Predicting Vinho Verde White Wine Quality with ML Regression Model

Presented by,
Harsh.

BACKGROUND

Around 4898 observations related to Vinho Verde White Wine samples from the north of Portugal were collected.

For dataset, the fields are:

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol
- quality

Objective

To predict the white wine quality, given the measurements fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfer dioxide, density, pH, sulphates, alcohol, quality.

Data description

RangeIndex: 4898 entries, 0 to 4897
Data columns (total 12 columns):

1. fixed acidity 4898 non-null float64
2. volatile acidity 4898 non-null float64
3. citric acid 4898 non-null float64
4. residual sugar 4898 non-null float64
5. chlorides 4898 non-null float64
6. free sulfur dioxide 4898 non-null float64
7. total sulfur dioxide 4898 non-null float64
8. density 4898 non-null float64
9. pH 4898 non-null float64
10. sulphates 4898 non-null float64
11. alcohol 4898 non-null float64
12. quality 4898 non-null int64

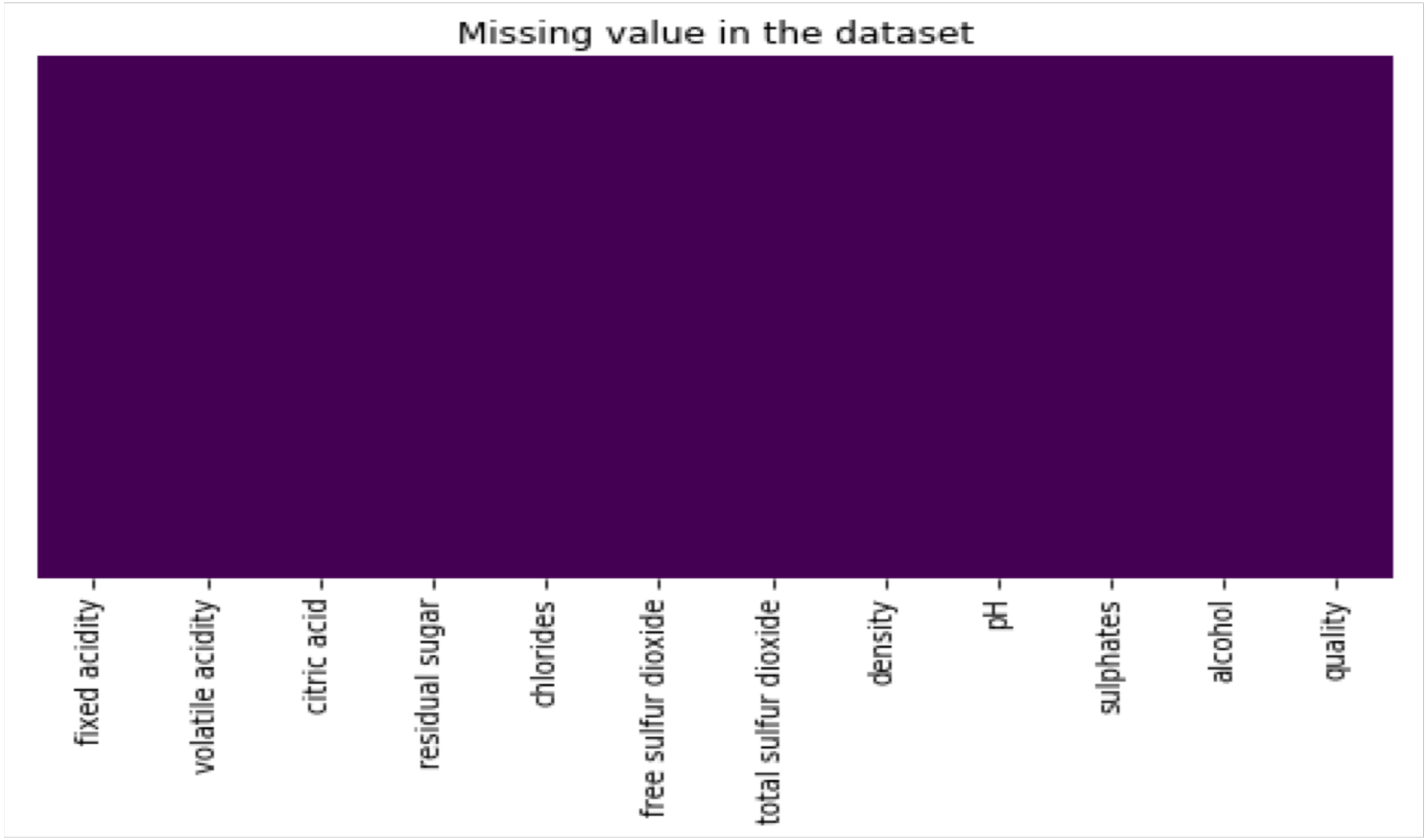
Dependent & Independent Variables

Dependent variable –Wine quality which has 7 Sub classes.

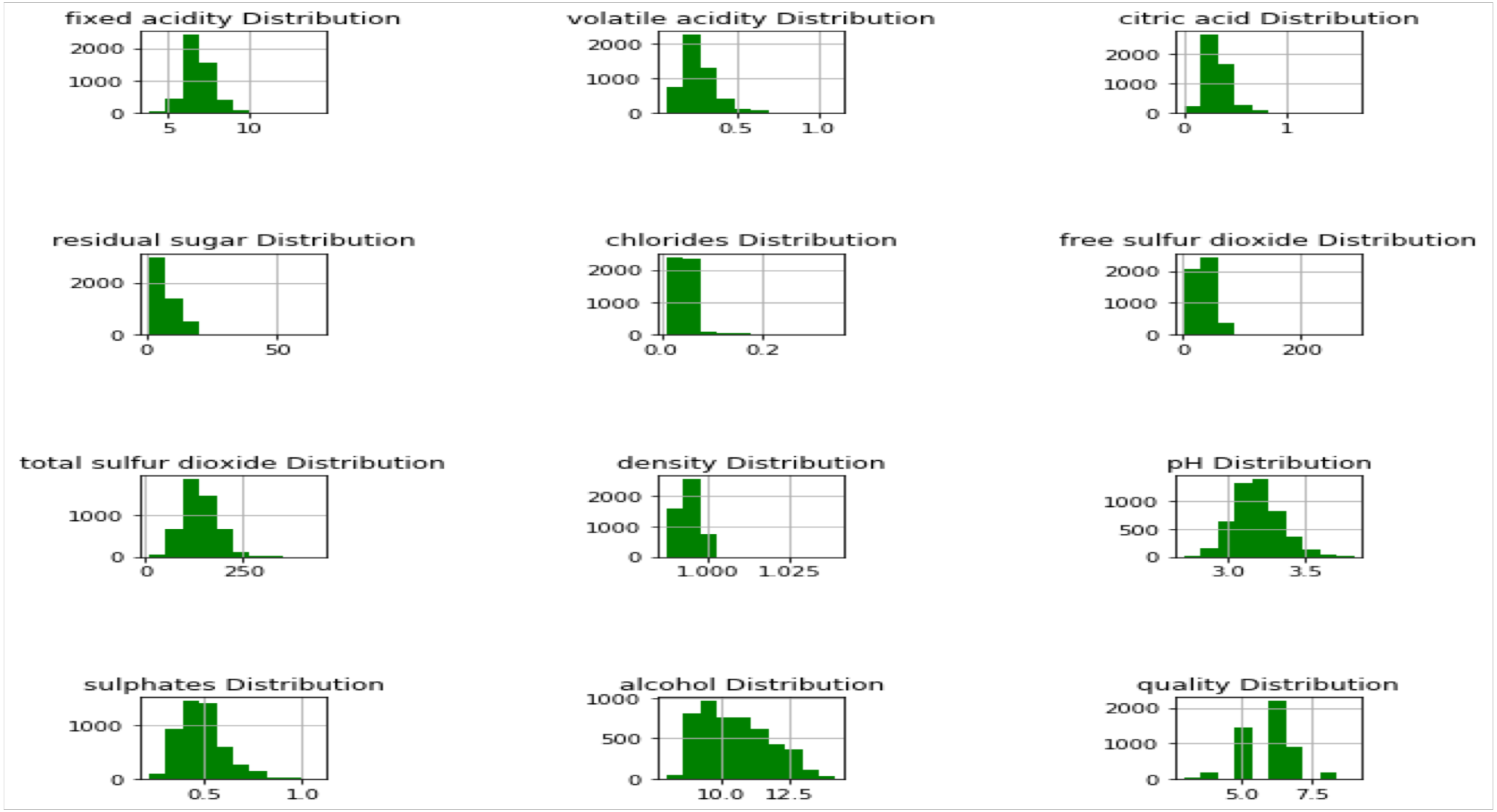
S.No	Quality class	Frequency
1	3	20
2	4	163
3	5	1457
4	6	2198
5	7	880
6	8	175
7	9	5

Independent variables- fixed acidity,volatile acidity ,citric acid ,residual sugar,chlorides,free sulfur dioxide,total sulfur dioxide,density,pH, sulphates,alcohol

#Checking for Missing values

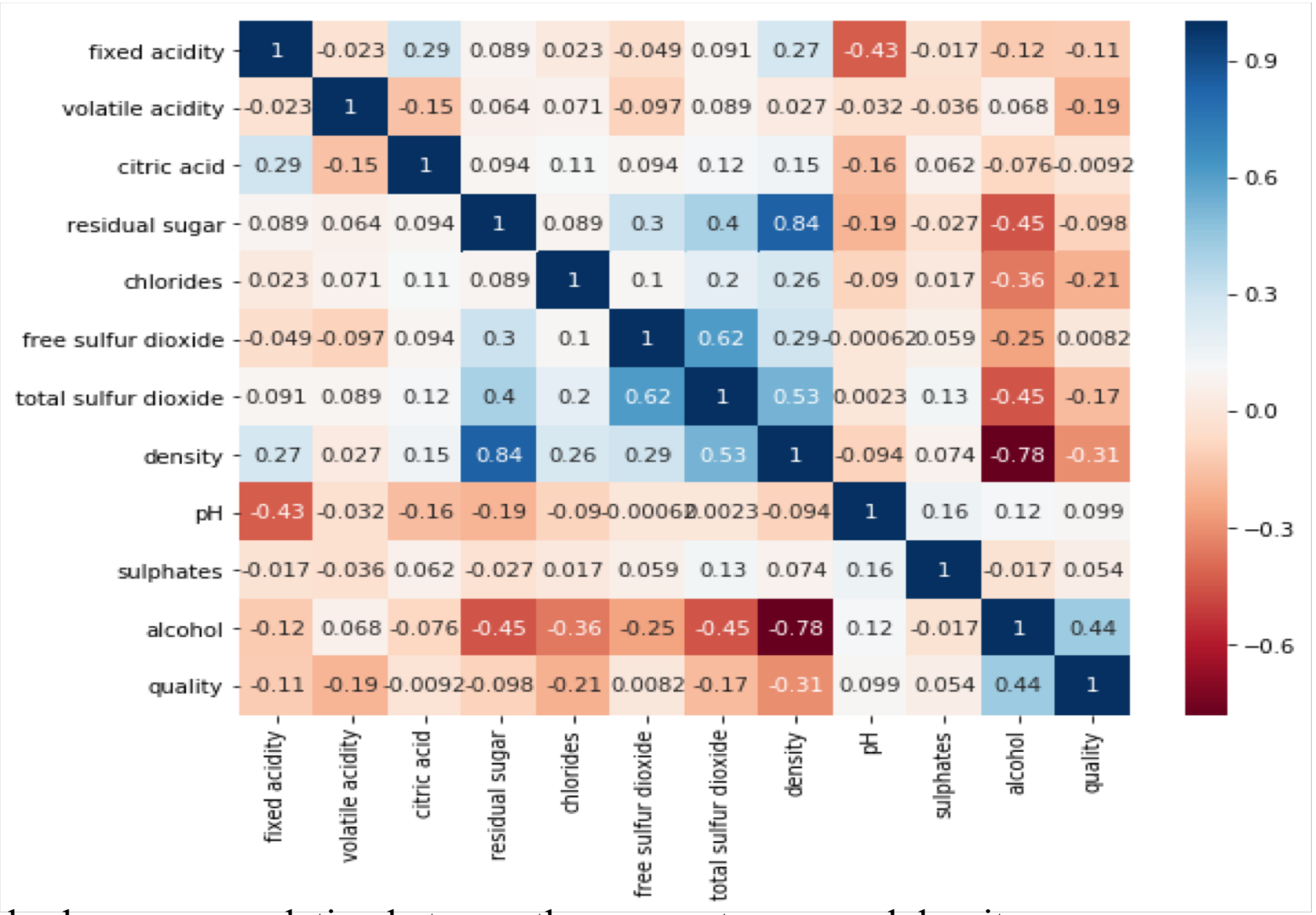


DESCRIBING FEATURE DISTRIBUTION



Above Graphs shows that ph is normally distributed and acidity are right skewed.

Pearson correlation heatmap (matrix) was plotted for correlation.



mostly variables have no correlation between them accept sugar and density.

#K-fold Cross Validation

- For k fold validation we have split data into 80:20 train and test and then we have trained our model with linear regression model,
- K- fold validation has been performed with 10 folds for which average accuracy score is 0.53

Confusion Matrix

$$
\begin{bmatrix} 0 & 0 & 2 & 2 & 0 & 1 \\ 0 & 0 & 13 & 12 & 0 & 0 \\ 0 & 0 & 144 & 146 & 1 & 0 \\ 0 & 0 & 80 & 349 & 3 & 0 \\ 0 & 0 & 7 & 173 & 12 & 0 \\ 0 & 0 & 1 & 30 & 4 & 0 \end{bmatrix}
$$

Here we can infer that all the values on a red line diagonal are true predicted values which are in total high in numbers.

Conclusion

- Model has accuracy score of 0.53 which means that there is 53% chance of predicting accurate quality of white wine with this model.
- As per confusion matrix we can see that true positive values on diagonal are quite high shows high number of true predictions.
- This model is not good for prediction of quality of wine but with feature engineering and using other models with hyperparameter tuning we can increase the accuracy of the model.

Thank you!