AY2024/2025 SEMESTER 1

DBA5106: Foundations of Business Analytics

Group Project 2

Chen Wei A0298073U

Jessica Yap Yi May A0297752M

Lee Jisu A0298088H

Yu Wangcong A0139463Y

1. Abstract

In the wine industry, assessing quality is crucial for determining market value, aligning with consumer preferences, and ensuring the success of a wine product. This project utilized the Wine Quality Dataset, which contains physicochemical properties and sensory quality data for red wines, to build a classification model for predicting wine grades. The report explored both glassbox models such as shallow Decision Trees, and blackbox models like LightGBM, comparing their interpretability and performance. Additionally, SHAP (SHapley Additive exPlanations) was applied to tree-based blackbox models to provide insights into feature importance at both the global and local levels. This analysis aimed to identify a balance between model accuracy and interpretability for practical applications in wine quality prediction.

2. Introduction

Predictive wine grading models have significant real-world applications. In wineries, such models can be integrated into quality control processes, allowing producers to maintain consistent standards while minimizing reliance on subjective sensory evaluations. Moreover, predictive grading systems offer a cost-effective alternative to traditional grading methods by leveraging chemical data from routine lab tests, making them especially beneficial for smaller producers. Lastly, these models provide objective measures of quality that wineries can use for market positioning and pricing strategies, helping them stay competitive. By combining accuracy, efficiency, and interpretability, wine grade prediction models hold the potential to transform the industry at multiple levels.

3. Data Pre-processing and Key Assumptions

This project utilizes the Wine Quality Dataset, a widely studied dataset available on Kaggle. The dataset contains 1,144 observations of red wine samples, each described by 11 features such as fixed acidity, alcohol content, etc. Additionally, each sample includes a quality score which serves as the target variable.

The quality column in the dataset is rated on a scale from 0 to 10, but the actual quality scores in the dataset range primarily from 3 to 8. This distribution is realistic, as most wines fall within the middle range in practice. Extreme quality scores are rare, with 0 and 1 representing poor or faulty wines, and 9 and 10 typically reserved for exceptional or rare wines. Also, the data is related to red variants of the Portuguese "Vinho Verde" wine, meaning the wines share similar features. As a result, the dataset has less variability in quality. To convert the dataset into classification, quality scores were grouped into three discrete categories: low, medium and high.

This process allows the model to classify wines into different quality classes rather than predicting an exact score. For the purpose of this project, all misclassifications of wine grades are treated equally, regardless of the degree of deviation from the true grade. This approach allows the project to prioritize overall model performance and interpretability without introducing additional complexity related to the weighting of errors.

The grouping was done as follows:

- Low Quality: Wines with quality scores less than 6 were categorized as low quality.

- Medium Quality: Wines with quality score equal to 6 were categorized as medium quality.
- High Quality: Wines with quality scores higher than 6 were categorized as high quality.

Model used an 80/20 split for dividing the dataset into training and test sets, where 80% of the data was used to train the model, and 20% was held back for testing. This returned 915 datapoints in the training set and 229 datapoints in the test set. This division was achieved using train_test_split from the sklearn.model_selection module.

4. Glassbox Model – Shallow Decision Tree

In this model, DecisionTreeClassifier from the sklearn.tree module was used to classify data with a shallow decision tree. DecisionTreeClassifier model was trained on the training set using Gini impurity to measure quality of split to identify the best possible split at each node. The model was evaluated using accuracy scores on the test set. To be able to explain the model, the max_depth was set to be low at 3.

Using this baseline model, an accuracy score of 0.6 was obtained. This shallow model is easily interpreted by the Tree Visualisation in Figure 1 of the appendix. The best splits are alcohol percentage and sulphates, which is the first layer of splits that led to the greatest loss reduction.

Model's accuracy changed with increasing max_depth which shows its impact on overfitting and underfitting. As the max_depth parameter increased, test accuracy peaked at max_depth = 8 and then began to decrease, indicating overfitting as the model became more complex and tailored to the training data. To avoid overfitting, max_depth value was chosen where the test accuracy is maximized. In this case, a max_depth of 8 seemed optimal, balancing model complexity and generalization.

Max_depth	Accuracy Score
6	0.63
8	0.66
10	0.64
12	0.63

After training the DecisionTreeClassifier, the model was evaluated on the test set using multiple metrics, including prediction counts, correct prediction percentages, and the confusion matrix. The following table summarizes the model's performance for each quality category:

Actual Label	Count of Actual Samples (test set)	Predicted Labels	Count of Prediction	Prediction (%)
High	35	[High, Medium, Low]	[19,15,1]	54.29%
Low	108	[Medium, Low, High]	[33,71,4]	65.74%
Medium	86	[Medium, Low, High]	[61,17,8]	70.93%

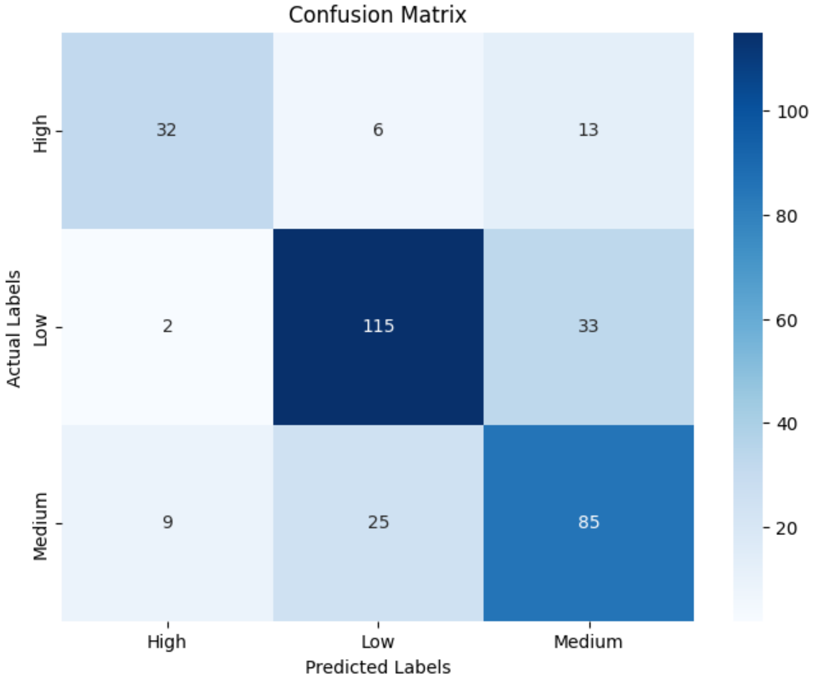
Medium quality wines are classified significantly more accurate than other quality categories. Meanwhile, high quality wines and low quality wines are often misclassified as medium quality wines, resulting in lower accuracy. The model achieved correct prediction percentages of 65.74%, 70.93% and 54.29% for low, medium, and high quality wines, respectively. Overall accuracy score returned 0.66.

5. Blackbox Model - LightGBM

The goal of this analysis was to build a model that can predict the quality of wines based on various features. LightGBM classifier was used to predict these wine quality categories from the given features.

100 tree iterations were generated, each tree improving on the loss using gradient boost. As this was a multiclass classification, the surrogate loss of multi logarithmic loss was used to represent misclassification loss.

Actual Label	**Count of Actual Samples (test set)**	**Predicted Labels**	**Count of Prediction**	**Prediction %**
High	51	[High, Medium, Low]	[32,13,6]	62.75%
Low	150	[Medium, Low, High]	[2,115,33]	76.67%
Medium	119	[Medium, Low, High]	[85,25,9]	71.43%



The model performs best for accurately classifying low quality wines, achieving the highest correct prediction rate of 76.67%, likely due to their higher prevalence in the dataset. Medium quality wines exhibit a slightly lower correct prediction rate of 71.43%, with notable misclassifications as low quality wines (25 cases). The model struggles the most with high quality wines, achieving a correct prediction rate of only 62.75%, with frequent misclassifications as medium quality wine (13 cases). Overall accuracy score of the model is 0.72.

Blackbox Interpretation Using SHAP (local and global):

i. Local SHAP

The SHAP value of features allows us to see how much it contributes to the decision-making process. Features with higher SHAP values (positive or negative) exert a stronger influence on the predicted wine grades, while features with near-zero SHAP values have minimal impact.

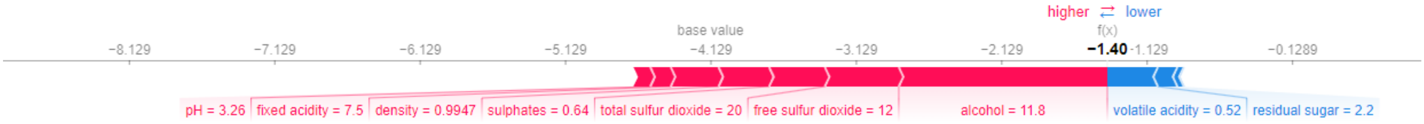
We looked at 3 random test predictions to see how each feature contributed to the final prediction. From our code, here is a correct prediction of a medium quality wine, along with its features.

Features of sample

fixed acidity	7.50000
volatile acidity	0.52000
citric acid	0.40000
residual sugar	2.20000
chlorides	0.06000
free sulfur dioxide	12.00000
total sulfur dioxide	20.00000
density	0.99474
pH	3.26000
sulphates	0.64000
alcohol	11.80000

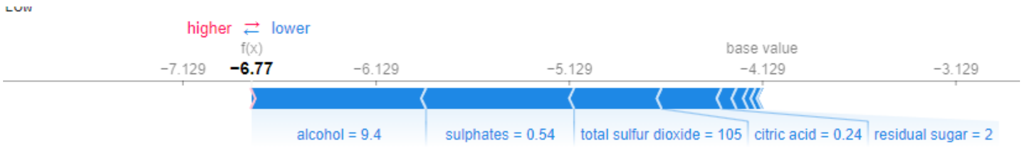
Actual Quality: Medium

Predicted Quality: Medium



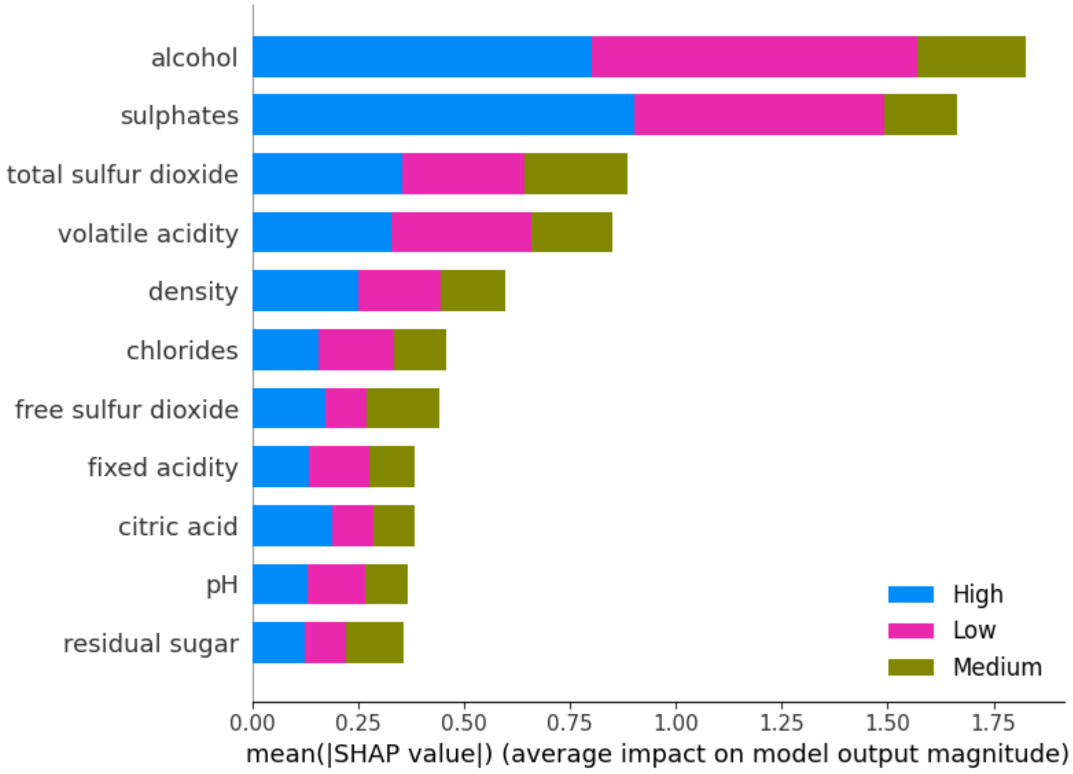
Using the shap package in Python, the following plot is obtained. The log-odds of –1.4 is the output that leads to the prediction of "medium". This is the highest odds among the 3 classes; hence medium is the quality prediction for this datapoint. The most important features that led to the prediction are the high SHAP values of alcohol percentage, and the free sulphur dioxides in the wine. These features pushed the prediction of "medium" to a higher probability, over the other two classes.

Next, an example of a wrong prediction is examined. The predicted quality was Low, but the actual quality is medium. In this case, alcohol and sulphates are the two most important features for this predicted class with low SHAP values. While this gives a low probability, the other classes have an even lower probability, resulting in this class being predicted.



ii. Global SHAP

The global SHAP results provide insights into the overall importance of each feature in the model. This is done by summing up all the individual local SHAP values. Plotting the graph generated by the shap package, we obtain the graph below.



From the results, **alcohol** content is one of the most influential features in determining wine quality, with higher alcohol levels often associated with higher predicted grades. Other significant features include **sulphates**, which generally contribute positively to the quality prediction; and **volatile acidity**, which negatively impacts quality as it increases. Features like **density** and **total sulfur dioxide** also play notable roles, albeit with varying directions of influence depending on their values.

In contrast, features such as **residual sugar** and **chlorides** exhibit relatively lower SHAP values, indicating they have a smaller impact on the model's predictions. These insights not only help in understanding the model's behavior but also align with domain knowledge, where factors like alcohol and acidity are known to be critical in determining wine quality. Such interpretations can guide winemakers in prioritizing adjustments to specific wine properties to achieve desired quality levels.

6. Model Selection

Two different types of models were explored for predicting wine quality - a **Glassbox** model (Shallow Decision Tree) and a **Blackbox** model (LightGBM). These models were evaluated based on their performance in terms of accuracy and their ability to generalize to new data. The Glassbox model provided insights into the decision-making process and classification result with accuracy score of 0.66. The Blackbox model, specifically LightGBM, is a more complex ensemble model that ran multiple iterations allowing the model to make decision based on the defined parameters. This resulted in an accuracy score of 0.72 and confusion matrix provided details on the classification output. Based on the comparison of accuracy scores, the Blackbox (LightGBM) model outperformed the Glassbox (Shallow Decision Tree) model. SHAP analysis on the Blackbox (LightGBM) showed that ‘Alcohol’ and ‘Sulphates’ are key features that impact the predicted class probability. Blackbox model's better performance makes it the optimal choice for predicting wine quality in this case, where maximizing predictive accuracy is the primary objective along with interpretability which is provided through SHAP.

7. Recommendations

Given the above, we recommend exploring the following to further improve the accuracy of the models.

i. **Adjusting the bin ranges**
The current binning approach creates large gap in low (0-5) and high (7-10) which may not be ideal for distinguishing wines in the range of 5 and 7. Additionally, wines with a quality score of 7 may share more in common with wines in the medium range, while wines with a score of 9 or 10 may significantly differ from the rest. Adjusting the bin ranges to address these characteristics may improve the classification accuracy of the model.

ii. **Changing the categories**
Instead of using three broad categories - low, medium and high, categories can be refined to have more granularity. This will enable the model to capture finer distinctions between quality scores. This may reduce the chances of misclassification between adjacent categories.

iii. **Review Feature Correlation**
Although LightGBM is considered to be less sensitive with feature correlation, analyzing feature correlation will be valuable to make the model simpler, faster, and more interpretable. This can be done by running the model with all features and carefully selecting features to remove based on the model performance, feature importance and wine domain knowledge.

8. Conclusion

This study aimed to predict wine quality using two different types of models: a Glassbox model (Shallow Decision Tree) and a Blackbox model (LightGBM). Both models were evaluated based on accuracy and interpretability. To address the lack of transparency, SHAP values were used to interpret the Blackbox model. This allowed the study to better understand the factors influencing wine quality and the behavior of the trained model.

A. Appendices:

Figure 1: Shallow Decision Tree with depth of 3

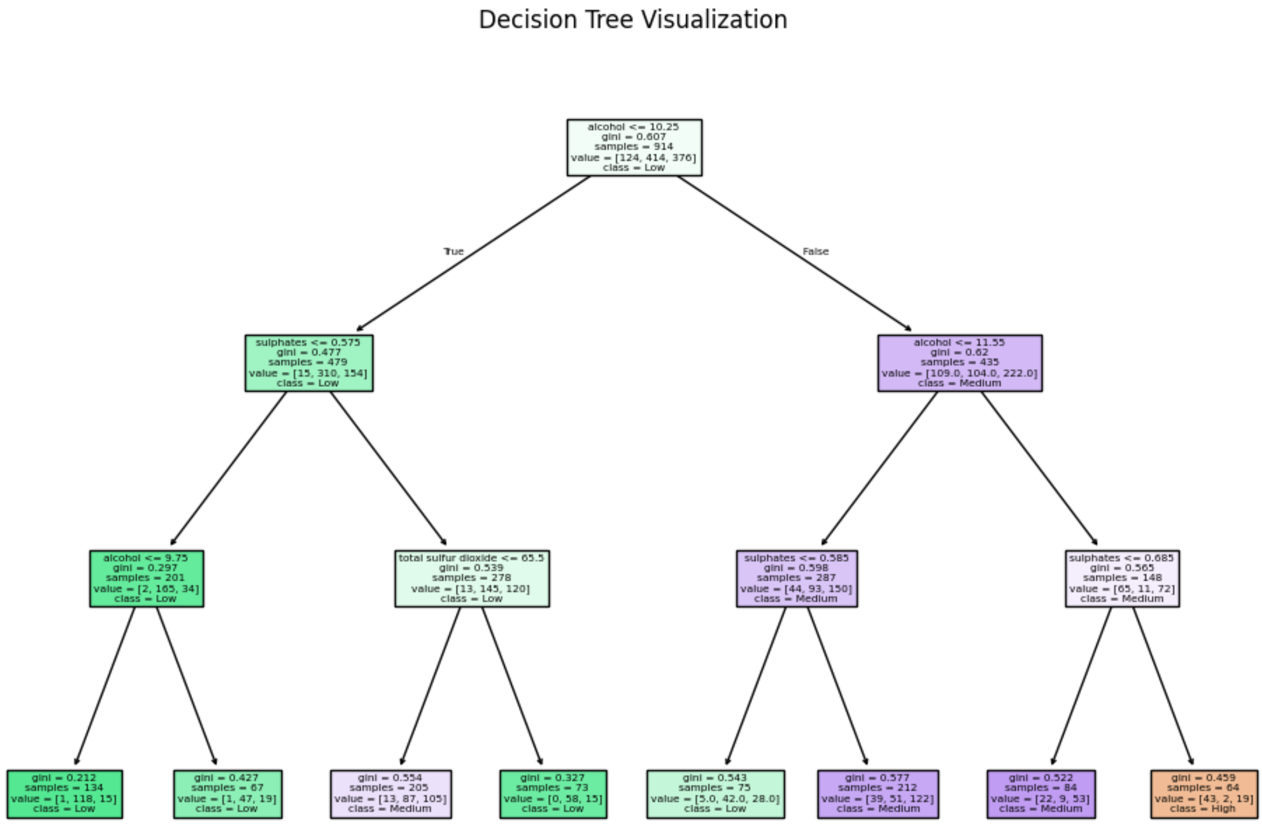
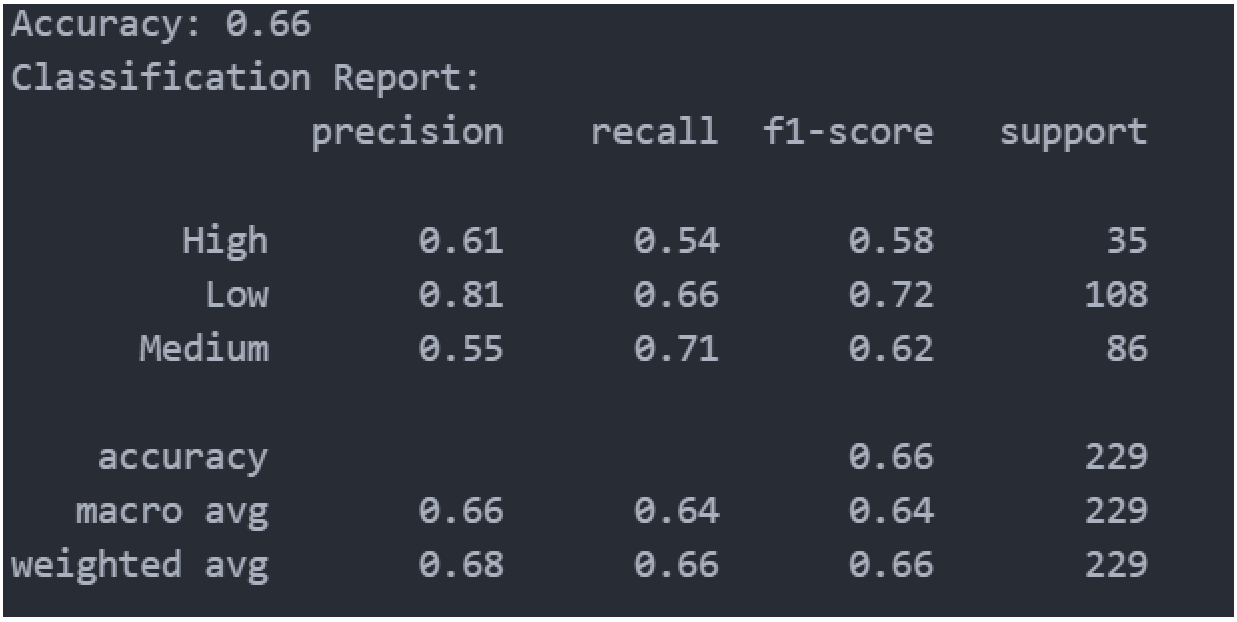

Decision Tree Visualization
alcohol <= 10.25
gini = 0.607
samples = 914
value = [124, 414, 376]
class = Low
True
False
sulphates <= 0.575
gini = 0.477
samples = 479
value = [15, 310, 154]
class = Low
alcohol <= 11.55
gini = 0.62
samples = 435
value = [109.0, 104.0, 222.0]
class = Medium
alcohol <= 9.75
gini = 0.297
samples = 201
value = [2, 165, 34]
class = Low
total sulfur dioxide <= 65.5
gini = 0.539
samples = 278
value = [13, 145, 120]
class = Low
sulphates <= 0.585
gini = 0.598
samples = 287
value = [44, 93, 150]
class = Medium
sulphates <= 0.685
gini = 0.565
samples = 148
value = [65, 11, 72]
class = Medium
gini = 0.212
samples = 134
value = [1, 118, 15]
class = Low
gini = 0.427
samples = 67
value = [1, 47, 19]
class = Low
gini = 0.554
samples = 205
value = [13, 87, 105]
class = Medium
gini = 0.327
samples = 73
value = [0, 58, 15]
class = Low
gini = 0.543
samples = 75
value = [5.0, 42.0, 28.0]
class = Low
gini = 0.577
samples = 212
value = [39, 51, 122]
class = Medium
gini = 0.522
samples = 84
value = [22, 9, 53]
class = Medium
gini = 0.459
samples = 64
value = [43, 2, 19]
class = High

Figure 2. Accuracy of Decision Tree Classification with max_depth = 8

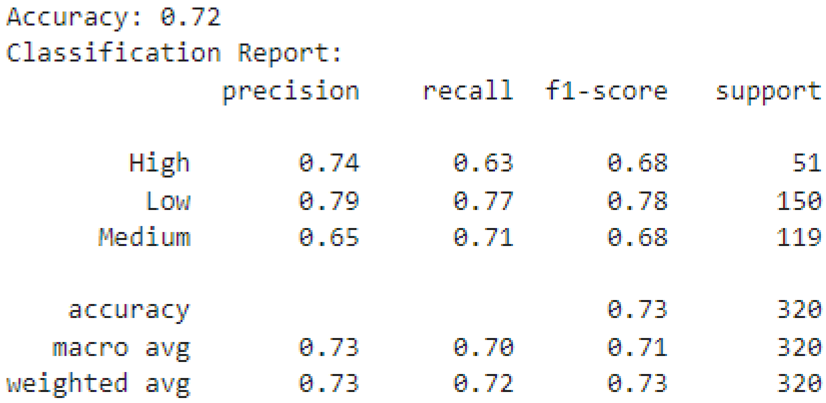


```
Accuracy: 0.66
Classification Report:
              precision    recall  f1-score   support

        High       0.61      0.54      0.58        35
         Low       0.81      0.66      0.72       108
      Medium       0.55      0.71      0.62        86

    accuracy                           0.66       229
   macro avg       0.66      0.64      0.64       229
weighted avg       0.68      0.66      0.66       229
```

Figure 3. LightGBM classification accuracy



```
Accuracy: 0.72
Classification Report:
              precision    recall  f1-score   support

        High       0.74      0.63      0.68        51
         Low       0.79      0.77      0.78       150
      Medium       0.65      0.71      0.68       119

    accuracy                           0.73       320
   macro avg       0.73      0.70      0.71       320
weighted avg       0.73      0.72      0.73       320
```