Predicting High-Quality Wines Using Logistic Regression Based on Chemical Properties

GITHUB Repository Link:

https://github.com/UmerAsif435/Predicting-High-Quality-Wines

Table of Contents

Introduction

1.1 Project overview

This work has been focused on predicting wine quality with Logistic Regression from their chemical composition. Expert human tasters classically determined wine quality, but using machine learning can offer an objective and scalable solution (Arshad, 2024). With the physicochemical parameters of wines, including acidity, sugar levels, alcohol content, and pH, this research seeks to create a prediction model that will predict whether the wines are high- or low-quality. The data for this project comes from publicly provided wine quality data sets, such as the UCI Wine Quality Dataset, which includes chemical measurements and related quality scores. The CRISP-DM methodology is followed in the project, which includes data gathering, preprocessing, feature selection, training of models, and evaluation.

1.2 Project aim

The major objective of this project is to create a predictive model based on Logistic Regression for classifying wines as high or low quality depending on their chemical makeup. The objectives are as follows,

- Discovering the relationship between chemical properties and the quality of wine.
- Preprocessing the data, managing missing data, feature scaling, and feature selection.
- Applying Logistic Regression, a common classification algorithm, to make wine quality predictions.
- Measuring model performance in terms of accuracy, precision, recall, and F1-score.
- Complementation of outcomes against other ML models, such as Random Forest or SVM, to enhance the effectiveness.

With these objectives accomplished, this project offers a data-driven methodology of wine quality estimation, which could be helpful for wine makers, sommeliers, and the beverage business since it will automate the determination of quality as well as make wine grading consistent. This project predicts high-quality wines via Logistic Regression, depending on their chemical composition (Niyogisubizo *et al.* 2025). Expert tasting has conventionally been used to determine wine quality, but machine learning provides a more objective, scalable, and data-driven method.

Literature review

2.1 Machine Learning in Wine Quality Prediction

Wine quality assessment through professional taster evaluations relies on subjective human judgments that show variation. The ML methods including "***Logistic regression, support vector machines, decision trees, and neural networks"*** have been studied for determining wine quality from chemical composition (Jain et al. 2023). Logistic regression serves as one of the extensively used methods for binary classification problems therefore making it suitable for wine quality separation. Studies have proven that ML models are capable of detecting patterns in physicochemical characteristics, enhancing the effectiveness of wine quality evaluation.

2.2 Influence of Chemical Properties on Wine Quality

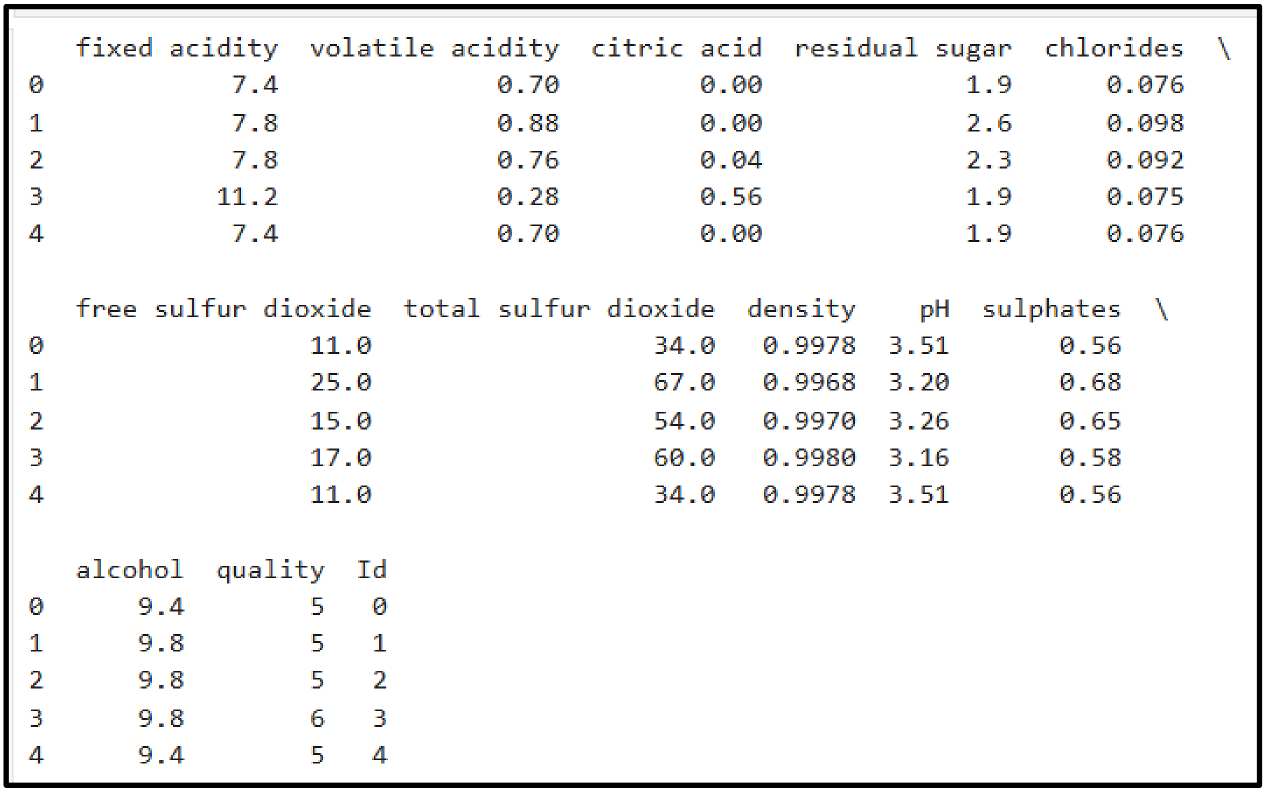
The physicochemical characteristics of wine determine its quality to a great extent, as they impact taste, smell, and general sensory quality. The most important factors are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol (Gutiérrez-Escobar *et al.* 2021). Research shows that alcohol content and volatile acidity correlate well with wine quality, such that increased alcohol content generally improves sensory perception, while high volatile acidity can cause unwanted sourness. Knowledge of these correlations assists in the choice of the most suitable features for predictive modeling, improving classification accuracy.

2.3 Logistic Regression for Wine Quality Classification

Logistic regression is a popular statistical model for binary classification, and thus it is a good option for predicting whether a wine is high or low quality. It predicts the likelihood of an outcome by using predictor variables and transforming values to between 0 and 1 with the sigmoid function (Clarin, 2022). The simplicity and interpretability of the model are why it is also a method of choice for studying the effect of specific chemical characteristics on wine quality. But logistic regression requires the relationships between independent variables and log odds to be linear, which could restrict its ability to perform well on highly complicated or nonlinear data. Scientists usually compare its accuracy with other ML models and find that though it is good for baseline classification, ensemble techniques such as random forests can provide greater predictive accuracy.

Methodology

Data Collection:

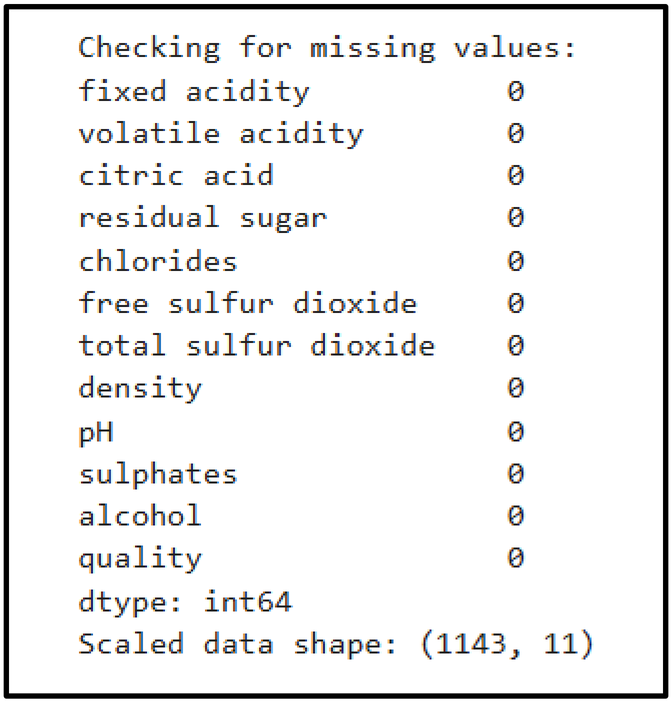


```
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0            7.4              0.70         0.00             1.9      0.076
1            7.8              0.88         0.00             2.6      0.098
2            7.8              0.76         0.04             2.3      0.092
3           11.2              0.28         0.56             1.9      0.075
4            7.4              0.70         0.00             1.9      0.076

   free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                 11.0                  34.0   0.9978  3.51       0.56
1                 25.0                  67.0   0.9968  3.20       0.68
2                 15.0                  54.0   0.9970  3.26       0.65
3                 17.0                  60.0   0.9980  3.16       0.58
4                 11.0                  34.0   0.9978  3.51       0.56

   alcohol  quality  Id
0      9.4        5   0
1      9.8        5   1
2      9.8        5   2
3      9.8        6   3
4      9.4        5   4
```

Figure 1: Data Loading

This image shows the first few rows of data, which is the wine quality dataset. Secondly, it contains the various "chemical properties of wine such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, as well as other properties like alcohol content and quality. There is also information on free sulfur dioxide, total sulfur dioxide, density, pH, and sulfates available on the dataset. The quality rating of each wine sample ranges from 0 to 10. These chemical features have been collected from Kaggle (Kaggle, 2025) and then modeled and predicted to qualify the wine, so this data has been used. For predictive analysis to predict wine quality using machine learning techniques, the dataset is a very valuable one.

Data Preprocessing:



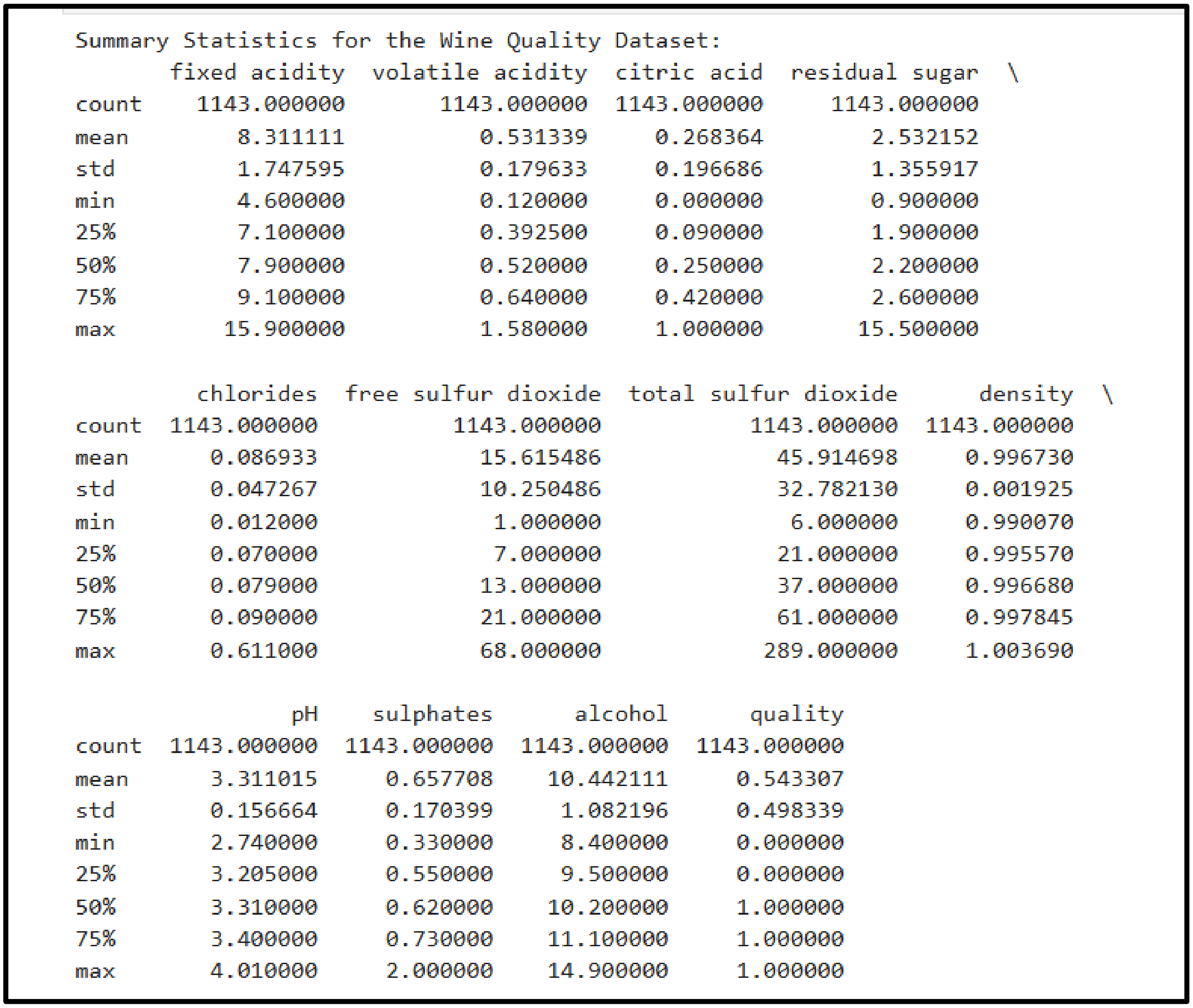
```
Checking for missing values:
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
dtype: int64
Scaled data shape: (1143, 11)
```

Figure 2: Checking the missing values and scaling the data

The following figure illustrates the missing value check and scaling process on the wine quality dataset. It was checked through the columns of each feature and made sure there aren't any blank values. All features, such as "fixed acidity, volatile acidity, citric acid, etc", have no missing values (as shown, 0). After which, it scales the dataset, there is a need for the same scale features that algorithms like Logistic Regression require. Based on the scaled data shape, the dataset has 1143 samples by 11 features. The preprocessing steps taken include:

- **Missing Value Check:** Ensured no missing values in any feature.
- **Features Scaling:** Used these features, and applied these through StandardScaler to normalize the features, making sure all variables could be centered around zero with a standard deviation of one.

Exploratory Data Analysis:



```
Summary Statistics for the Wine Quality Dataset:
       fixed acidity  volatile acidity  citric acid  residual sugar  \
count    1143.000000       1143.000000  1143.000000     1143.000000   
mean        8.311111          0.531339     0.268364        2.532152   
std         1.747595          0.179633     0.196686        1.355917   
min         4.600000          0.120000     0.000000        0.900000   
25%         7.100000          0.392500     0.090000        1.900000   
50%         7.900000          0.520000     0.250000        2.200000   
75%         9.100000          0.640000     0.420000        2.600000   
max        15.900000          1.580000     1.000000       15.500000   

         chlorides  free sulfur dioxide  total sulfur dioxide      density  \
count  1143.000000          1143.000000           1143.000000  1143.000000   
mean      0.086933            15.615486             45.914698     0.996730   
std       0.047267            10.250486             32.782130     0.001925   
min       0.012000             1.000000              6.000000     0.990070   
25%       0.070000             7.000000             21.000000     0.995570   
50%       0.079000            13.000000             37.000000     0.996680   
75%       0.090000            21.000000             61.000000     0.997845   
max       0.611000            68.000000            289.000000     1.003690   

                pH    sulphates      alcohol      quality  
count  1143.000000  1143.000000  1143.000000  1143.000000  
mean      3.311015     0.657708    10.442111     0.543307  
std       0.156664     0.170399     1.082196     0.498339  
min       2.740000     0.330000     8.400000     0.000000  
25%       3.205000     0.550000     9.500000     0.000000  
50%       3.310000     0.620000    10.200000     1.000000  
75%       3.400000     0.730000    11.100000     1.000000  
max       4.010000     2.000000    14.900000     1.000000  
```

Figure 3: Summary Statistics

"The summary statistics of the wine quality dataset are given in this figure. For each feature, an important set of statistical measures is brought with in the table by including mean, standard deviation, min, max, and percentiles". Suppose the mean of fixed acidity is given to be 8.31 and of alcohol is 10.44. It is a good summary that summarizes central tendency, spread and range values of each feature, which can lead to identifying possible outliers, skewness, or distributions on which transformation is required before modeling.

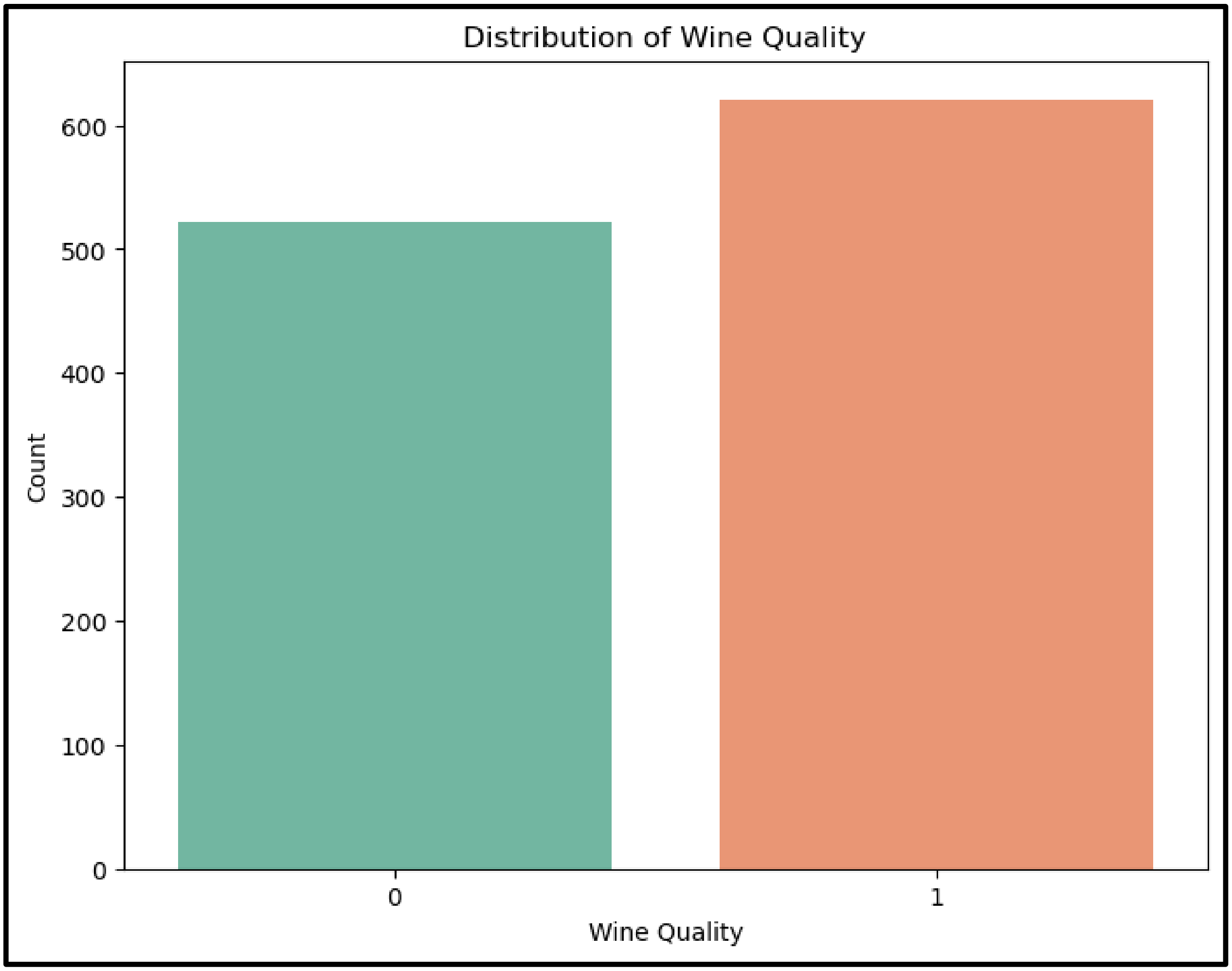
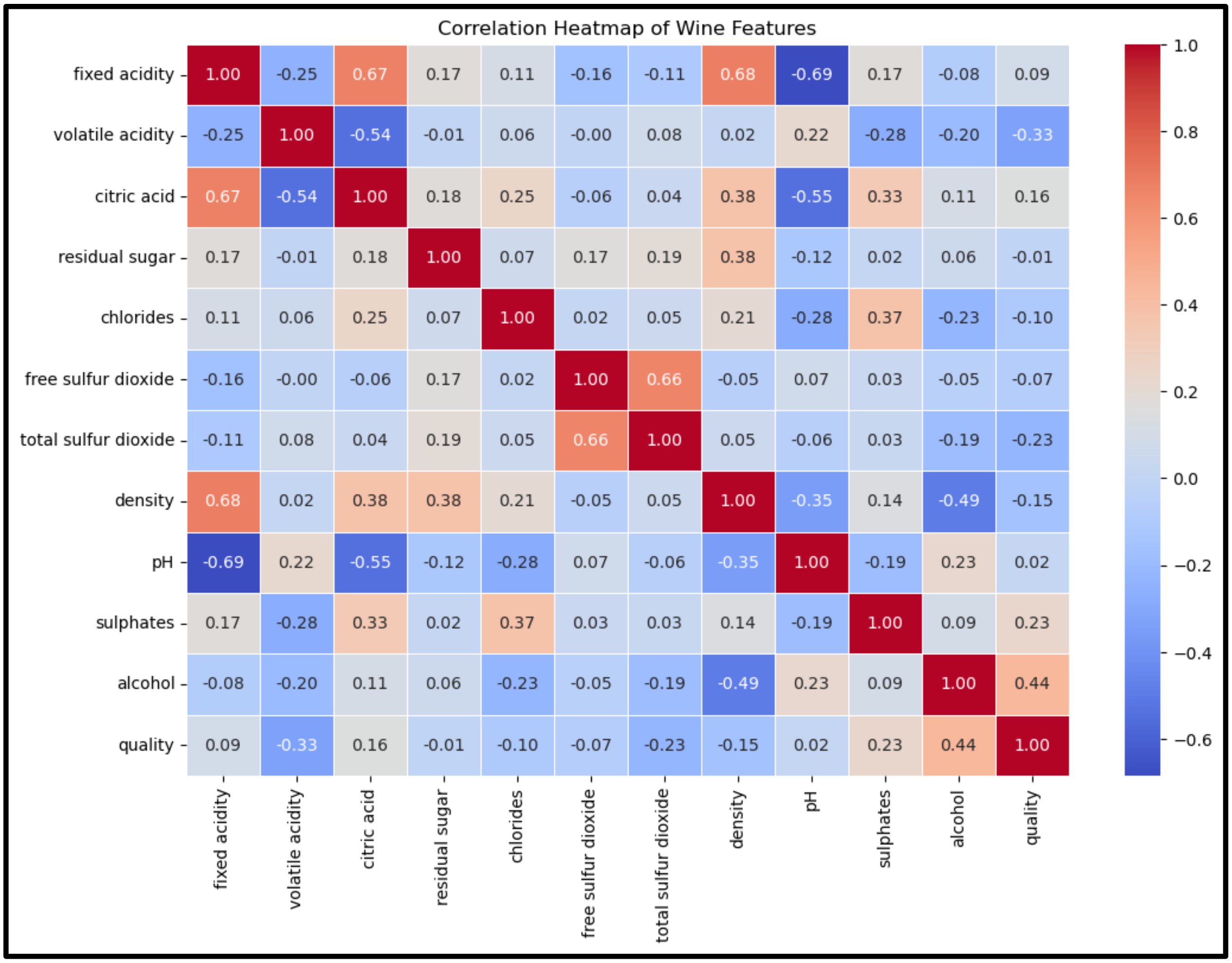


Figure 4: Distribution of Wine Quality

This visualizes how the “wine quality ratings are distributed in the dataset. The wine quality in the x axis is zero, one is high or low quality. In this case, the x axis describes different categories of wine, and the y axis describes the count of wines in that category”. The figure also shows that the dataset is fairly balanced between low and high quality wines. This will inform the modelling process and establish whether or not the dataset is imbalanced, in wherein case resampling or changing classification thresholds will be required.



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.00	-0.25	0.67	0.17	0.11	-0.16	-0.11	0.68	-0.69	0.17	-0.08	0.09
volatile acidity	-0.25	1.00	-0.54	-0.01	0.06	-0.00	0.08	0.02	0.22	-0.28	-0.20	-0.33
citric acid	0.67	-0.54	1.00	0.18	0.25	-0.06	0.04	0.38	-0.55	0.33	0.11	0.16
residual sugar	0.17	-0.01	0.18	1.00	0.07	0.17	0.19	0.38	-0.12	0.02	0.06	-0.01
chlorides	0.11	0.06	0.25	0.07	1.00	0.02	0.05	0.21	-0.28	0.37	-0.23	-0.10
free sulfur dioxide	-0.16	-0.00	-0.06	0.17	0.02	1.00	0.66	-0.05	0.07	0.03	-0.05	-0.07
total sulfur dioxide	-0.11	0.08	0.04	0.19	0.05	0.66	1.00	0.05	-0.06	0.03	-0.19	-0.23
density	0.68	0.02	0.38	0.38	0.21	-0.05	0.05	1.00	-0.35	0.14	-0.49	-0.15
pH	-0.69	0.22	-0.55	-0.12	-0.28	0.07	-0.06	-0.35	1.00	-0.19	0.23	0.02
sulphates	0.17	-0.28	0.33	0.02	0.37	0.03	0.03	0.14	-0.19	1.00	0.09	0.23
alcohol	-0.08	-0.20	0.11	0.06	-0.23	-0.05	-0.19	-0.49	0.23	0.09	1.00	0.44
quality	0.09	-0.33	0.16	-0.01	-0.10	-0.07	-0.23	-0.15	0.02	0.23	0.44	1.00

Figure 5: Correlation Heatmap of Wine Features

The correlation is used to visualize the relationship between different wine features. Values of the correlation are shown on a color gradient “with blue indicating negative correlation and red indicating positive correlation. Alcohol has a positive correlation of 0.44 with quality, and volatile acidity has a negative correlation of -0.28 with quality”. This helps you to easily identify which features are most contributing to quality in order to select features and interpret this model.

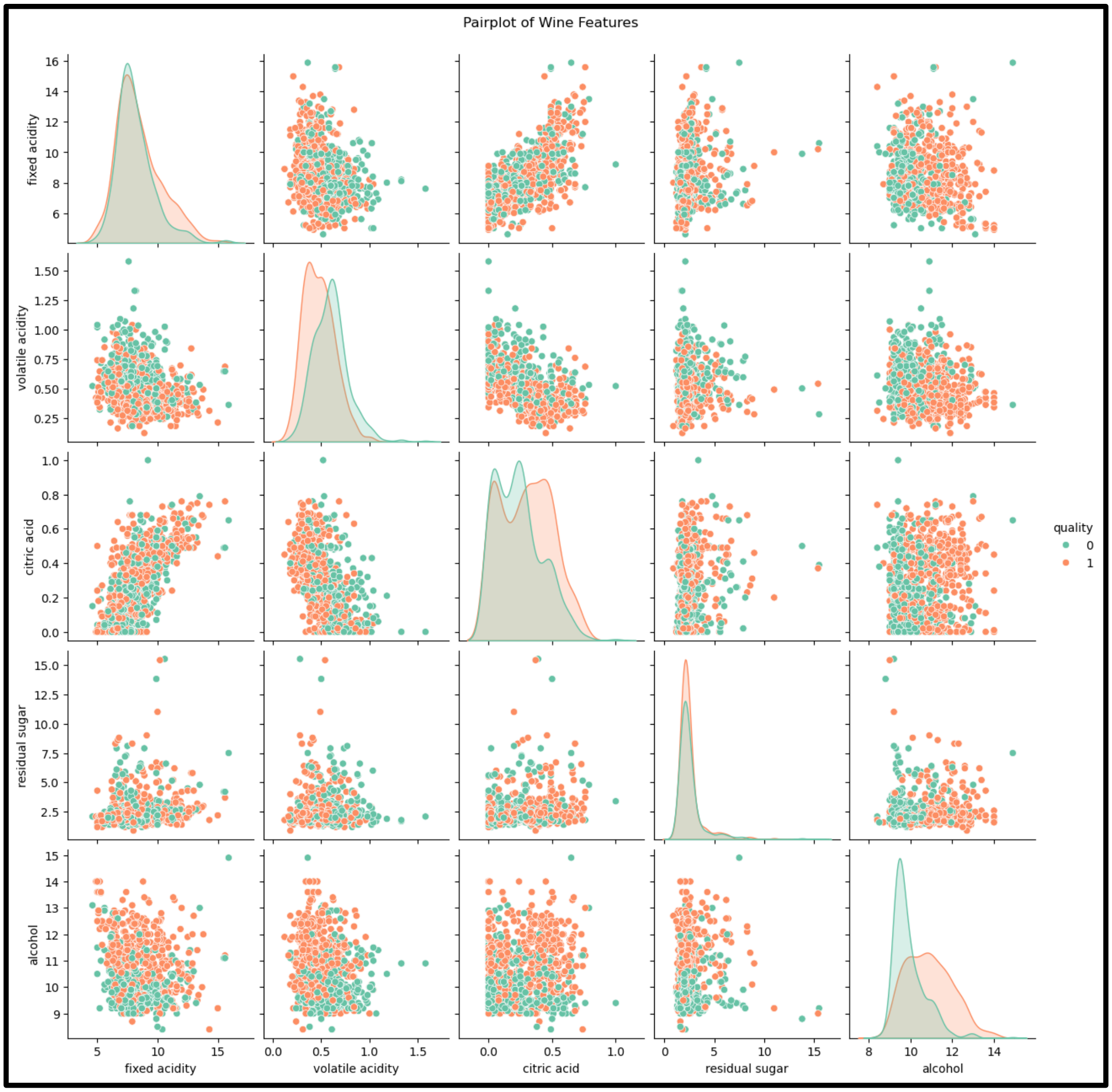


Figure 6: Pairplot of Wine Features

This is a pairplot between certain wine features and wine quality (0, 1). Scatter plots colored by wine quality show the relationship of two features and the diagonal plots are of the distribution of each feature. For instance, “features such as alcohol and fixed acidity are found to have contrasting distributions between the low and high quality wines”. This visualization can identify potential patterns or separations in the data that can be helpful in training a classification model.

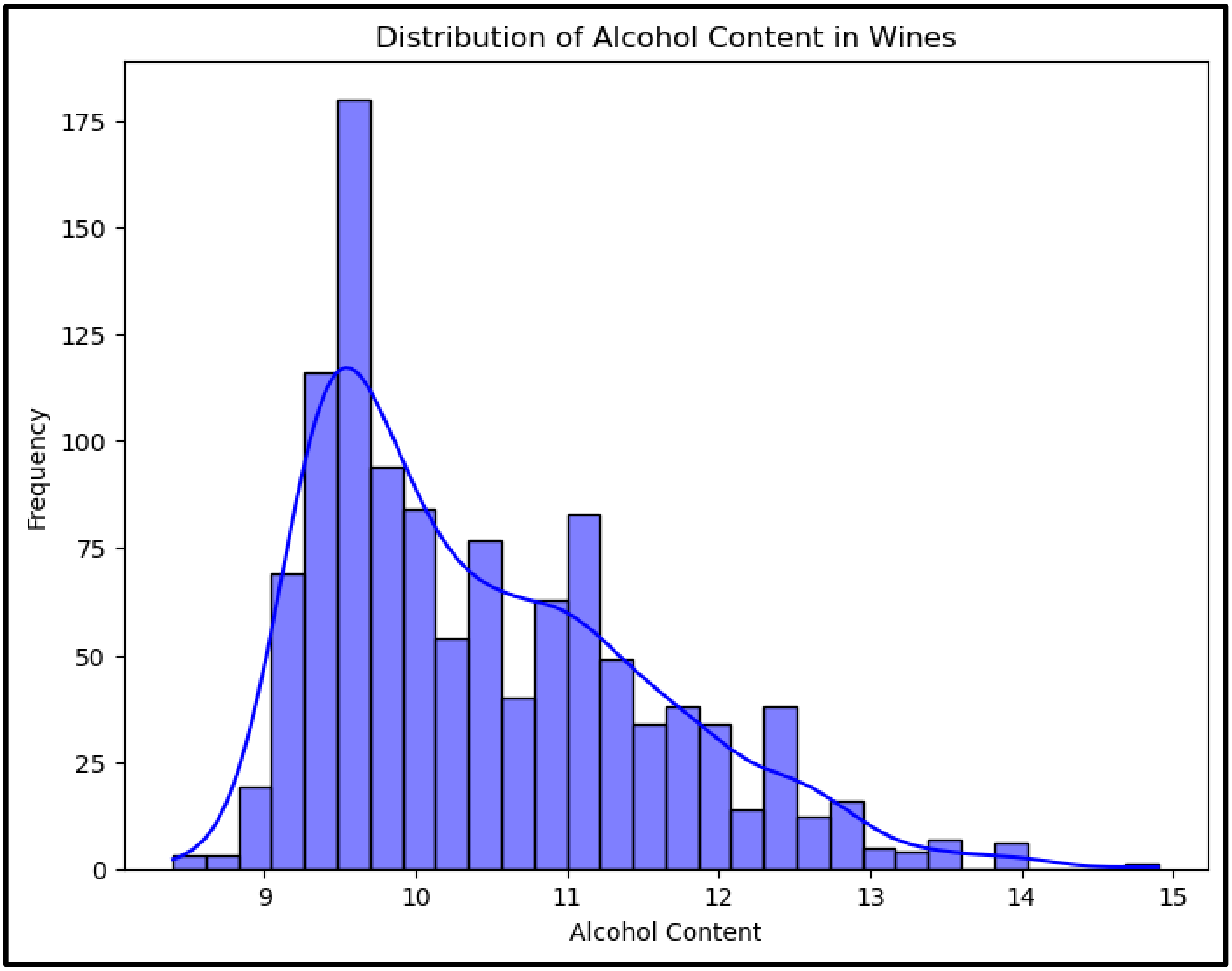
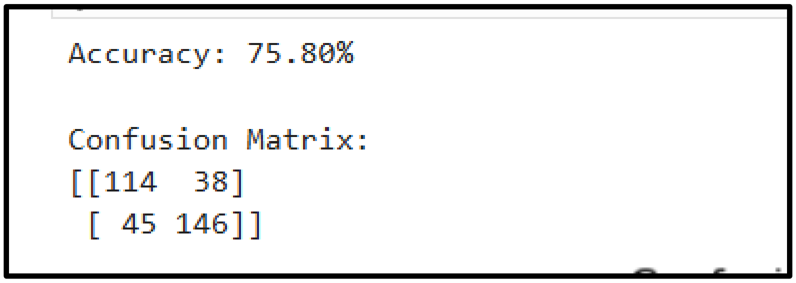


Figure 7: Distribution of Alcohol Content in Wines

In this figure, it can be seen the distribution of the value of alcohol content in the wine dataset. A Kernel Density Estimate (KDE) plot and histogram together show how most of the alcohol content is concentrated around 9.5 - 11.5, with a spike at 10 is concentrated within. The distribution is slightly right skewed such that higher alcohol content wines are less common. “The distribution shape allows information about how alcohol concentration varies across the wines and may indicate how alcohol concentration is related to wine quality in predictive modeling.

Result and analysis



```
Accuracy: 75.80%

Confusion Matrix:
[[114  38]
 [ 45 146]]
```

Figure 8: Logistic Regression Model Accuracy and Confusion Matrix

This confusion matrix, along with the accuracy of the Logistic Regression model on this data, is 75.80%. It is a matrix that contains the number of true positives, false positives, true negatives, and false negatives. The values show that for 114 low-quality wines, the model classified them as low, and for 146 high-quality wines, the model also classified them as high, misclassifying 45 high-quality wines as low, and 38 low-quality wines as high. The classification error in this matrix is what this matrix evaluates".

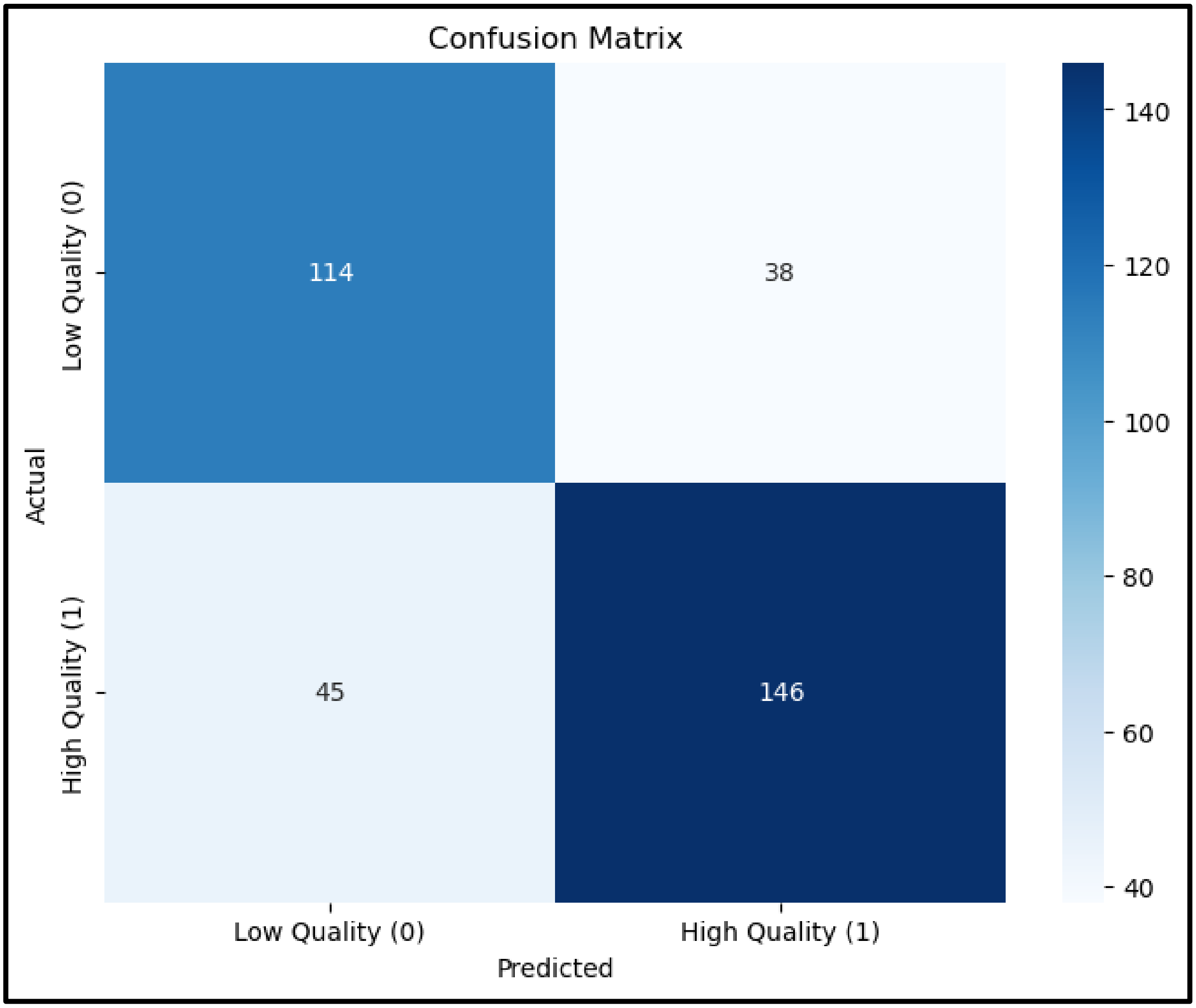
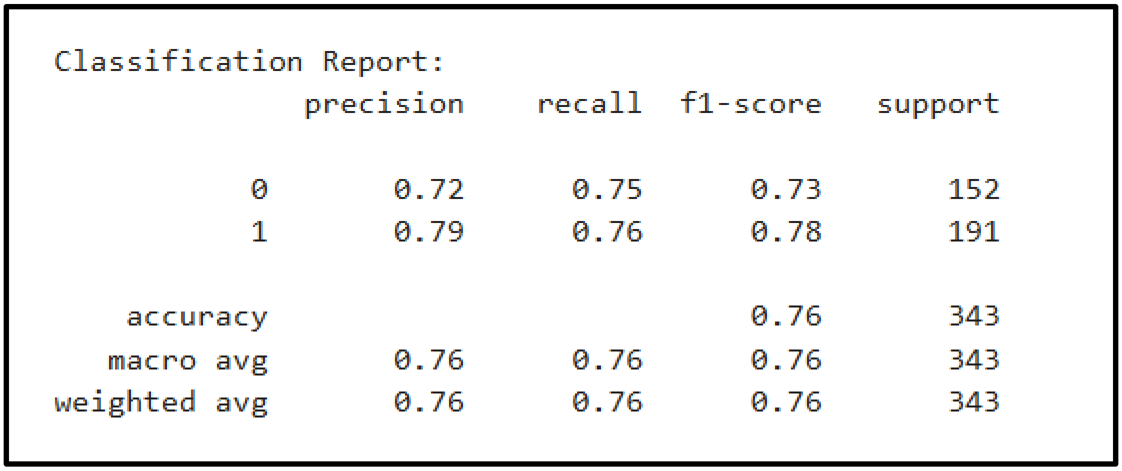


Figure 9: Visualizing the Confusion Matrix of the Logistic Regression Model

This figure corresponds to this heatmap, which is a visual depiction of the confusion matrix. This is making the color intensity that shows how many predictions were in each category, and the darker the color more predictions were present in it. "The confusion matrix makes the model's accuracy obvious in classifying high quality wines (146 true positives) and high quality wines (114 true positives), proving why it sometimes fails 45 negatives of false and 38 of false positives. By the design of the visualization, there is less intent to infer the model but rather to spot its weaknesses.



```
Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.75      0.73       152
           1       0.79      0.76      0.78       191

    accuracy                           0.76       343
   macro avg       0.76      0.76      0.76       343
weighted avg       0.76      0.76      0.76       343
```

Figure 10: Classification Report of Logistic Regression Model

The classification report gives you the performance metric of the result of the logistic regression model for both classes (low quality – 0, high quality – 1), that is, precision, recall, and F1 score. Precision defines the accuracy of positive predictions, and recall stands for what portion of the model that can find true positives. The F1 score is the harmonic mean of precision and recall and is a balanced metric". The report also includes an accuracy (76%) and averages for both classes. Breaking each class down in this way helps assess the model's effectiveness in each.

Ethical consideration

The project complies with ethical requirements on data integrity, transparency, and proper AI utilization. Data sourcing is done from publicly available sets, conforming to privacy norms and best practice data acquisition processes. Bias against model prediction outcomes is corrected for through balancing of training datasets and testing of outcomes through diverse means. Misconstruing the predictions, especially for commercial purposes, is limited through an assertion of model boundary and human management. Further, the research fosters equity in AI decision-making to avoid possible biases in wine quality determination that may affect producers.

Recommendation

Some changes to the Logistic Regression model can be applied to improve its performance level. Professional practitioners should attempt feature engineering techniques involving interaction terms and polynomial features to reveal advanced patterns between wine quality characteristics and chemical composition. A combination of minority class oversampling and the utilization of SMOTE contributes to reducing misclassification errors according to Dumitrescu et al (2022). Model performance benefits from being improved through hyperparameter tuning by applying both grid search method and cross-validation. The analysis should establish Random Forest or SVM models to verify their potential for greater accuracy. The addition of wine connoisseur domain expertise to feature selection processes would build better predictions while reducing classification mistakes.

Overall conclusion

The research implements Logistic Regression as a forecasting method which achieves 75.80% accuracy in predicting wine quality based on chemical characteristics. The model demonstrated strong accuracy in its ability to classify wine quality according to high or low standards based on the analysis results shown in the confusion matrix and classification report. An improved performance can result from creating new features as well as balancing the data alongside parameter optimization. Random Forest and SVM models should be compared for potential improvement of predictive accuracy. On the whole, the research illustrates how machine learning is capable of impacting wine quality ratings, providing the wine industry with a data-intensive, scalable mechanism for sustaining and enhancing quality controls.

Reference

Arshad, H., 2024. The Wine Quality Prediction Using Machine Learning. *Journal of Innovative Computing and Emerging Technologies*, *4*(2). Available at https://jicet.org/index.php/JICET/article/view/146

Clarin, A., 2022. Comparison of the performance of several regression algorithms in predicting the quality of white wine in WEKA. *Int. J. Emerg. Technol. Adv. Eng.*, *12*(7), pp.20-26. Available at https://www.researchgate.net/profile/Jeffrey-Clarin/publication/361765795_Comparison_of_the_Performance_of_Several_Regression_Algorithms_in_Predicting_the_Quality_of_White_Wine_in_WEKA/links/62ca7e3a3bbe636e0c51b616/Comparison-of-the-Performance-of-Several-Regression-Algorithms-in-Predicting-the-Quality-of-White-Wine-in-WEKA.pdf

Dumitrescu, E., Hué, S., Hurlin, C. and Tokpavi, S., 2022. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), pp.1178-1192. Available at https://www.sciencedirect.com/science/article/pii/S0377221721005695

Gutiérrez-Escobar, R., Aliaño-González, M.J. and Cantos-Villar, E., 2021. Wine polyphenol content and its influence on wine quality and properties: A review. *Molecules*, *26*(3), p.718. Available at https://www.mdpi.com/1420-3049/26/3/718

Jain, K., Kaushik, K., Gupta, S.K., Mahajan, S. and Kadry, S., 2023. Machine learning-based predictive modelling for the enhancement of wine quality. *Scientific Reports*, *13*(1), p.17042. Available at https://www.nature.com/articles/s41598-023-44111-9

Kaggle, 2025. Wine Quality Dataset, Viewed on 26th March 2025. From https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

Niyogisubizo, J., de Dieu Ninteretse, J., Nziyumva, E., Nshimiyimana, M., Murwanashyaka, E. and Habiyakare, E., 2025. Towards Predicting the Quality of Red Wine Using Novel Machine Learning Methods for Classification, Data Visualization, and Analysis. In *Artificial Intelligence* *and Applications* (Vol. 3, No. 1, pp. 31-42). Available at http://ojs.bonviewpress.com/index.php/AIA/article/view/1999