THE OHIO STATE UNIVERSITY

HackAI 2023 Challenge: Wine Quality

Technical Report

Team: Maverick

Mayur Patil
Jaeha Lee
Jonathan

Problem Description

This year, AI CLub provided whe Wine Quality Dataset for the AI Club challenge. The goal of this challenge is to build a machine-learning model that predicts the quality of the wine based on various input features.

Dataset

The training and testing dataset is shown in the figures below:

	Id	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	0	8.0	0.50	0.39	2.20	0.073	30.0	39.0	0.99572	3.33	0.77	12.1	6
1	1	9.3	0.30	0.73	2.30	0.092	30.0	67.0	0.99854	3.32	0.67	12.8	6
2	2	7.1	0.51	0.03	2.10	0.059	3.0	12.0	0.99660	3.52	0.73	11.3	7
3	3	8.1	0.87	0.22	2.60	0.084	11.0	65.0	0.99730	3.20	0.53	9.8	5
4	4	8.5	0.36	0.30	2.30	0.079	10.0	45.0	0.99444	3.20	1.36	9.5	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2051	2051	6.6	0.31	0.13	2.00	0.056	29.0	42.0	0.99388	3.52	0.87	12.0	7
2052	2052	9.7	0.59	0.21	1.80	0.079	27.0	65.0	0.99745	3.14	0.58	9.4	5
2053	2053	7.7	0.43	0.42	1.70	0.071	19.0	37.0	0.99258	3.32	0.77	12.5	8
2054	2054	9.1	0.50	0.00	1.75	0.058	5.0	13.0	0.99670	3.22	0.42	9.5	5
2055	2055	6.2	0.31	0.18	2.30	0.059	12.0	28.0	0.99520	3.56	0.88	11.4	7

2056 rows × 13 columns

Figure 1: Snippet showing training set

	Id	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
0	2056	7.2	0.510	0.01	2.0	0.077	31.0	54.0	0.99748	3.39	0.59	9.8
1	2057	7.2	0.755	0.15	2.0	0.102	14.0	35.0	0.99586	3.33	0.68	10.0
2	2058	8.4	0.460	0.40	2.0	0.065	21.0	50.0	0.99774	3.08	0.65	9.5
3	2059	8.0	0.470	0.40	1.8	0.056	14.0	25.0	0.99480	3.30	0.65	11.7
4	2060	6.5	0.340	0.32	2.1	0.044	8.0	94.0	0.99356	3.23	0.48	12.8
...	...	...	...	...	...	...	...	...	...	...	...	...
1367	3423	8.8	0.745	0.18	2.7	0.084	41.0	115.0	0.99823	3.38	0.70	9.8
1368	3424	15.6	0.240	0.55	2.9	0.062	11.0	25.0	0.99724	2.99	0.77	10.1
1369	3425	7.3	0.760	0.00	2.2	0.095	6.0	19.0	0.99880	3.67	0.60	9.4
1370	3426	7.6	0.780	0.26	2.6	0.118	17.0	104.0	0.99616	3.30	0.53	9.9
1371	3427	8.8	0.780	0.22	3.0	0.085	21.0	43.0	0.99790	3.37	0.56	10.2

1372 rows × 12 columns

Figure 2: Snippet showing testing set

Data Analysis

Analysing the data by visualizing the correlation matrix using the corr() method. The resulting matrix contains the pairwise correlations between all the features in the dataset using heatmap() function from seaborn library to plot the correlation matrix. The resulting plot shows the strength and direction of the pairwise correlations between the features in the dataset. The correlation plot is shown below:

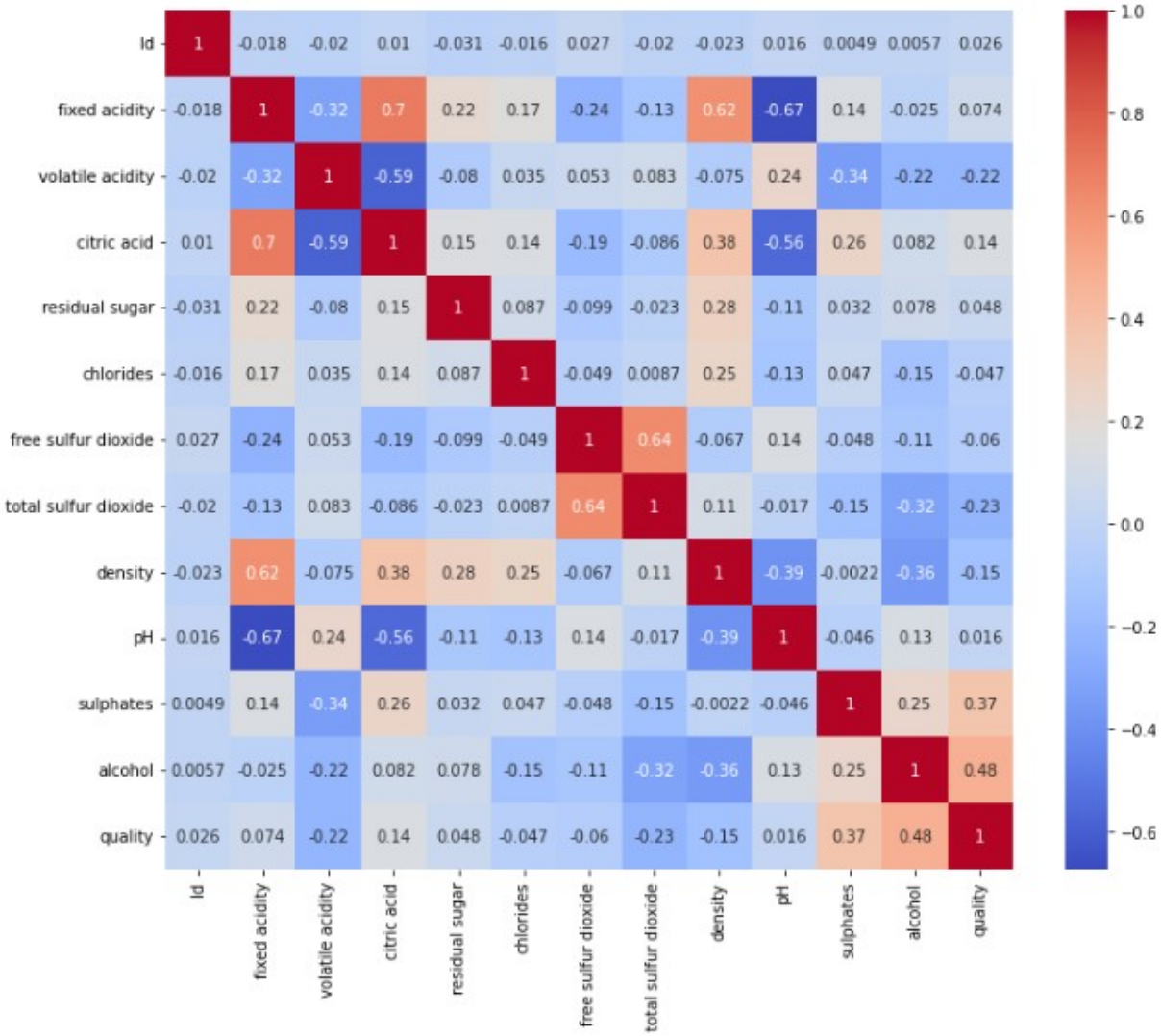

Figure 3: Correlation Heatmap

Based on the heatmap, here are some inferences we can make about the correlations between the features:

1. Quality has the strongest positive correlation with alcohol, and a moderate positive correlation with volatile acidity, sulphates, and citric acid. It has a moderate negative correlation with total acidity, and a weak negative correlation with density and fixed acidity. These correlations suggest that higher levels of alcohol, volatile acidity, sulphates, and citric acid are associated with higher quality wine, while higher levels of total acidity, density, and fixed acidity are associated with lower quality wine.

2. Alcohol has a strong positive correlation with density, and a moderate positive correlation with residual sugar, pH, and sulphates. It has a weak positive correlation with total sulfur dioxide, free sulfur dioxide, and citric acid. These correlations suggest that higher levels of alcohol are

associated with higher density and residual sugar, while lower pH and higher sulphates may also contribute to higher alcohol levels.

3. Density has a strong positive correlation with residual sugar, and a moderate positive correlation with pH, fixed acidity, and total sulfur dioxide. It has a weak positive correlation with free sulfur dioxide and a weak negative correlation with citric acid. These correlations suggest that higher density is associated with higher levels of residual sugar, pH, fixed acidity, and total sulfur dioxide.

4. Residual sugar has a moderate positive correlation with pH and total sulfur dioxide, and a weak positive correlation with free sulfur dioxide and citric acid. These correlations suggest that higher levels of residual sugar may be associated with higher pH and total sulfur dioxide.

5. Total sulfur dioxide has a strong positive correlation with free sulfur dioxide, and a moderate positive correlation with pH and citric acid. These correlations suggest that higher levels of total sulfur dioxide are associated with higher levels of free sulfur dioxide, and that pH and citric acid may also contribute to higher levels of total sulfur dioxide.

6. Free sulfur dioxide has a weak positive correlation with citric acid and a weak negative correlation with pH. These correlations suggest that higher levels of free sulfur dioxide may be associated with higher levels of citric acid, and that lower pH may also contribute to higher levels of free sulfur dioxide.

7. pH has a weak negative correlation with fixed acidity, and a weak positive correlation with citric acid. These correlations suggest that lower pH may be associated with higher levels of fixed acidity and lower levels of citric acid.

The plot showing multiple pairwise bivariate distributions in the dataset is shown below:

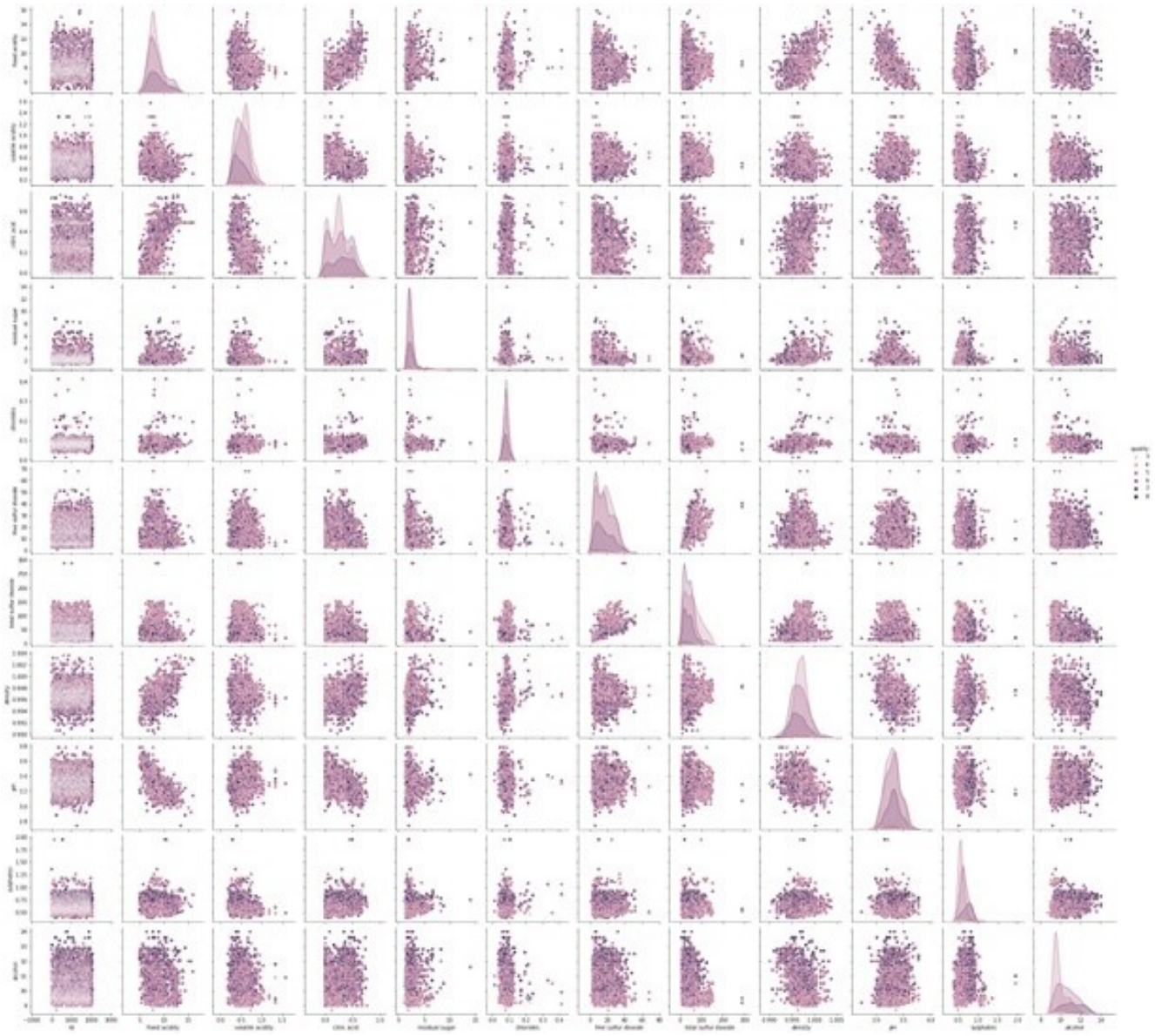

Figure 4: Correlation Heatmap

Pre-processing

The pre-processing step is performed for:
1. Checking any outliers and dropping/substituting N/A values.
2. Using scaler technique to standardize the data features within a fixed range.
The boxplot approach was adopted to analyse and remove the outliers as shown in the figure below and the missing values were dropped from the dataset.

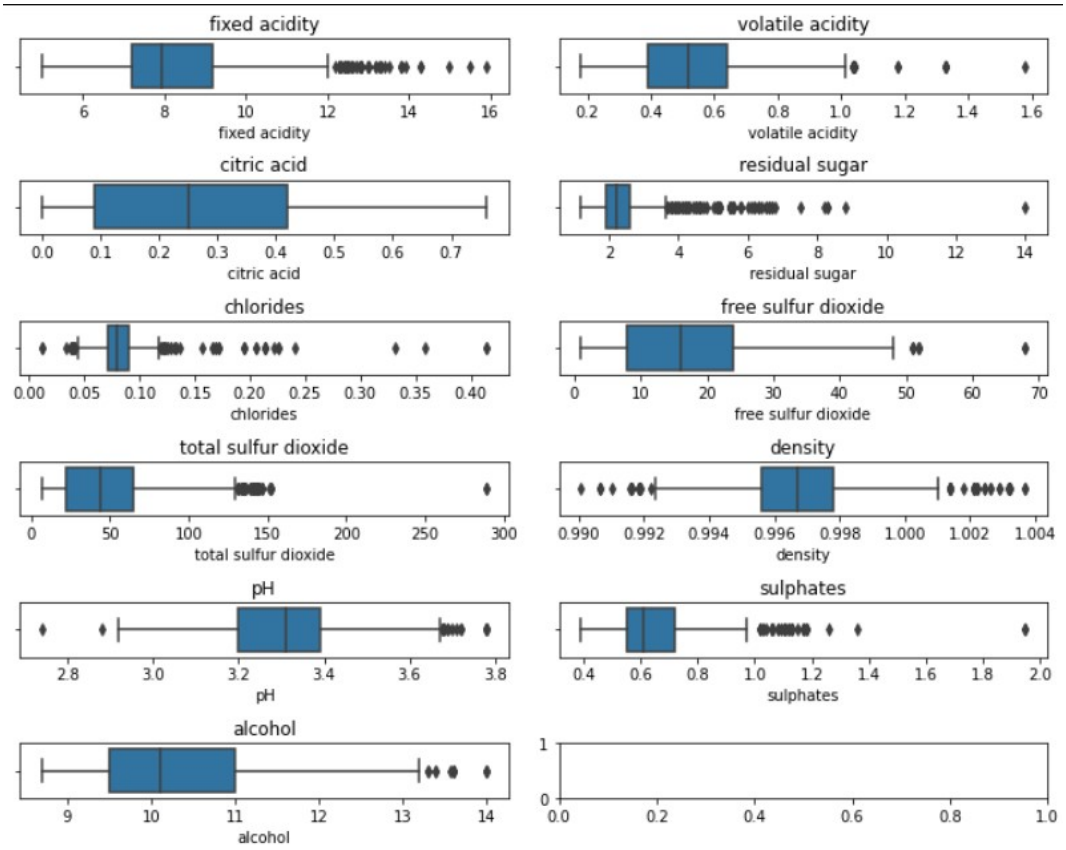

Figure 5: Boxplot

Based on the box plots, here are some inferences we can make about the distribution of the features in the dataset:
1. Fixed acidity has a skewed distribution with many outliers on the upper end of the distribution.
2. Volatile acidity has a skewed distribution with a moderate outliers on the upper end of the distribution.
3. Citric acid has a moderately skewed distribution with no outliers.
4. Residual sugar has a very skewed distribution with a large number of outliers on the upper end of the distribution.
5. Chlorides has a moderately skewed distribution with a large numner of outliers on the upper end of the distribution.
6. Free sulfur dioxide has a moderately skewed distribution with a few outliers on the upper end of the distribution.
7. Total sulfur dioxide has a moderately skewed distribution with a large number of outliers on the upper end of the distribution.
8. Density has a moderately skewed distribution with a few outliers on the both sides of the distribution.
9. pH has a symmetric distribution with a few outliers on both sides of the distribution.
10. Sulphates has a skewed distribution with a large number of outliers on the upper end of the distribution.
11. Alcohol has a skewed distribution with a few outliers on the upper end of the distribution.

Feature Engineering

The following augmentation were performed:
The total acidity feature was added that may capture important information about the overall acidity of the wine that is not fully captured by the fixed acidity and volatile acidity features separately. The sweetness feature was added that may capture information about the perceived level of sweetness of the wine, which may be different from the actual residual sugar content due to the influence of other factors such as acidity and bitterness. And finally, the total sulfur dioxide feature was added that may capture important information about the stability and shelf-life of the wine, which may be different from the free sulfur dioxide and total sulfur dioxide features separately.

Hyperparameter Tuning

We used grid search to perform hyperparameter tuning for the RandomForestClassifier. Grid search is a technique that allows us to systematically search for the best hyperparameters for a machine learning algorithm by testing all possible combinations of hyperparameters in a predefined range. The Random Forest Classifier is often used for classification tasks. It is a type of ensemble learning method that combines multiple decision trees to make a final prediction. There are several reasons why we might choose to use Random Forest Classifier for a given classification task:
1. **High accuracy**: Random Forest Classifier is known for its high accuracy, especially when compared to other algorithms such as logistic regression or decision trees. This is because it is able to combine the predictions of multiple decision trees to reduce overfitting and improve generalization performance.
2. **Robustness to outliers**: Random Forest Classifier is relatively robust to outliers and noise in the data, which can be an important consideration when working with real-world data.
3. **Handles large datasets**: Random Forest Classifier can handle large datasets with many features, making it suitable for many real-world applications.
4. **Provides feature importance**: Random Forest Classifier provides a measure of feature importance, which can be useful for understanding which features are most relevant to the classification task.

Then, we used GridSearchCV to perform a grid search on the hyperparameter space followed by a 5-fold cross-validation using the cv argument to evaluate the performance of each combination of hyperparameters. The best parameters obtained were the following:
Best hyperparameters: 'maxDepth': 10, 'minSamplesLeaf': 4, 'nEstimators': 300

Performance

The overall predicted labels showed an accuracy of **52.44%** on the test set.

References

1. Horning, N. (2010). Random Forests : An algorithm for image classification and generation of continuous fields data sets.
2. https://www.mygreatlearning.com/blog/gridsearchcv/