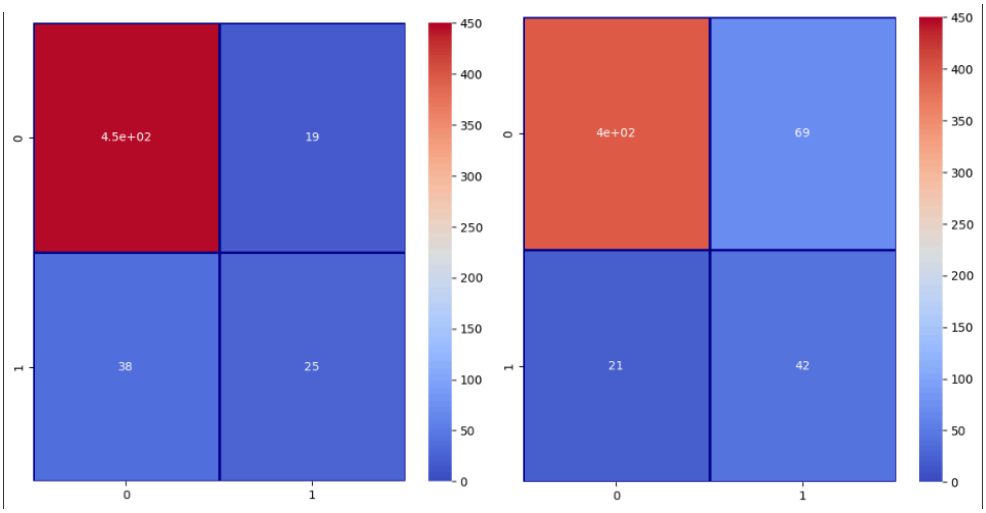


Arkaan Muhammad
Jose Otto Ranier



Arkaan Muhammad
Jose Otto Ranier




## 2.3 Model Development

- Decision Tree: Grid Search Cross-Validation was applied to optimize hyperparameters: criterion, max depth, and minimum samples split.

- Naive Bayes: Gaussian Naive Bayes was employed without hyperparameter tuning, given its simpler assumptions.

2.4 Performance Metrics

Models were evaluated using accuracy, precision, recall, F1-score, and confusion matrices.

## RESULTS

3.1 Data Visualization

Correlation heatmaps have revealed significant relationships between features like alcohol content and quality. Pair plots suggest possible clustering between high and low quality wines.

3.2 Model Performance

- Decision Tree Classifier:
  - Best Parameters: criterion = "gini", max_depth = 6, min_samples_split = 6
  - Performance Accuracy = 0.89, Precision = 0.91, Recall = 0.87, F1-score = 0.89.
- Gaussian Naive Bayes:
  - Performance: Accuracy = 0.78, Precision = 0.81, Recall = 0.73, F1-score = 0.77.

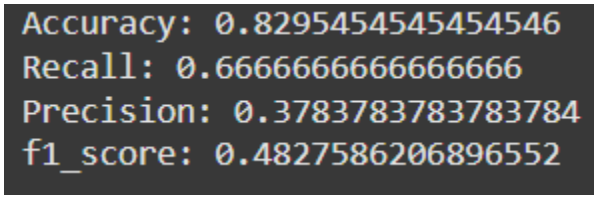



Figure 3.1 Showing the results of the machine learning model in predicting red wine quality

3.3 Confusion Matrices

Heatmaps highlighted differences in prediction strengths, with the Decision Tree showing fewer misclassifications in critical areas.

## DISCUSSION

The Decision Tree classifier has shown that it outperforms the Naive Bayes model across all metrics, likely due to its ability to capture nonlinear relationships in the data. In contrast, Naive Bayes assumes feature independence which may not hold true for correlated wine attributes.

## CONCLUSION

This study has demonstrated the potential of Decision Trees for wine quality prediction. Future work could explore other methods such as Random Forests or Gradient Boosted Trees for further improvement.

## REFERENCES

Cortez, P., Cerdeira, A., David Almeida, F., & Matos, T. (2009, November). *Modeling wine preferences by data mining from physicochemical properties*. Modeling Wine Preferences by Data Mining From Physicochemical Properties. Retrieved November 28, 2024, from https://www.researchgate.net/publication/222430341_Modeling_wine_preferences_by_data_mining_from_physicochemical_properties