RED WINE

Andreea Stroia
Hala Albahloul
Hritika Kathuria

PS: MADE DRINKING WHITE WINE

Tabel of contents

Wine is the most healthful and most hygienic of beverages

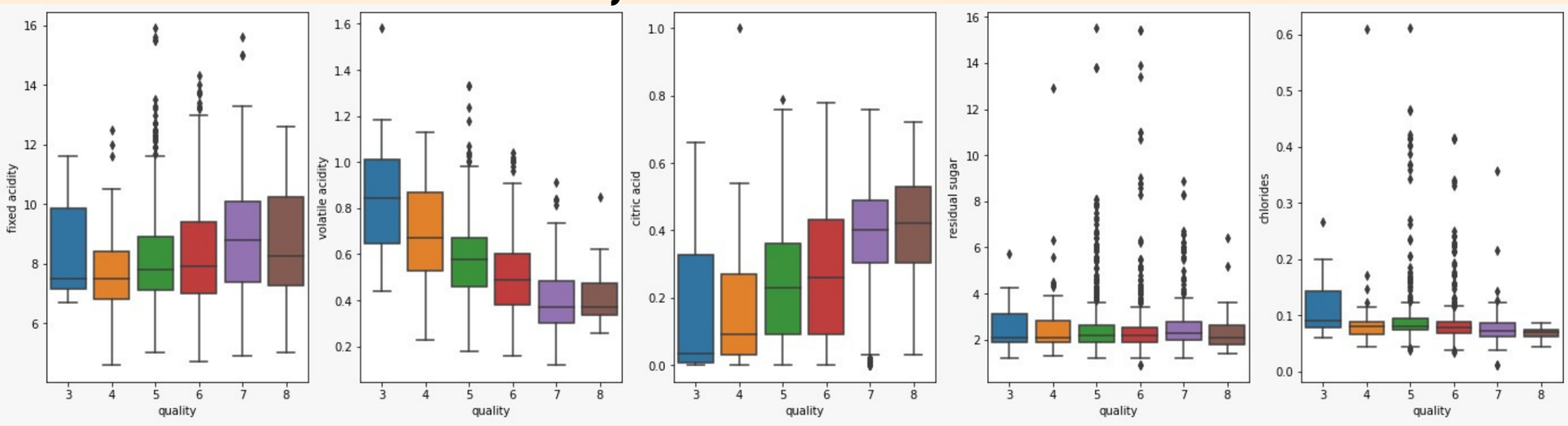
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

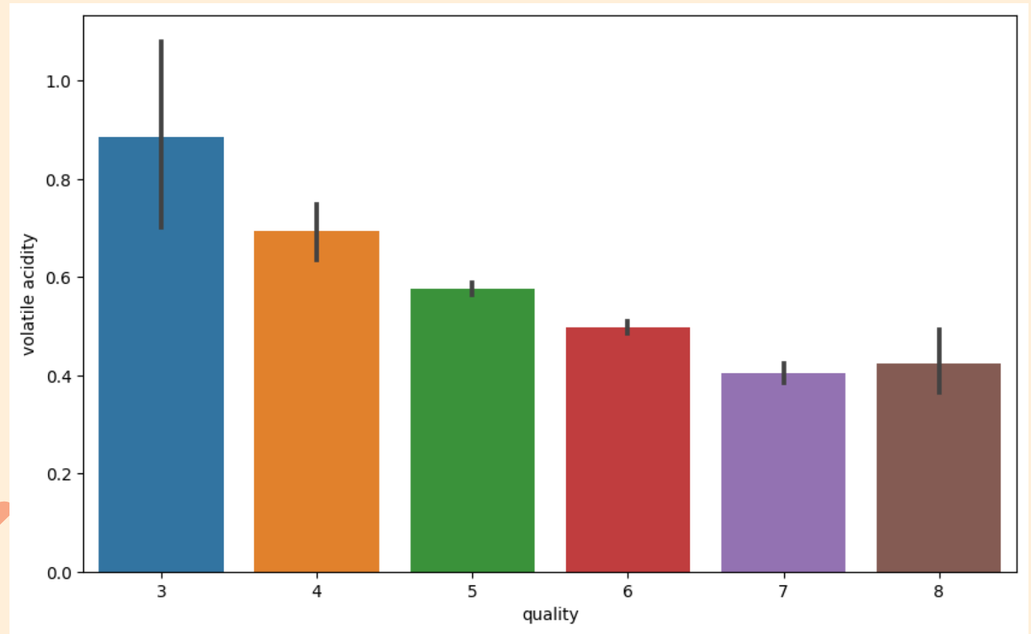
Some info on the dataset

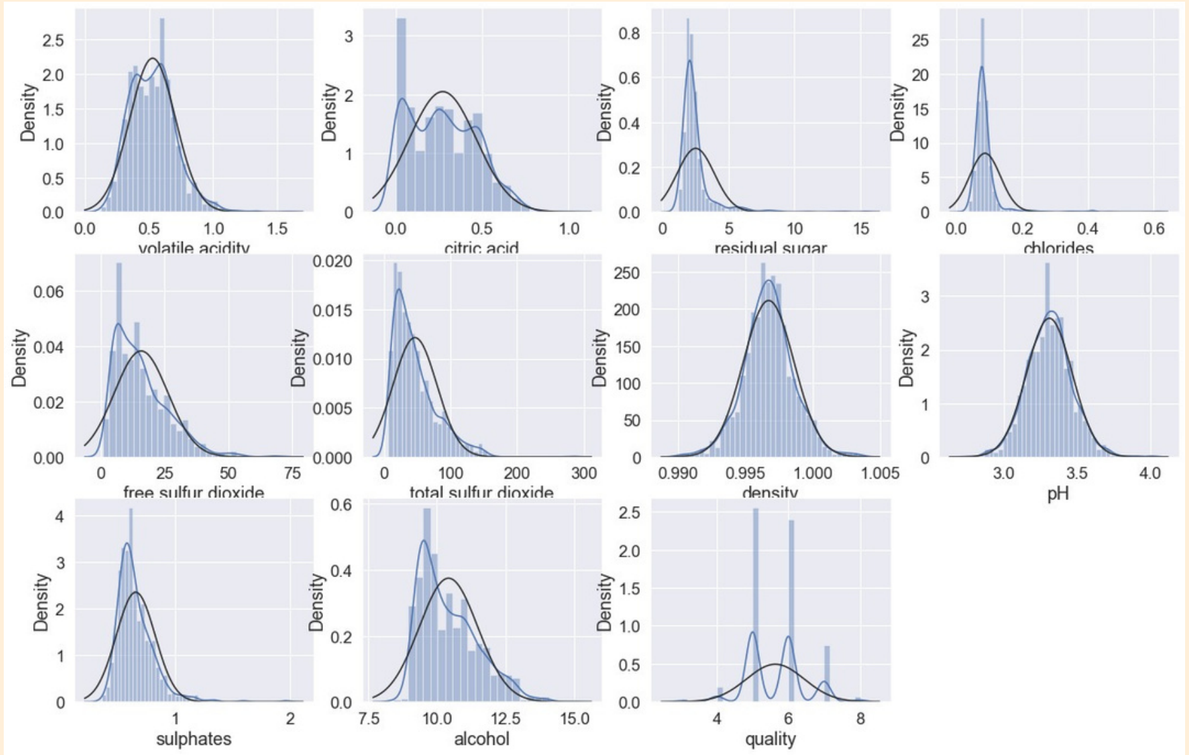
- No null values
- Only numerical data
- Duplicate values

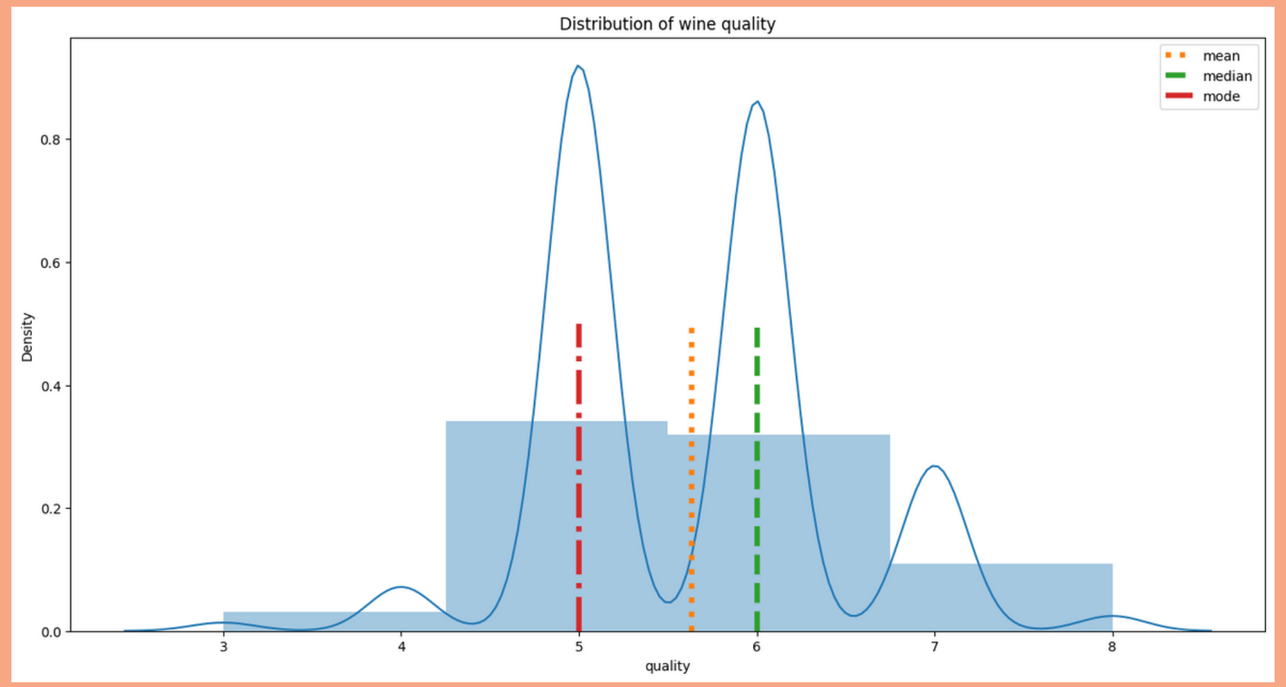
- 1559 Rows
- 12 Columns

QUALITY
(SEEN THROUGH OTHER FEATURES)

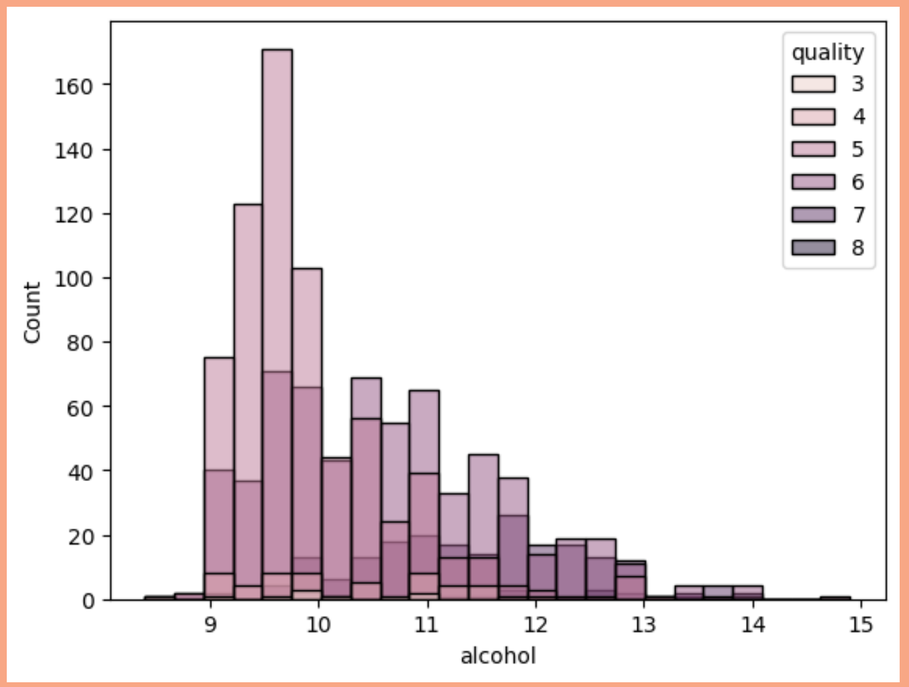




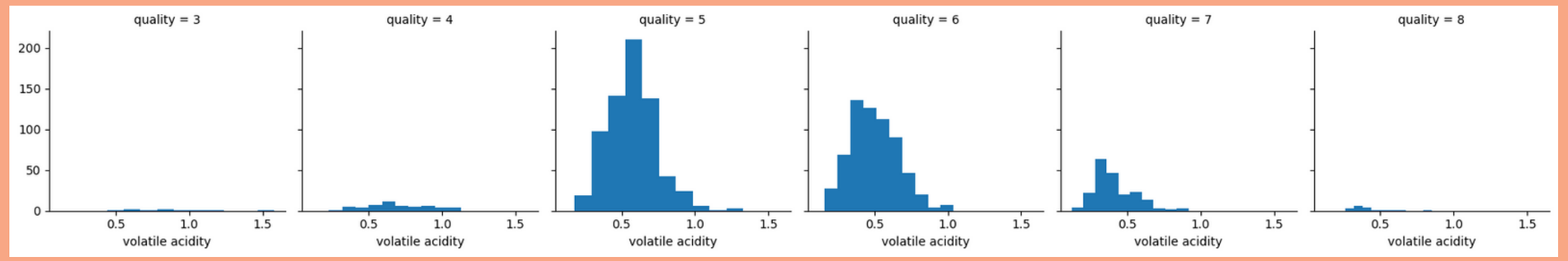




DENSITY

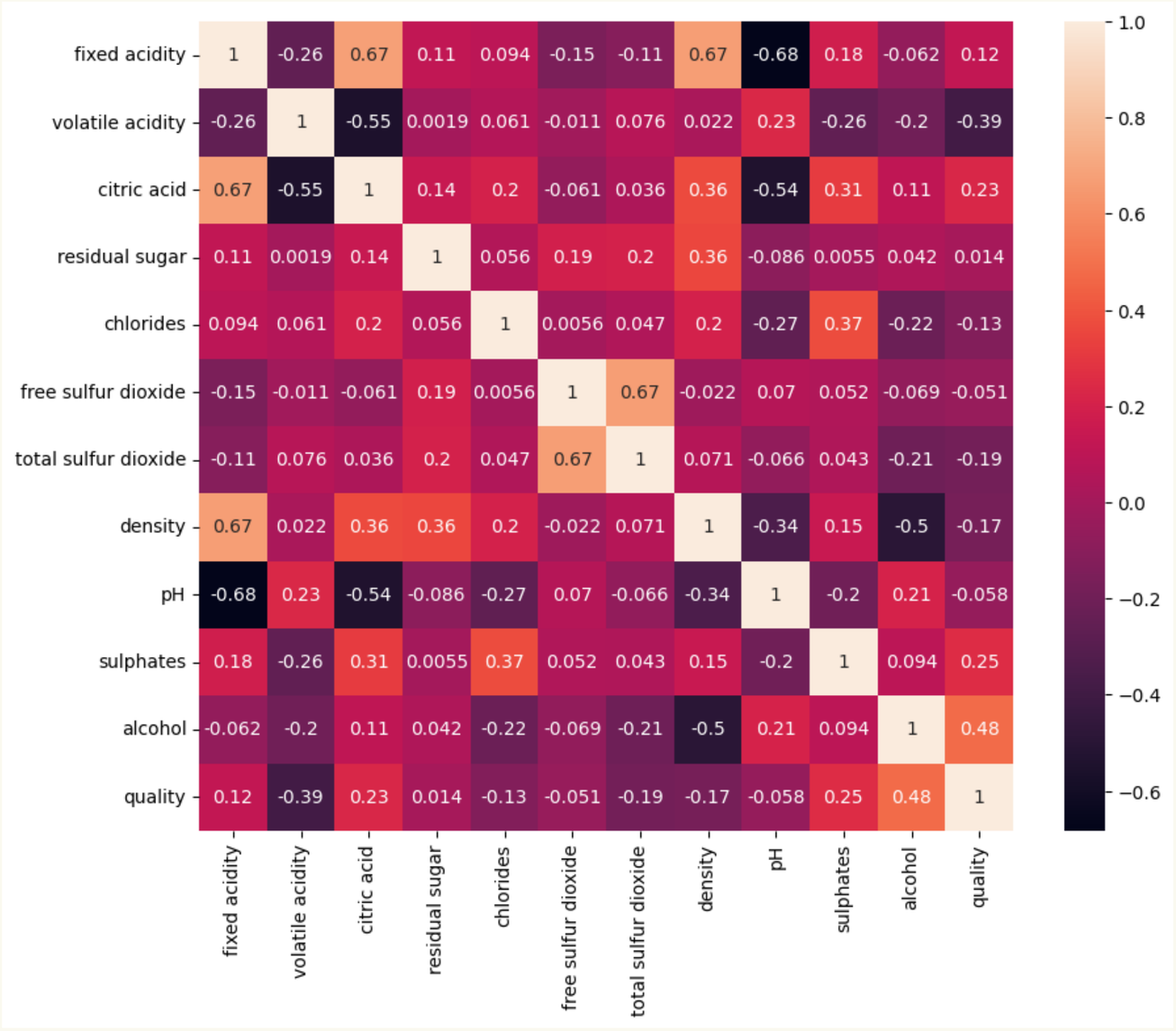


ALCOHOL



VOLATILITY ACID

CORRELATION



PREDICTING THE QUALITY OF WINE

- 0 if quality <=6
- 1 if quality >6

- Volatile Acidity
- Alcohol
- Sulphates
- Citric acid

STEPS

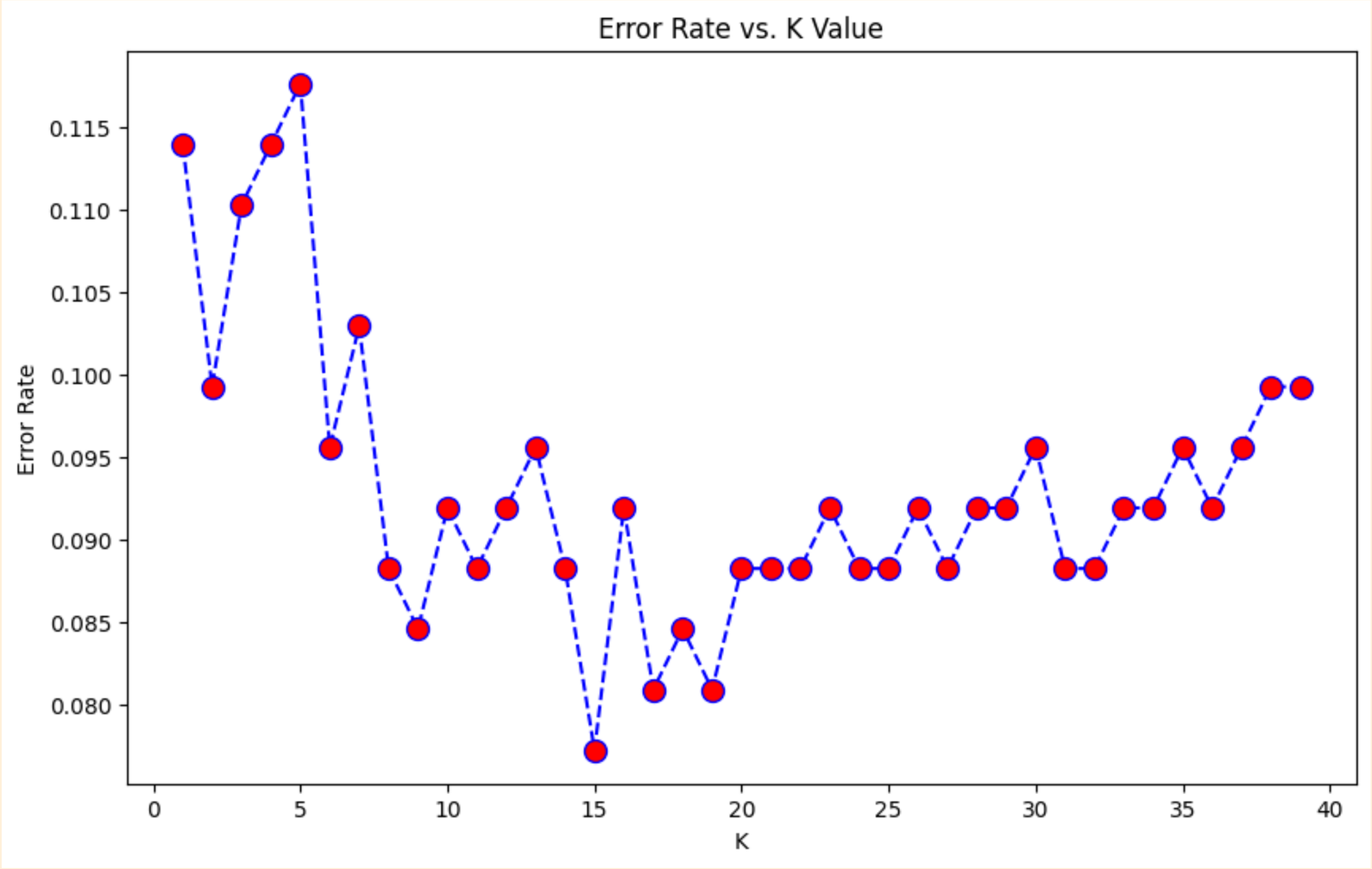
CLASSIFIER MODELS

- KNN
- Logistic Regression
- Decision Trees
- Naive Bayes

OPTIMIZATION

PERFORMANCE EVALUATION

KNN



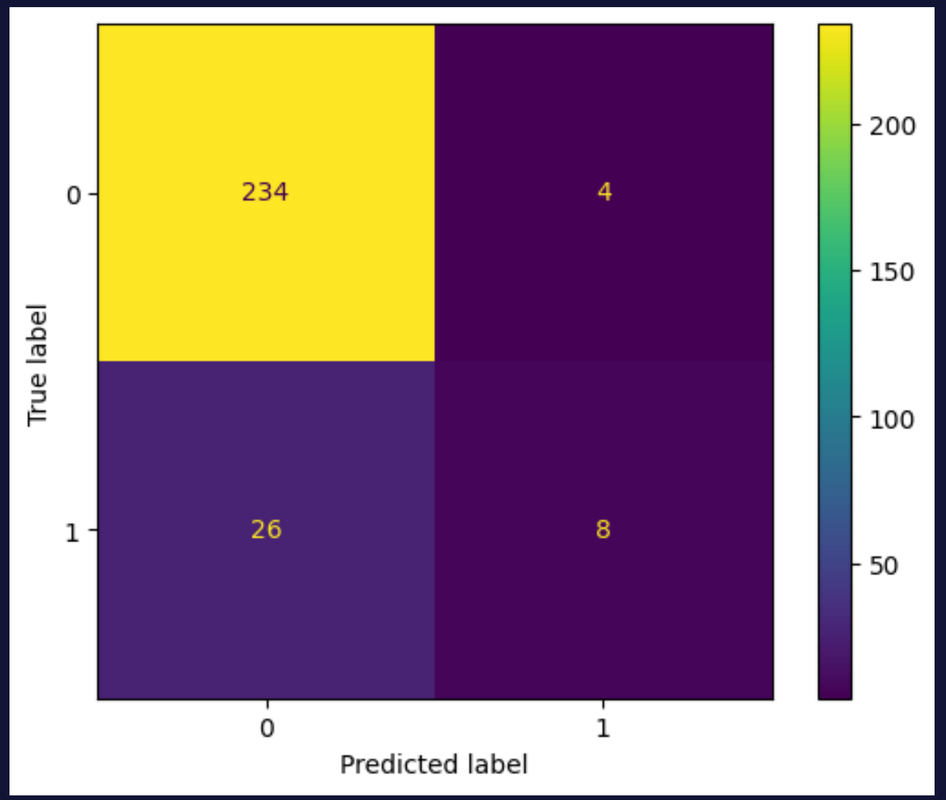
K=15 has lowest error rate

Accuracy score:

- train data set: 0.88
- test data set: 0.89

```
[0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0,
 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 1, 0, 0, 0, 0])
```

LOGISTIC REGRESSION



Cross validation

Accuracy:

- 5 K-fold: 0.86
- 10 K-fold: 0.87

	precision	recall	f1-score	support
0	0.90	0.98	0.94	238
1	0.67	0.24	0.35	34
accuracy			0.89	272
macro avg	0.78	0.61	0.64	272
weighted avg	0.87	0.89	0.87	272

DECISION TREES

Accuracy

Max depth=10
decision tree

0.82

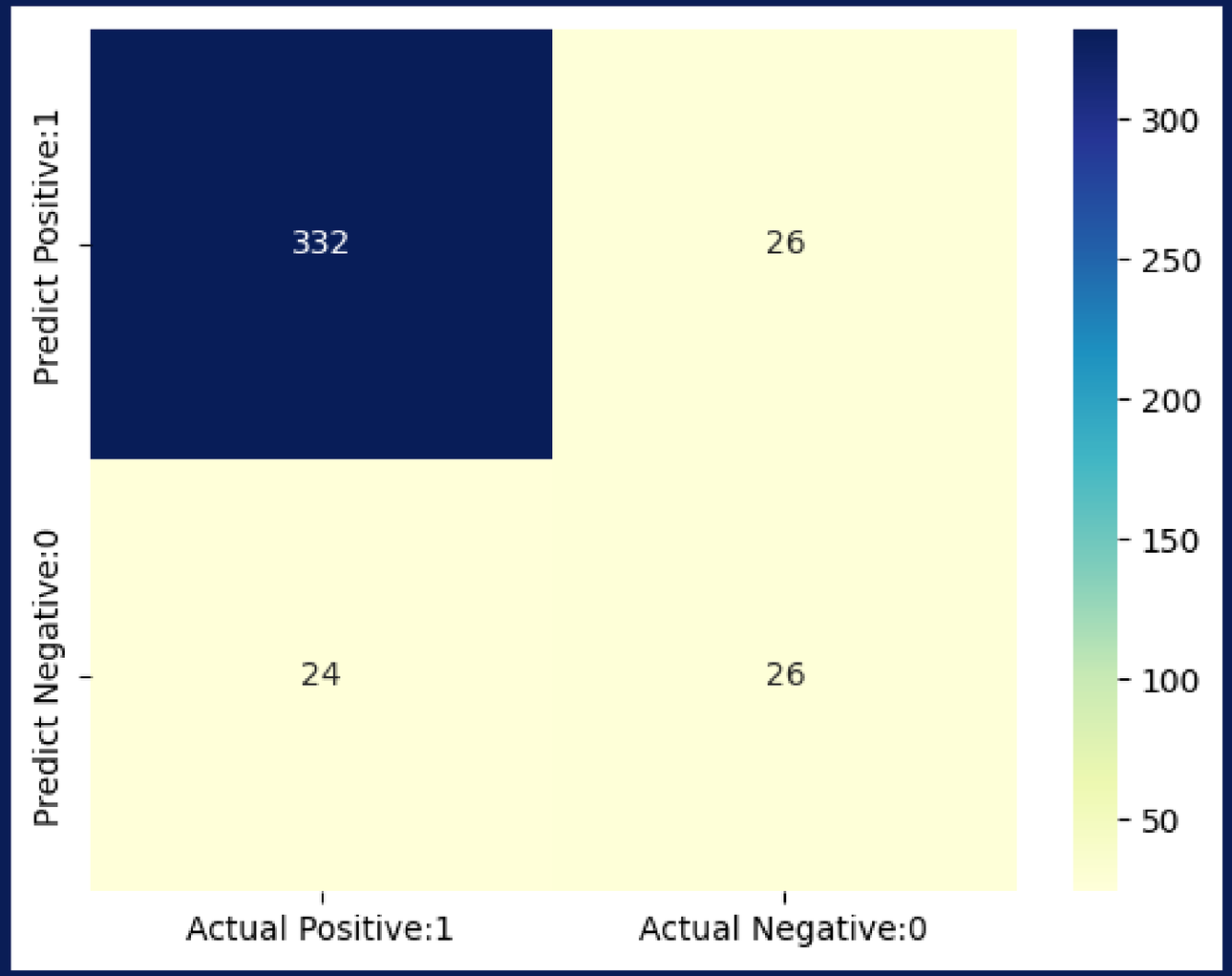
	Predicted BAD	Predicted GOOD
Actual BAD	313	45
Actual GOOD	27	23

After choosing max depth(4)
&
cross validation(5 Kfolds)

0.89

	Predicted BAD	Predicted GOOD
Actual BAD	339	19
Actual GOOD	37	13

Bootstrap Aggregation
Train: 0.86
Test: 0.89



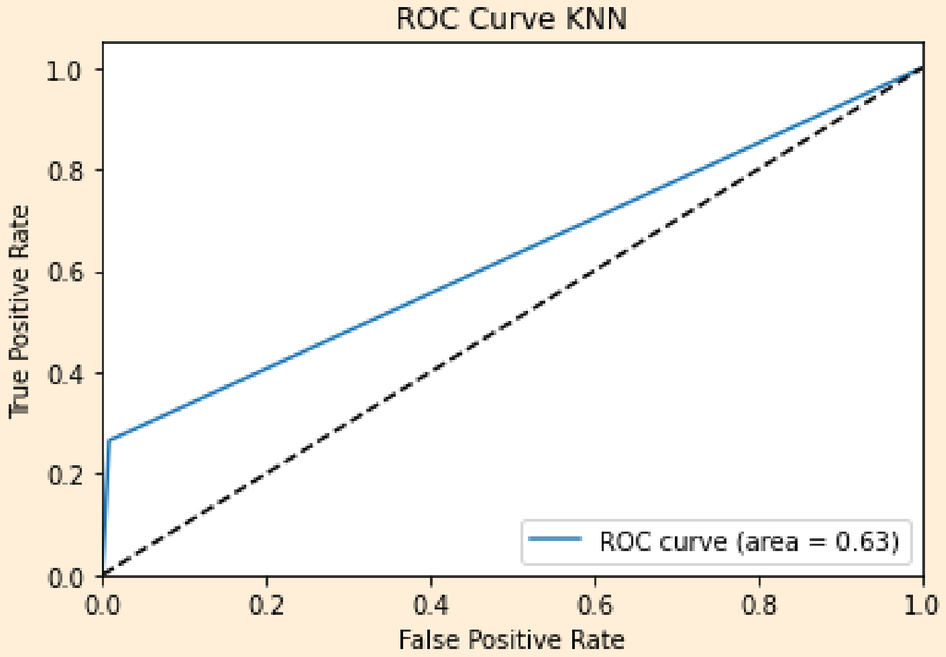
Naive Bayes

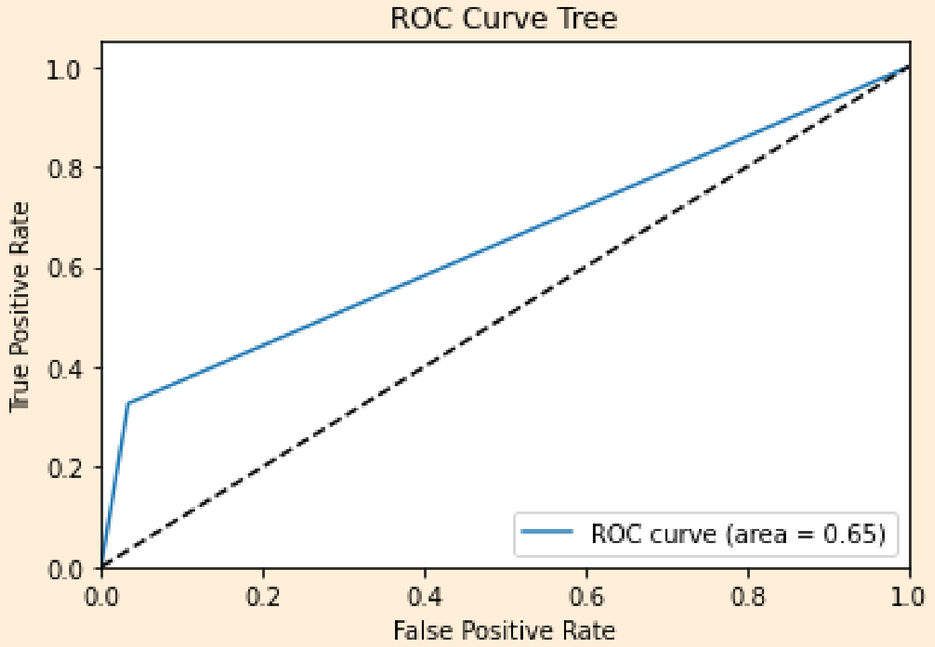
- Training set score: 0.8633
- Test set score: 0.8775

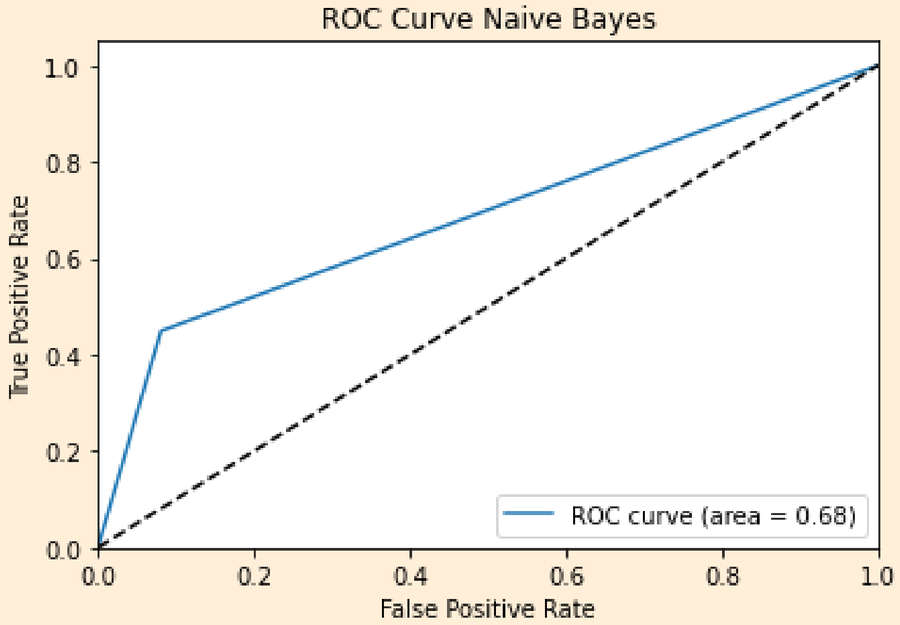
	precision	recall	f1-score	support
0	0.93	0.93	0.93	358
1	0.50	0.52	0.51	50
accuracy			0.88	408
macro avg	0.72	0.72	0.72	408
weighted avg	0.88	0.88	0.88	408

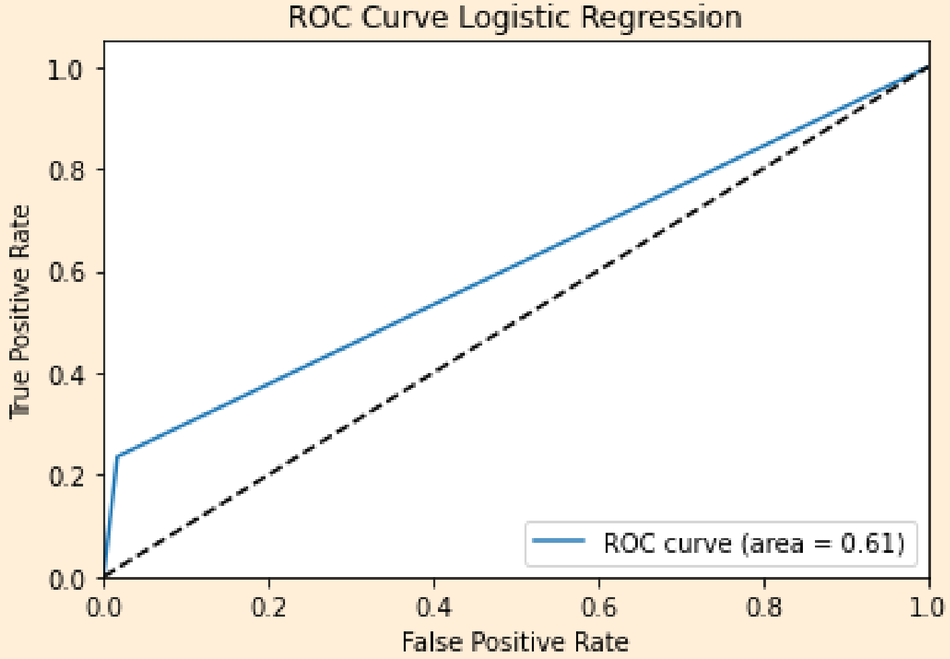
PERFORMANCE EVALUATION

Model	BER
Logistic Regression	0.391
KNN	0.371
Decision Trees	0.396
Naive Bayes	0.276


ROC Curve KNN
True Positive Rate
False Positive Rate
ROC curve (area = 0.63)
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve Tree
True Positive Rate
False Positive Rate
ROC curve (area = 0.65)
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve Naive Bayes
True Positive Rate
False Positive Rate
ROC curve (area = 0.68)
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve Logistic Regression
True Positive Rate
False Positive Rate
ROC curve (area = 0.61)
0.0
0.2
0.4
0.6
0.8
1.0

Thank you for your attention!