Wine Quality Classification

Bishal Adhikari
Rijan Pokhrel
Nirajan Bekoju
Manoj Khatri

Topics of discussion

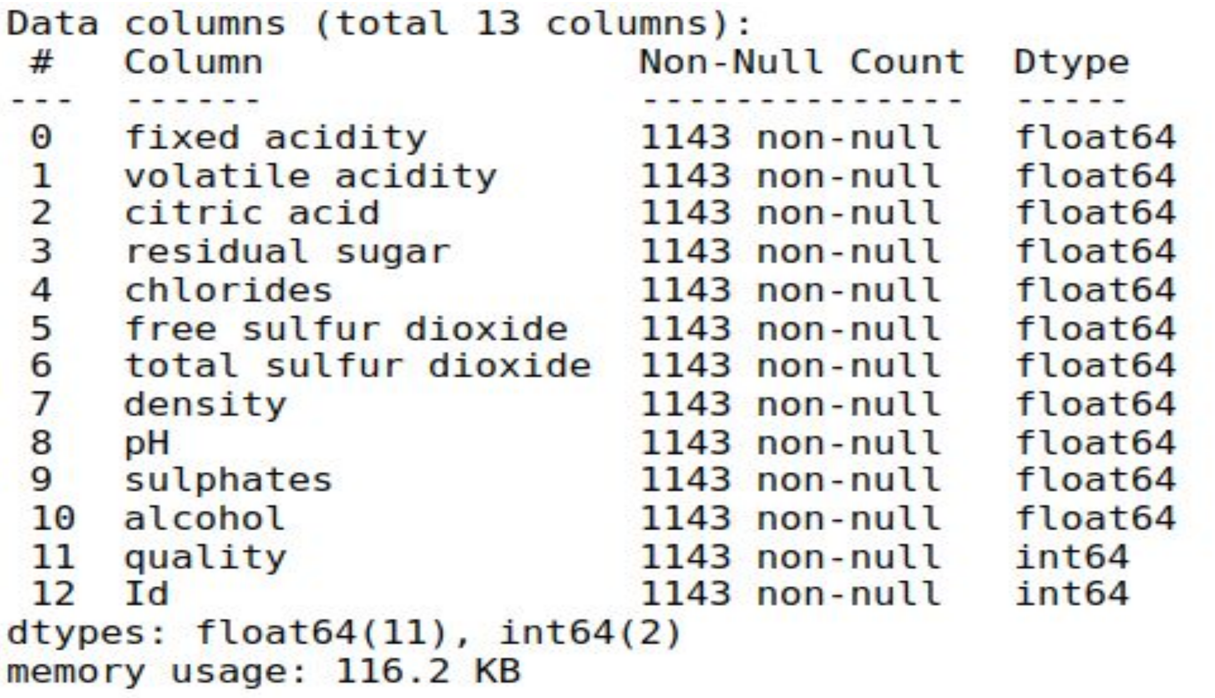
- Problem Statement
- Dataset Description
- Exploratory Data Analysis (EDA)
- Imbalance Data Handling
- Model Development
- Model Evaluation
- Experiment Tracking
- Tools and Technologies

Problem Statement

- The dataset describes the amount of various chemicals present in wine and their effect on it’s quality.
- The datasets can be viewed as classification or regression tasks.
- The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones)
- The complexity arises due to the fact that the dataset has fewer samples, & is highly imbalanced

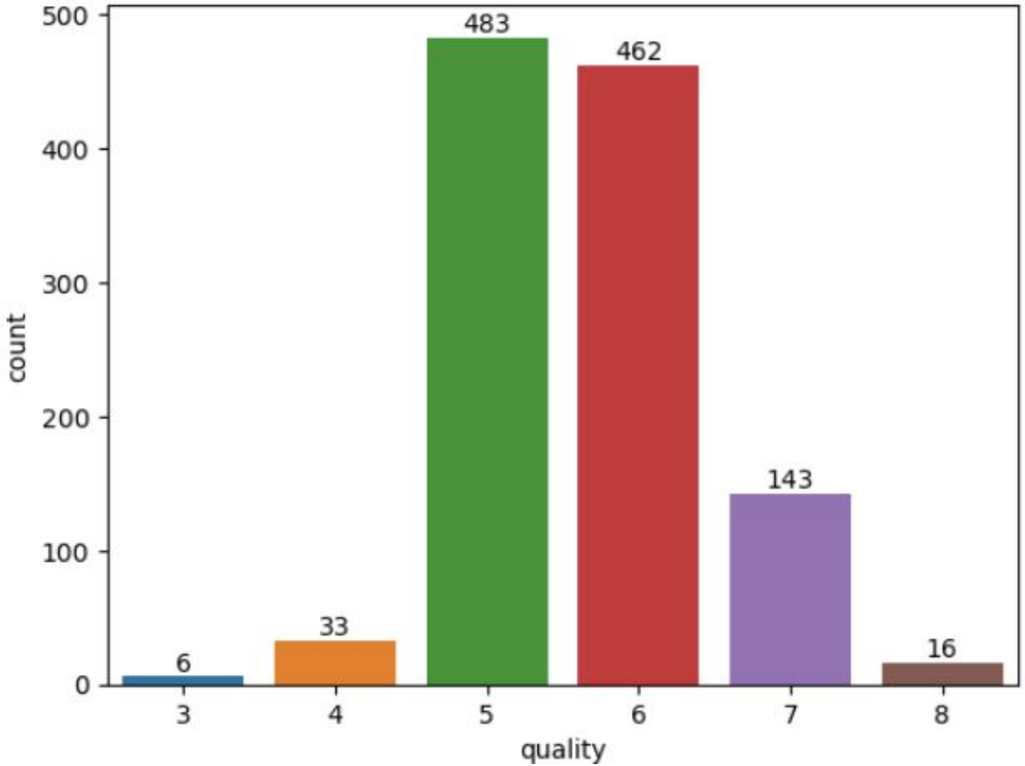
Dataset Description

- The dataset is related to red variants of the Portuguese ”Vinho Verde” wine.
- Data Source : https://archive.ics.uci.edu/ml/datasets/wine+quality

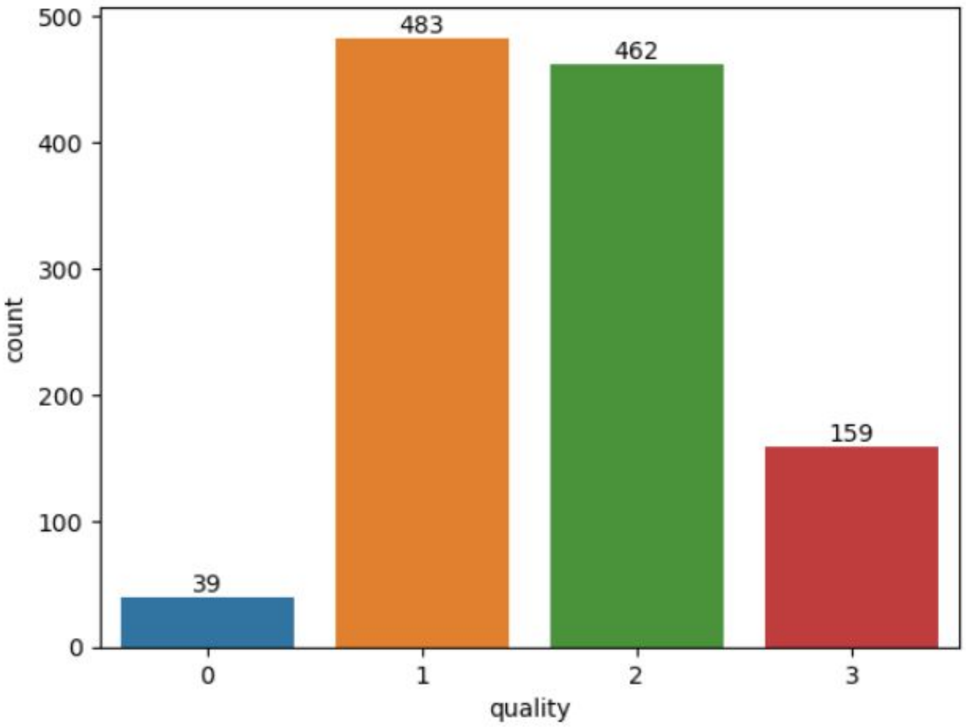


```
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         1143 non-null   float64
 1   volatile acidity      1143 non-null   float64
 2   citric acid           1143 non-null   float64
 3   residual sugar        1143 non-null   float64
 4   chlorides             1143 non-null   float64
 5   free sulfur dioxide   1143 non-null   float64
 6   total sulfur dioxide  1143 non-null   float64
 7   density               1143 non-null   float64
 8   pH                    1143 non-null   float64
 9   sulphates             1143 non-null   float64
 10  alcohol               1143 non-null   float64
 11  quality               1143 non-null   int64  
 12  Id                    1143 non-null   int64  
dtypes: float64(11), int64(2)
memory usage: 116.2 KB
```

The count plot of the whole dataset on the basis of the quality of wine is shown aside.

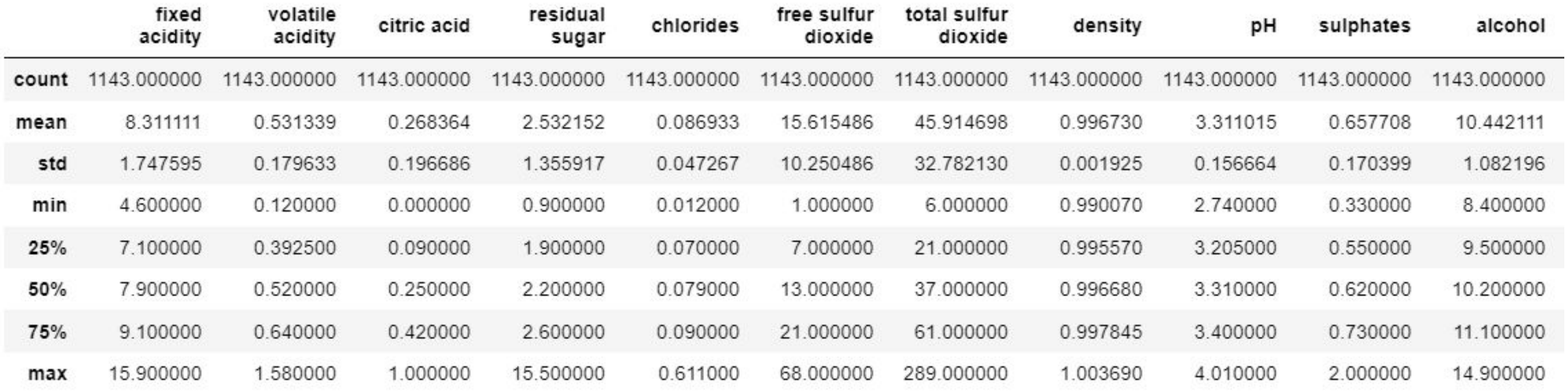


- label 3 and label 4 => label 0
- label 5 => label 1
- label 6 => label 2
- label 7 and label 8 => label 3



EDA

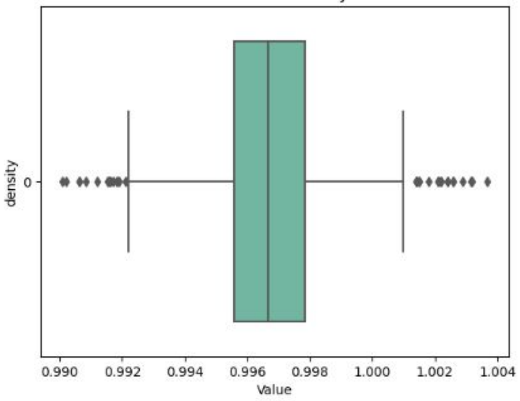
- Descriptive Statistics



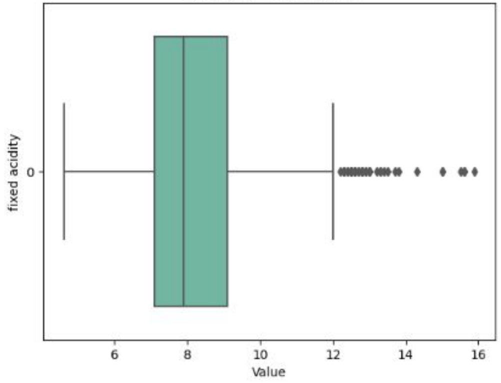
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	1143.000000	1143.000000	1143.000000	1143.000000	1143.000000	1143.000000	1143.000000	1143.000000	1143.000000	1143.000000	1143.000000
mean	8.311111	0.531339	0.268364	2.532152	0.086933	15.615486	45.914698	0.996730	3.311015	0.657708	10.442111
std	1.747595	0.179633	0.196686	1.355917	0.047267	10.250486	32.782130	0.001925	0.156664	0.170399	1.082196
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000
25%	7.100000	0.392500	0.090000	1.900000	0.070000	7.000000	21.000000	0.995570	3.205000	0.550000	9.500000
50%	7.900000	0.520000	0.250000	2.200000	0.079000	13.000000	37.000000	0.996680	3.310000	0.620000	10.200000
75%	9.100000	0.640000	0.420000	2.600000	0.090000	21.000000	61.000000	0.997845	3.400000	0.730000	11.100000
max	15.900000	1.580000	1.000000	15.500000	0.611000	68.000000	289.000000	1.003690	4.010000	2.000000	14.900000

- Box Plot of different features

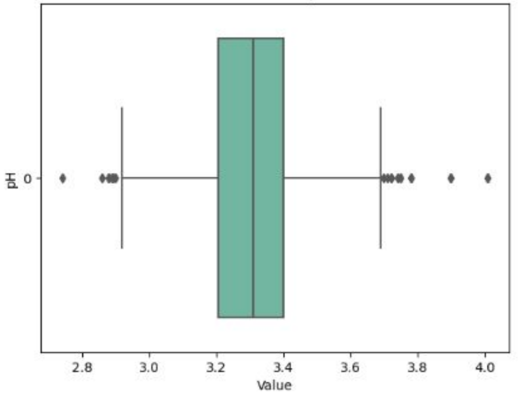
Box Plot of density



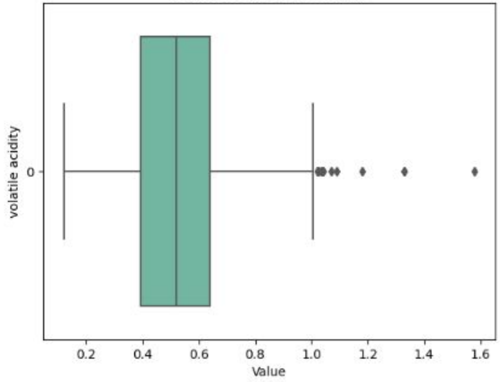
Box Plot of fixed acidity



Box Plot of pH

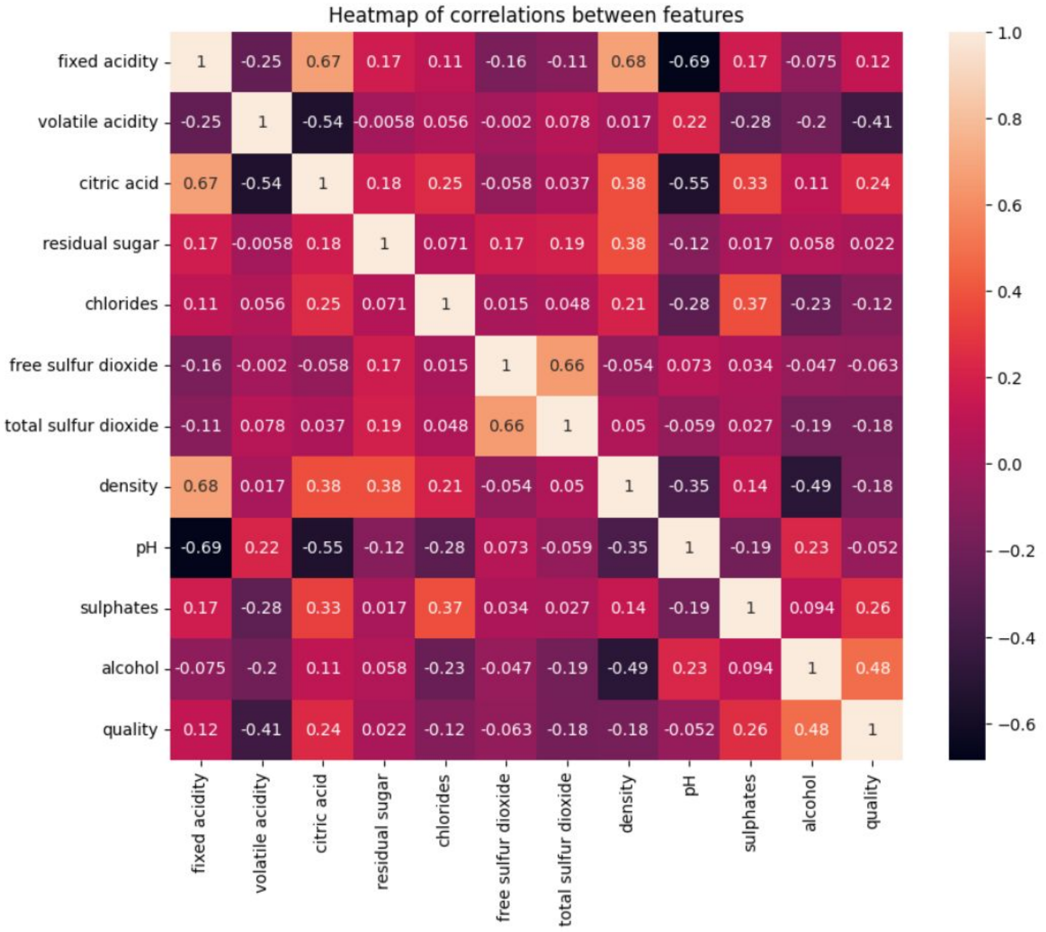


Box Plot of volatile acidity



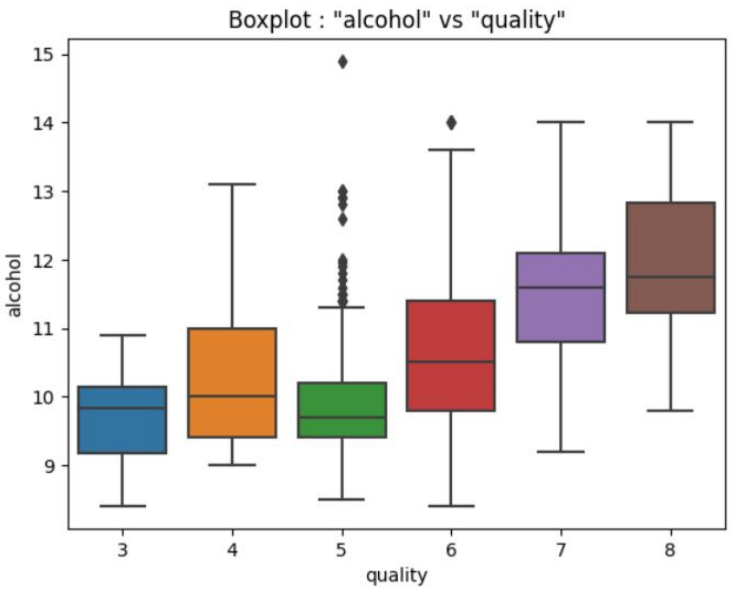
Box Plot of citric acid

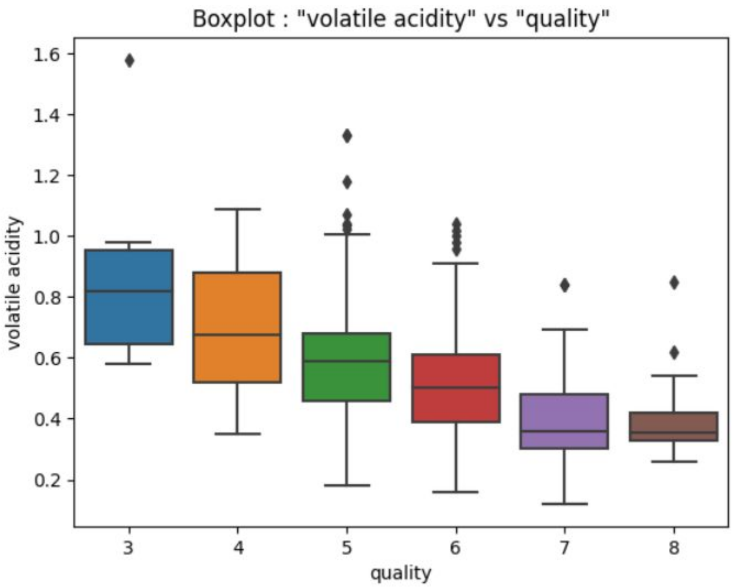
- Correlation matrix visualized as a heatmap



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	-0.25	0.67	0.17	0.11	-0.16	-0.11	0.68	-0.69	0.17	-0.075	0.12
volatile acidity	-0.25	1	-0.54	-0.0058	0.056	-0.002	0.078	0.017	0.22	-0.28	-0.2	-0.41
citric acid	0.67	-0.54	1	0.18	0.25	-0.058	0.037	0.38	-0.55	0.33	0.11	0.24
residual sugar	0.17	-0.0058	0.18	1	0.071	0.17	0.19	0.38	-0.12	0.017	0.058	0.022
chlorides	0.11	0.056	0.25	0.071	1	0.015	0.048	0.21	-0.28	0.37	-0.23	-0.12
free sulfur dioxide	-0.16	-0.002	-0.058	0.17	0.015	1	0.66	-0.054	0.073	0.034	-0.047	-0.063
total sulfur dioxide	-0.11	0.078	0.037	0.19	0.048	0.66	1	0.05	-0.059	0.027	-0.19	-0.18
density	0.68	0.017	0.38	0.38	0.21	-0.054	0.05	1	-0.35	0.14	-0.49	-0.18
pH	-0.69	0.22	-0.55	-0.12	-0.28	0.073	-0.059	-0.35	1	-0.19	0.23	-0.052
sulphates	0.17	-0.28	0.33	0.017	0.37	0.034	0.027	0.14	-0.19	1	0.094	0.26
alcohol	-0.075	-0.2	0.11	0.058	-0.23	-0.047	-0.19	-0.49	0.23	0.094	1	0.48
quality	0.12	-0.41	0.24	0.022	-0.12	-0.063	-0.18	-0.18	-0.052	0.26	0.48	1

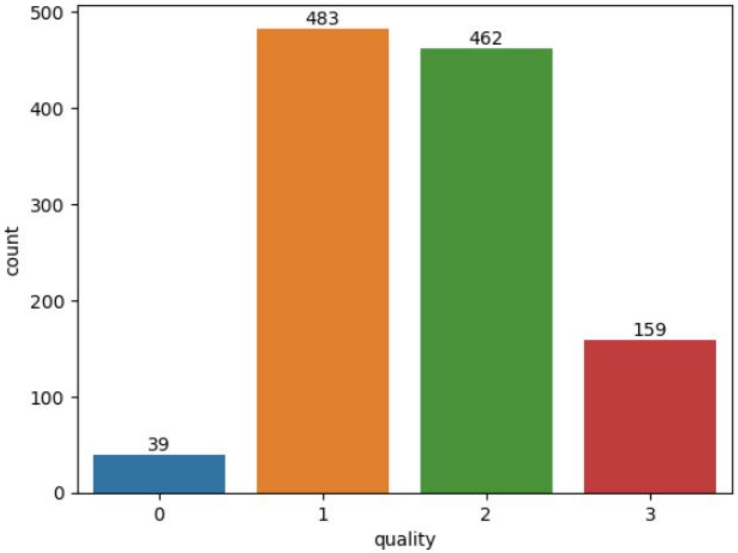
- Distribution of features having higher correlation with target, visualized with respect to ‘quality’



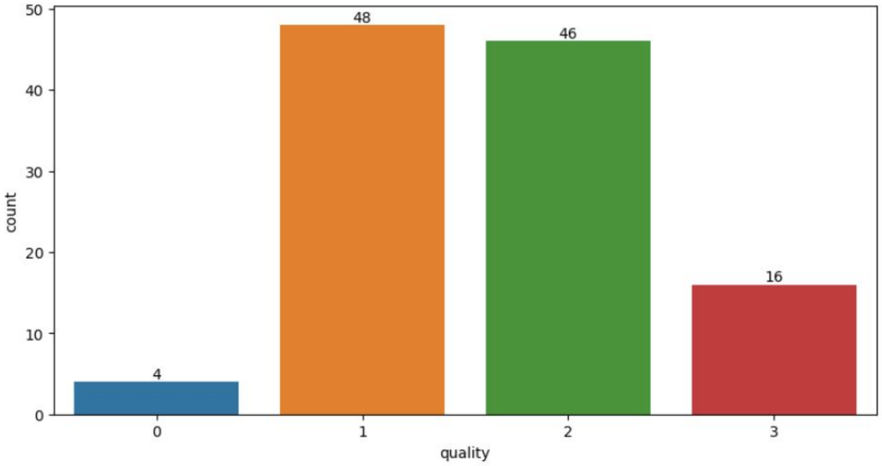


Train Test Split

80% Train Data



20% Train Data



Imbalance Data Handling

- Undersampling
- Oversampling
- Smote
- Class Weights

Model Development

- Logistic Regression
- Decision Trees (Experiment with early stopping, pruning)
- Ensembling methods
- Random Forest

Model Evaluation

- F1-score as the primary evaluation metric
- roc
- auc-score
- precision and recall

Experiment Tracking

- Tensorboard

Tools and Technologies

- Trello for agile methodology
- Discord for team meeting

CONCLUSION