True connoisseur? Features that Decide the Qualify of a Good White Wine using Machine Learning Techniques

Junjin(Candice) Wang, Vincent Liu, Zhiming Huang

CSE 163

March 11th, 2023

Summary of questions and results

1. The correlation between various features in the feature space.

   (Heatmap of correlation between various features in feature space)(In the visualization section)

   a. Most features of white wines do not have a strong correlation in feature space. Total sulfur dioxide and free sulfur dioxide have a strong positive relationship, while alcohol and density show a significant inverse relationship. Moreover, the strong correlation between the target quality and some features (such as alcohol, density, chlorides, and volatile acidity) suggests that these features might be crucial for prediction.

2. The relationship between different features and wine quality.

   a. Looking at Figure 2 (In the visualization section), we can see that in general sulfates, pH, density, citric acid, and fixed acidity do have not a significant relationship with the quality of the white wine. While the rest features: volatile acidity, residual sugar chlorides, free sulfur dioxide, and total sulfur dioxide have somewhat of an impact on the quality of the white wine's quality.

3. Which features are the most important for predicting the target, wine quality in this case?

   a. According to the code output from our lasso model

      ```
      ([-0.        -0.        -0.         0.        -0.         0.
       -0.        -0.         0.         0.         0.22496738]
      ['alcohol']),
      ```

The feature 'alcohol' is the most important for predicting wine quality. The corresponding coefficient value is 0.22496738, which indicates that this feature has a statistically significant positive effect on wine quality. The other features all have coefficient values of 0 or -0, indicating that they do not have a significant impact on the wine quality prediction in this model.

Motivation

- This is a very interesting topic to us. Most of our group members are under 21 with no prior experience with drinking alcohol. So we were simply out of curiosity to see what factors are crucial when it comes to judging alcohol quality.
- If we can find some relationships between the components of wine, we can judge the wine more professionally, not just by taste.
- Wine quality appreciation can be time and labor-intensive, with an effective machine learning algorithm helping automate, this process can provide significant cost savings and efficiency gains in assessing wine quality and designing supplying and pricing strategies.

Dataset

- We used the dataset which is called Wine Quality Data Set(http://archive.ics.uci.edu/ml/datasets/Wine+Quality). It contains two datasets, including red and white vinho verde wine samples. The goal of this dataset is to model wine quality based on physicochemical tests.
- We specifically focused on the investigation, analysis, and prediction of the White Wine Quality Dataset.

Method

1. The correlation between various features in the feature space.

   Methodology:

   1. Load the white wine quality dataset into a Jupyter notebook.
   2. Clean the dataset by checking for and handling any missing or incorrect data values.
   3. Use the pandas library correlation method to compute a correlation matrix of all the features in the dataset.
   4. Generate a correlation heatmap using the seaborn library to visually represent the correlation matrix.
   5. Analyze the heatmap to identify any strong correlations between features.
   6. Use the correlation coefficients to determine the strength and direction of the correlation between the features.
   7. Draw conclusions about the relationships between the features based on the correlation analysis.

2. The relationship between different features and wine quality.

   Methodology:

   1. Load the white wine quality dataset into a Jupyter notebook.
   2. Clean the dataset by checking for and handling any missing or incorrect data values.

3. Use descriptive statistics, such as mean, median, and standard deviation, to explore the relationships between each feature and wine quality.
4. Generate subplots of histograms to visualize the distribution and relationship between each feature and wine quality.
5. With the result of correlation coefficients, determine if there are statistically significant differences in wine quality between each specific feature.
6. Draw conclusions about the relationships between the features and wine quality based on the combination of descriptive statistics, visualizations, and correlation coefficients.

3. Which features are of the most important for predicting the target, wine quality in this case?

Methodology:

1. Load the white wine quality dataset into a Jupyter notebook.
2. Split the dataset into the feature space and target variable (white wine quality).
3. Train a Lasso Decision Tree model, which combines the Lasso regularization technique with decision tree-based feature selection, on the dataset.
4. Extract the feature importance scores from the trained model.
5. Rank the importance of the features based on their importance score.
6. Draw conclusions about the importance of each feature in predicting wine quality based on the results of the Lasso Decision Tree model, and identify the top features that contribute the most to wine quality prediction.

Results

How the result led us to the conclusion to our first research question(The correlation between various features in the feature space) :

The correlation analysis reveals how strongly the features in the white wine quality dataset are related to each other. The heatmap helps us to highlight areas of high and low correlation between features, with warmer colors indicating a stronger correlation. For example, total sulfur dioxide and free sulfur dioxide have a strong positive relationship, while alcohol and density show a significant inverse relationship. Moreover, the strong correlation between the target quality and some features (such as alcohol, density, chlorides, and volatile acidity) suggests that these features might be crucial for prediction.

Second research question:

From our findings from the combination of descriptive statistics, visualizations, and correlation coefficients, we conclude that in general sulfates, pH, density, citric acid, and fixed acidity do have not a significant relationship with the quality of the white wine. While the rest features– volatile acidity, residual sugar chlorides, free sulfur dioxide, and total sulfur dioxide– have somewhat of an impact on the quality of the white wine's quality. We also conclude that sulfates and alcohol content are positive drives to the white wine quality. Volatile acidity, on the other hand, has a strong negative association with white wine quality, which makes sense as quality tests favor a lower amount of acidity.

Third question:

By using the Lasso Decision Tree model, we identified the feature 'alcohol' being the most important for predicting wine quality. The corresponding coefficient value is 0.22496738, which indicates that this feature has a statistically significant positive effect on wine quality. The other features all have coefficient values of 0 or -0, indicating that they do not have a significant impact on the wine quality prediction in this model. We then used the Lasso Decision Tree model selected features to retrain the model and compared its performance with the baseline and the regular decision tree model.

Interpret the results

Our analysis's findings indicate that several characteristics in the white wine quality dataset are significantly correlated with one another and affect wine quality. For instance, we discovered that volatile acidity had a detrimental influence on wine quality whereas sulfates and alcohol concentration have good effects. These results provide confidence to our study since they are consistent with existing knowledge in the field and make intuitive sense.

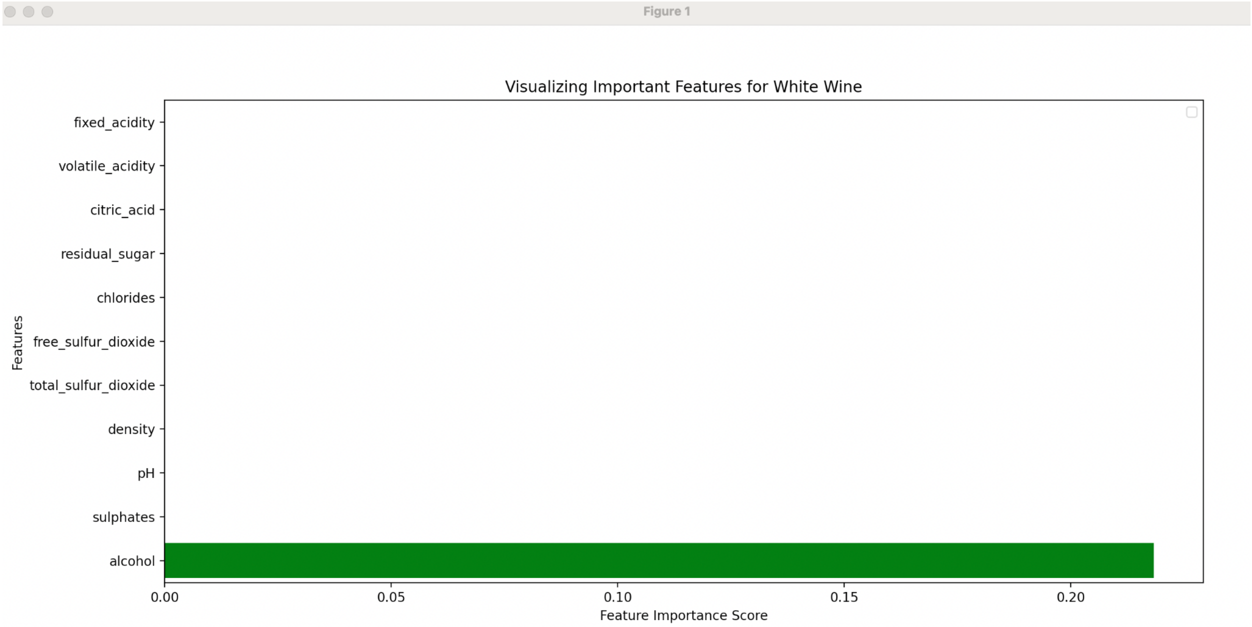
Also, we determined that alcohol concentration was crucial for predicting wine quality using the Lasso Decision Tree model. This result is consistent with our earlier discovery that there is a link between wine quality and alcohol concentration.

It may seem unexpected at first to learn that the only variable with a detectable effect on the quality of white wine is alcohol. It's crucial to bear in mind, though, that this outcome is predicated on a particular model, notably the Lasso Decision Tree model. By punishing the coefficients of less crucial features and thereby decreasing them to zero, this model is intended to select the most critical features for prediction.

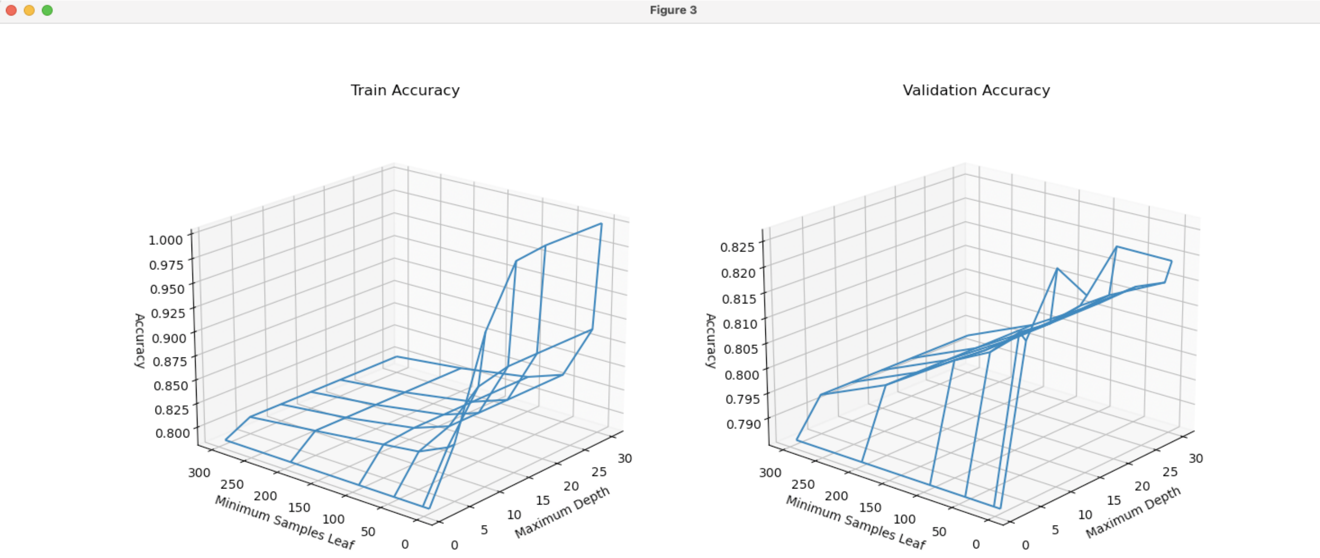
Furthermore, it's possible that the other parameters in the dataset are still crucial for forecasting wine quality but that the Lasso Decision Tree model is unable to recognize their relevance because to how complicated or nonlinear their connection is.

Also, it's crucial to note that although while alcohol plays the largest role in this model's ability to predict white wine quality, that doesn't imply other factors are necessarily unimportant. The quality of white wine may also be significantly influenced by additional elements such as grape varietal, winemaking processes, and storage conditions.

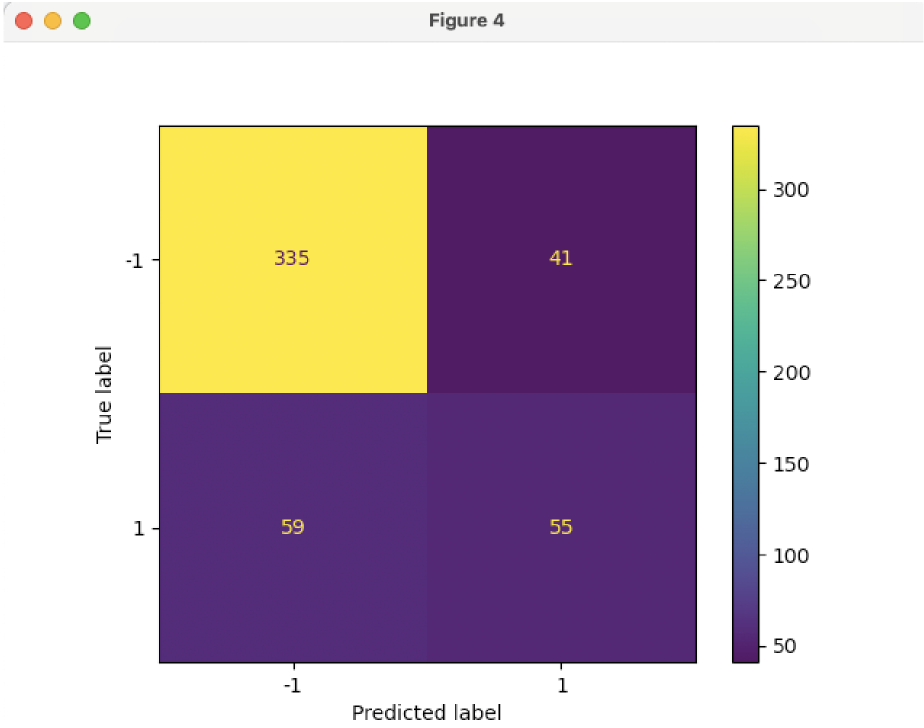
Visualizations



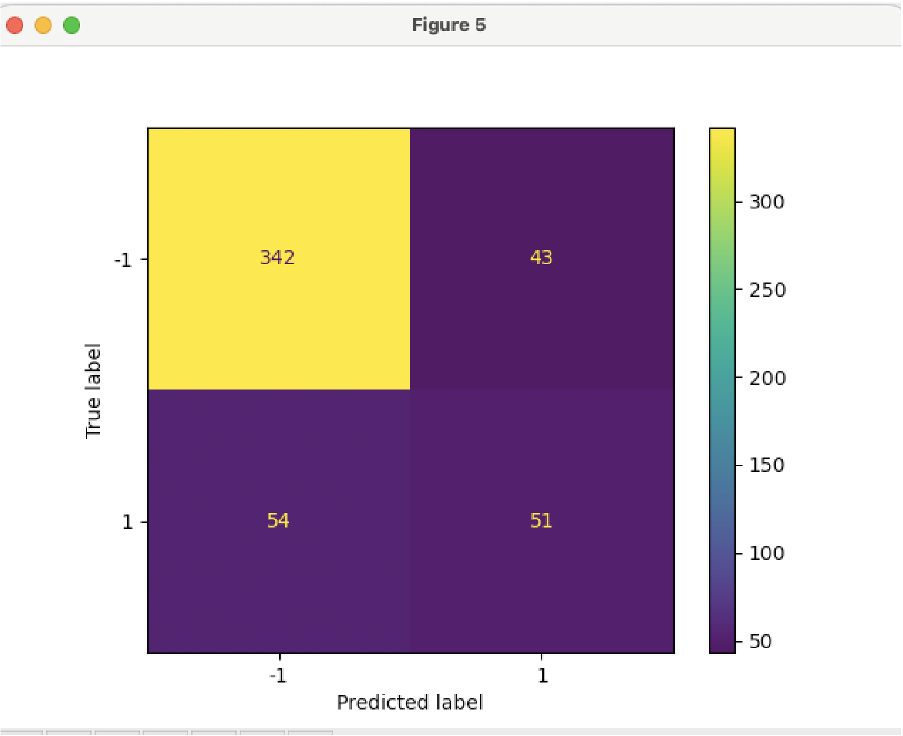
(picture of visualizing important features for white wine)



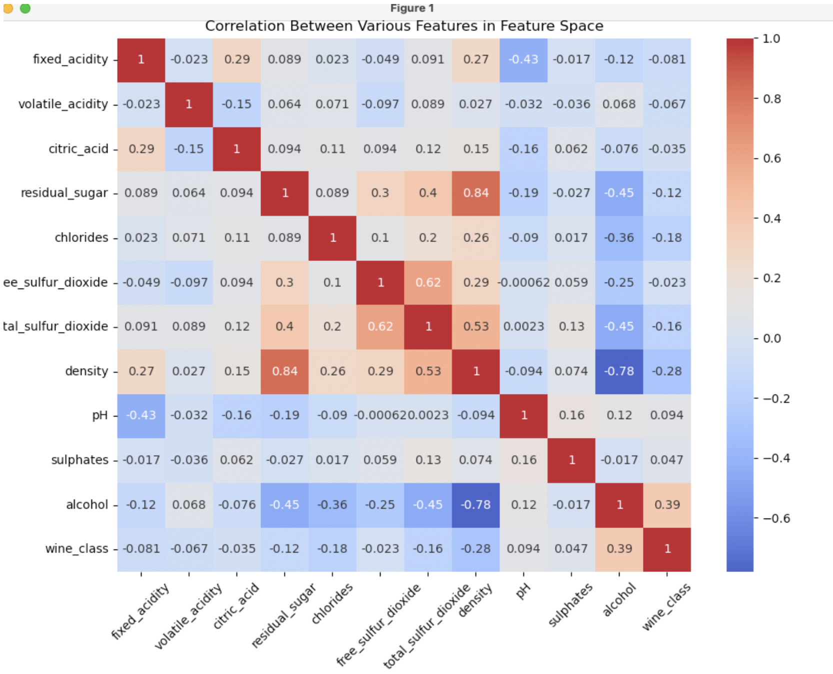
(picture of train accuracy and validation accuracy)



(the picture of validation confusion matrix)



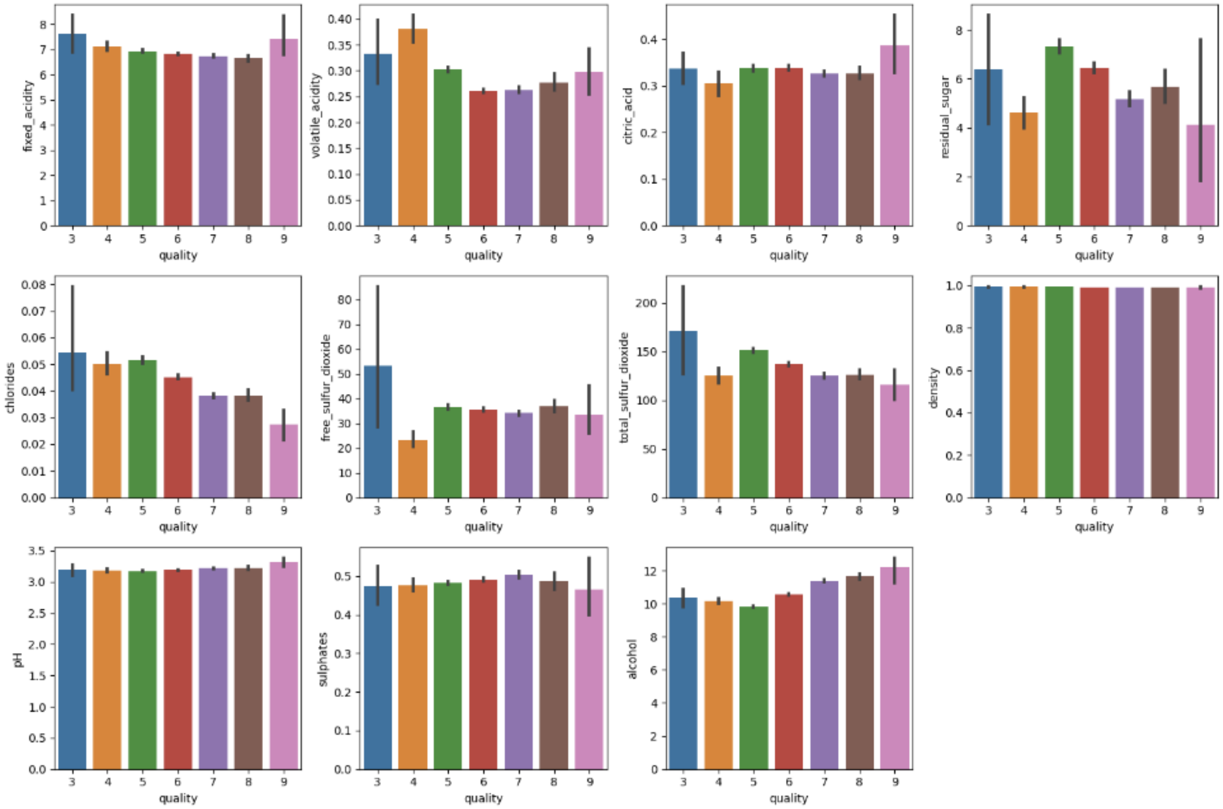
(The picture of test confusion matrix)



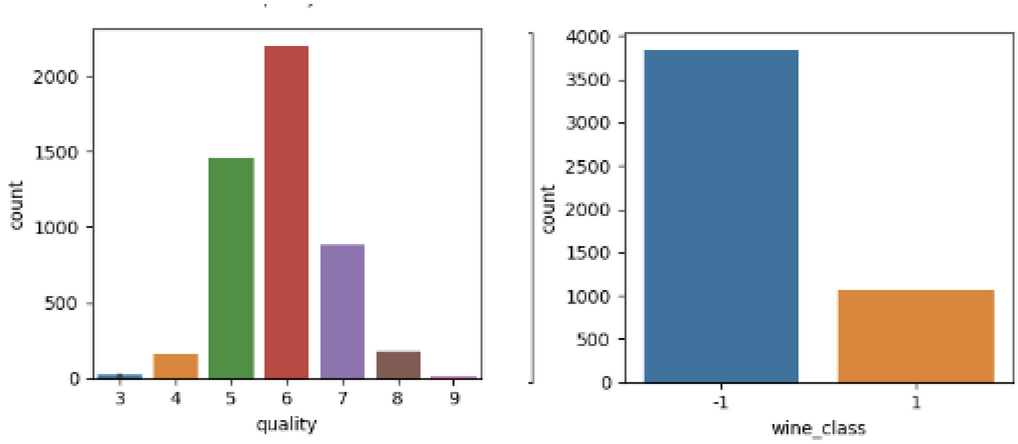
Correlation Between Various Features in Feature Space

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	wine_class
fixed_acidity	1	-0.023	0.29	0.089	0.023	-0.049	0.091	0.27	-0.43	-0.017	-0.12	-0.081
volatile_acidity	-0.023	1	-0.15	0.064	0.071	-0.097	0.089	0.027	-0.032	-0.036	0.068	-0.067
citric_acid	0.29	-0.15	1	0.094	0.11	0.094	0.12	0.15	-0.16	0.062	-0.076	-0.035
residual_sugar	0.089	0.064	0.094	1	0.089	0.3	0.4	0.84	-0.19	-0.027	-0.45	-0.12
chlorides	0.023	0.071	0.11	0.089	1	0.1	0.2	0.26	-0.09	0.017	-0.36	-0.18
ee_sulfur_dioxide	-0.049	-0.097	0.094	0.3	0.1	1	0.62	0.29	-0.00062	0.059	-0.25	-0.023
tal_sulfur_dioxide	0.091	0.089	0.12	0.4	0.2	0.62	1	0.53	0.0023	0.13	-0.45	-0.16
density	0.27	0.027	0.15	0.84	0.26	0.29	0.53	1	-0.094	0.074	-0.78	-0.28
pH	-0.43	-0.032	-0.16	-0.19	-0.09	-0.00062	0.0023	-0.094	1	0.16	0.12	0.094
sulphates	-0.017	-0.036	0.062	-0.027	0.017	0.059	0.13	0.074	0.16	1	-0.017	0.047
alcohol	-0.12	0.068	-0.076	-0.45	-0.36	-0.25	-0.45	-0.78	0.12	-0.017	1	0.39
wine_class	-0.081	-0.067	-0.035	-0.12	-0.18	-0.023	-0.16	-0.28	0.094	0.047	0.39	1

(the picture of the correlation between various features in feature space)



(Subplots of the relationship between white wine quality and each feature)



(target count plot & balanced target count plot)

Impact and Limitations

Our findings may be utilized to predict wine quality with a high degree of accuracy using the features chosen for the Lasso Decision Tree model. We think this will help wine producers and merchants improve the quality of their products and increase customer satisfaction. Customers may benefit from making better-informed wine purchase decisions.

Nonetheless, there may be some biases and inaccuracies in our analysis. Because the wine quality dataset only includes white wine, our results might not be applicable to red wine or other varieties of wine. The dataset is not representative of other white wines because it only includes samples from one region of Portugal. This can bias our results and diminish their generalizability. Also, we think that our dataset may not include all of the variables that influence wine quality. For instance, the wine's quality may be impacted by its production process, storage environment, and aging, but our study does not take these elements into consideration.

Therefore, while our results show that the Lasso Decision Tree model can be effective in predicting wine quality, it is also crucial to be aware of its limitations. Users should use caution when extending our findings to different situations or wine varieties without additional verification.

Challenge Goals

1. The value of each feature in the dataset is not balanced. For example, the value of chlorides is 0.045, and the value of fixed acidity is 7. Therefore, these two values are imbalanced. We therefore used MinMaxScaler to normalize all values between the range 0 to 1. By scaling the feature values to a standard range (usually between 0 and 1), we prevented features with large values from dominating the model training process and helped the model in learning more meaningful relationships between the features and the target wine quality variable.
2. The Lasso Decision Tree model was challenging for us as we had never used this model before. However, we used it to select the most important features in the wine quality dataset and re-trained the model with these selected features. The performance of this model was then compared with that of a standard decision tree model and the baseline model. This complex and challenging Lasso Decision Tree Model approach proved to be worth it for us. By focusing on the most essential features, we were able to streamline the model and improve its functionality. Additionally, it enhanced our knowledge of the relationships between the different characteristics and how they impact wine quality. When we compared the Lasso Decision Tree model's performance to that of the baseline model and the traditional decision tree model, we could see that it did better than the baseline model and was on par with the conventional decision tree model.

Work Plan Evaluation

Compare to the original work plan, we have followed the majority of the plan. One of the big changes was: originally we want to analyze both red wine and white wine data. However, during the process, we figured out that the data set have the same columns. And all we need is to change the dataset to redwine.csv then we can do data analysis for the red wine. So instead of writing redundant code, we only focused on the white wine for this project. Other than that, we followed our original plan. The result was ideal.

Testing

Since the majority of our code was on graph computing and machine learning, there is not really a good way to test these things. So we have decided to make sure the data processing phase was all correct. In the testing file, we have double-checked that the CSV files have no null values, and also the spilt on the quality worked as expected. We have created another dataset to test our files to assure our code was correct. We also implemented the assert_equals from the cse163_utils python file for easy comparison of the anticipated result with the outcome. With the clean and well-processed dataset, we are confident that our code has the anticipated result.

Collaboration

During the process, we were lucky to have a friend who has a lot of experience in data analysis and machine learning. So whenever we got stuck on the code, we will talk to our friend for some help. He will explain to us what went wrong and how to fix it. The process was very good practice for learning python and general coding and debugging.