Assignment 2
ACIT 4880
Machine Learning

Tamim Hemat and
Daryush Balsara

Agenda

- Dataset
- EDA
- K Means Clustering and Initial Clusters
- Random Forest Clustering
- Logistic Regression

Dataset - Red Wine Quality

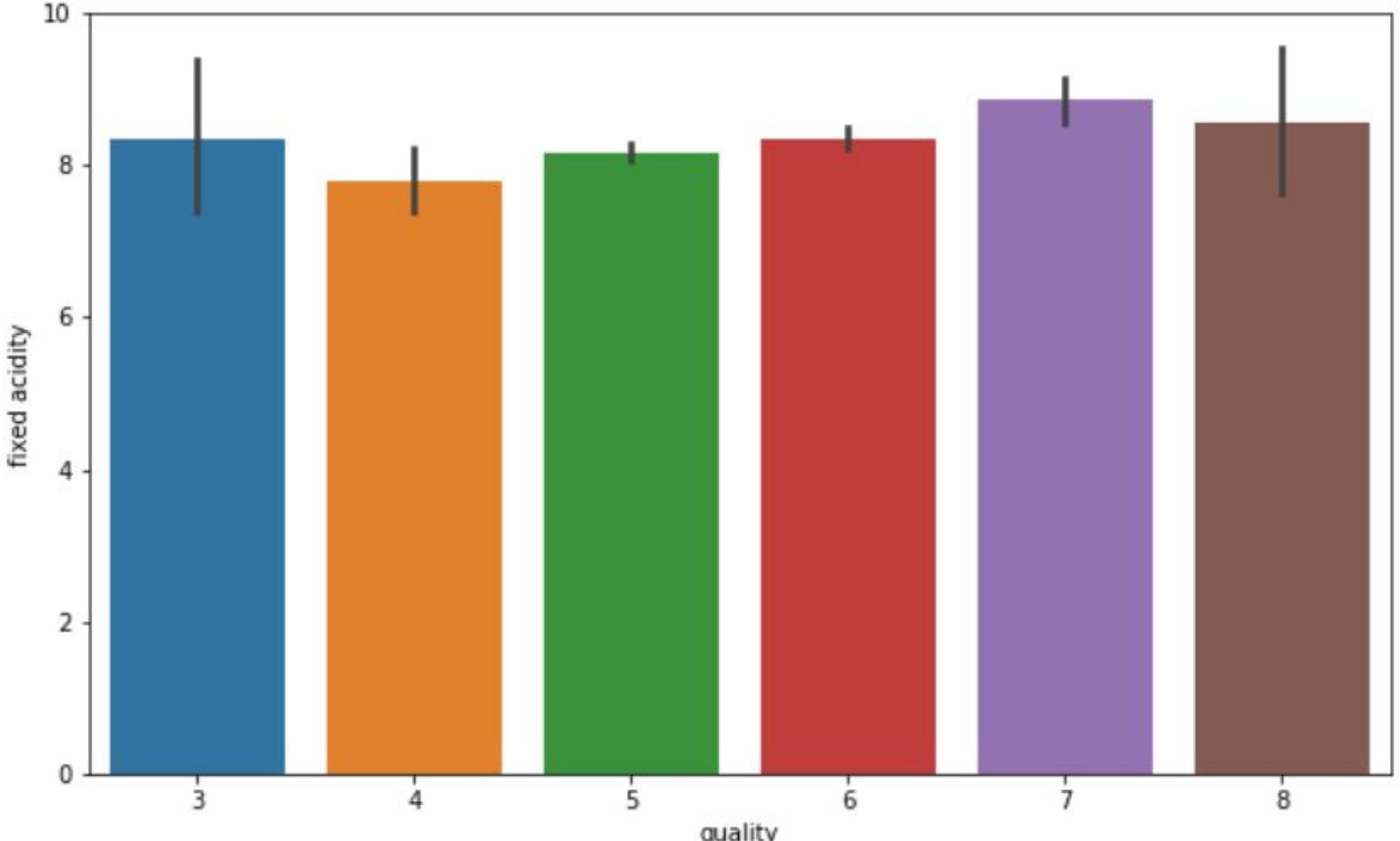
- Our dataset consists of 12 columns with 1600 records
- There was no missing data
- One of the columns was unusable so we worked with 11 of the columns
- We had 10 independent variables (features) and 1 dependent (discrete target value)

Features:

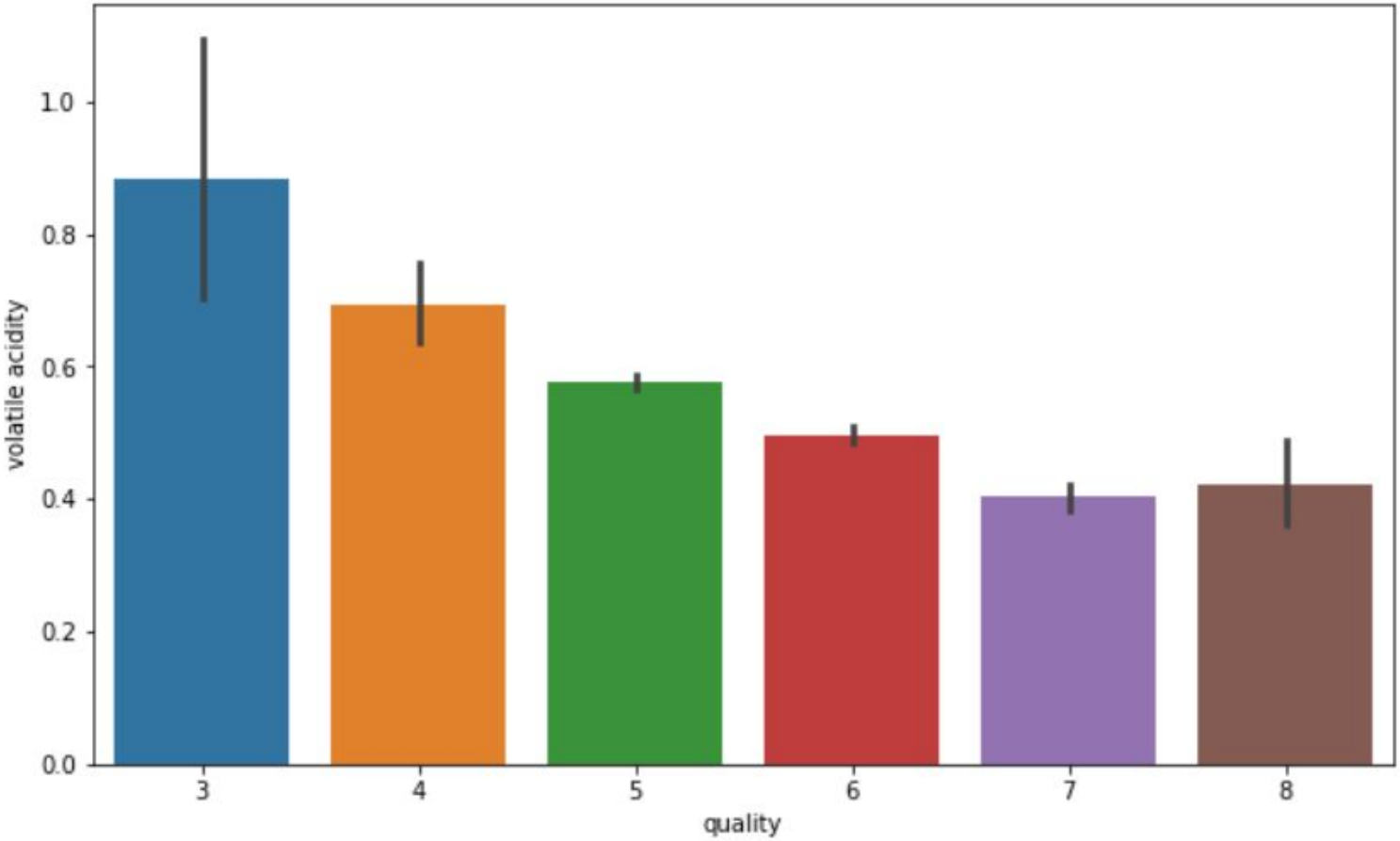
- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- density
- pH
- sulphates
- alcohol

Target Value: Wine quality

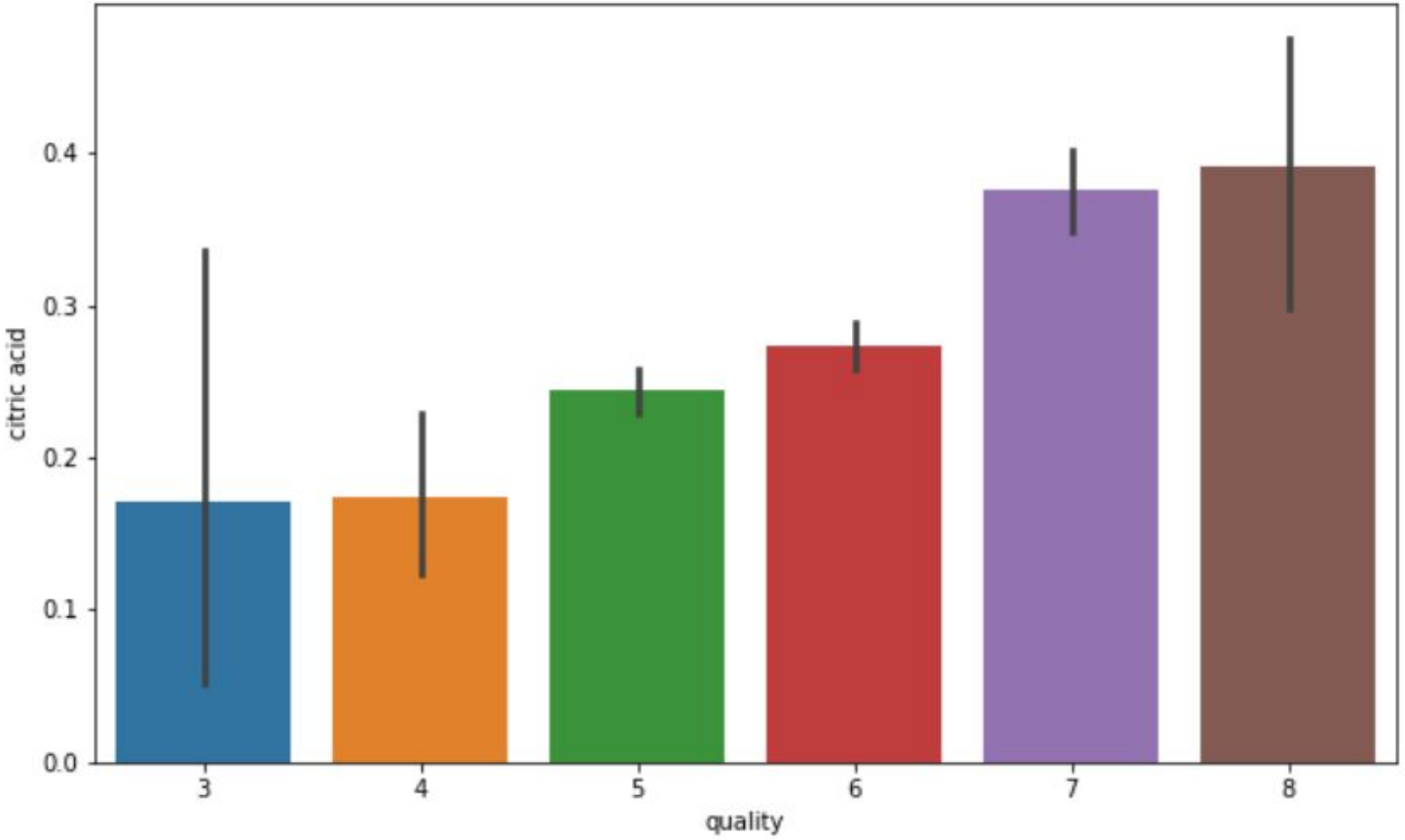
EDA



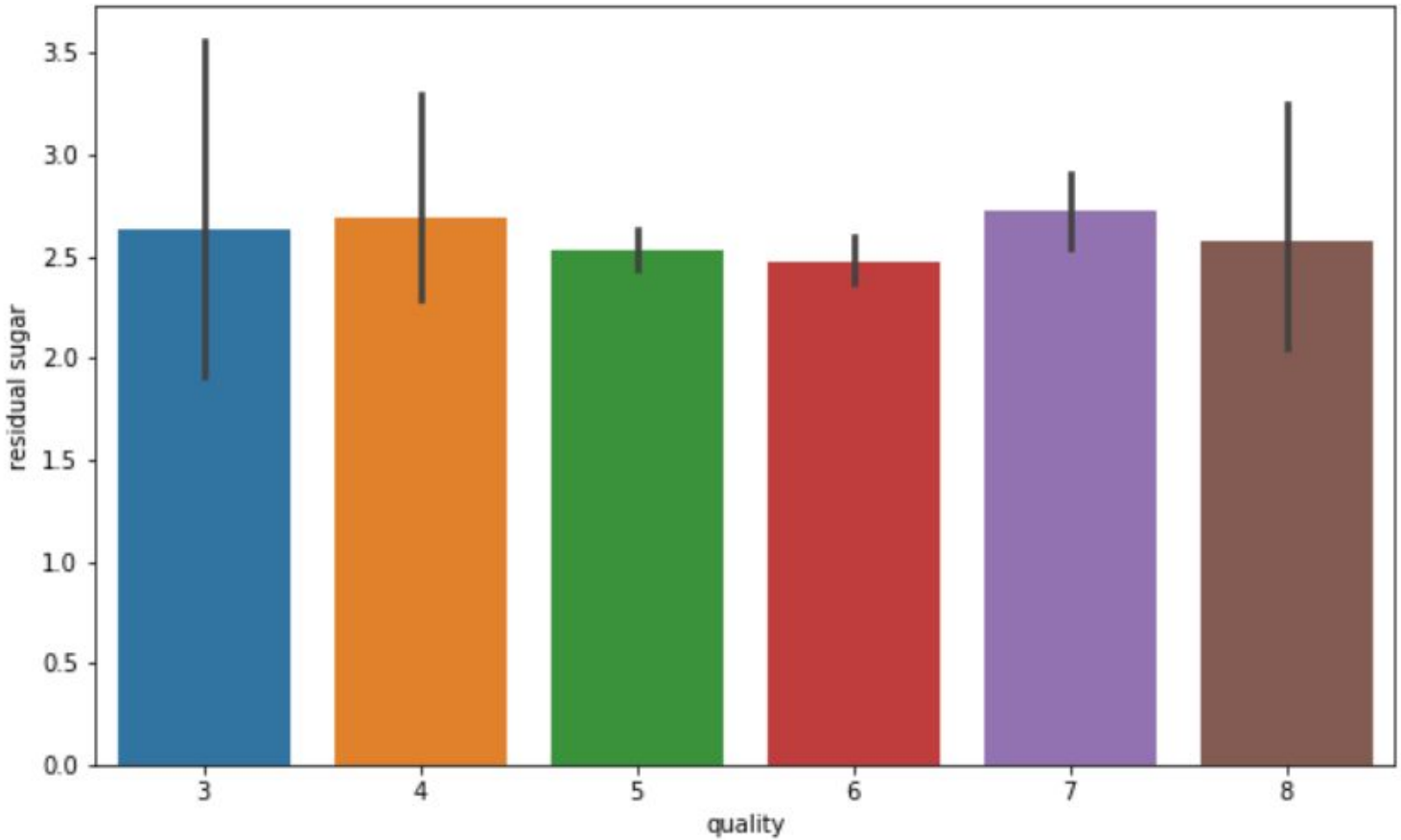
EDA



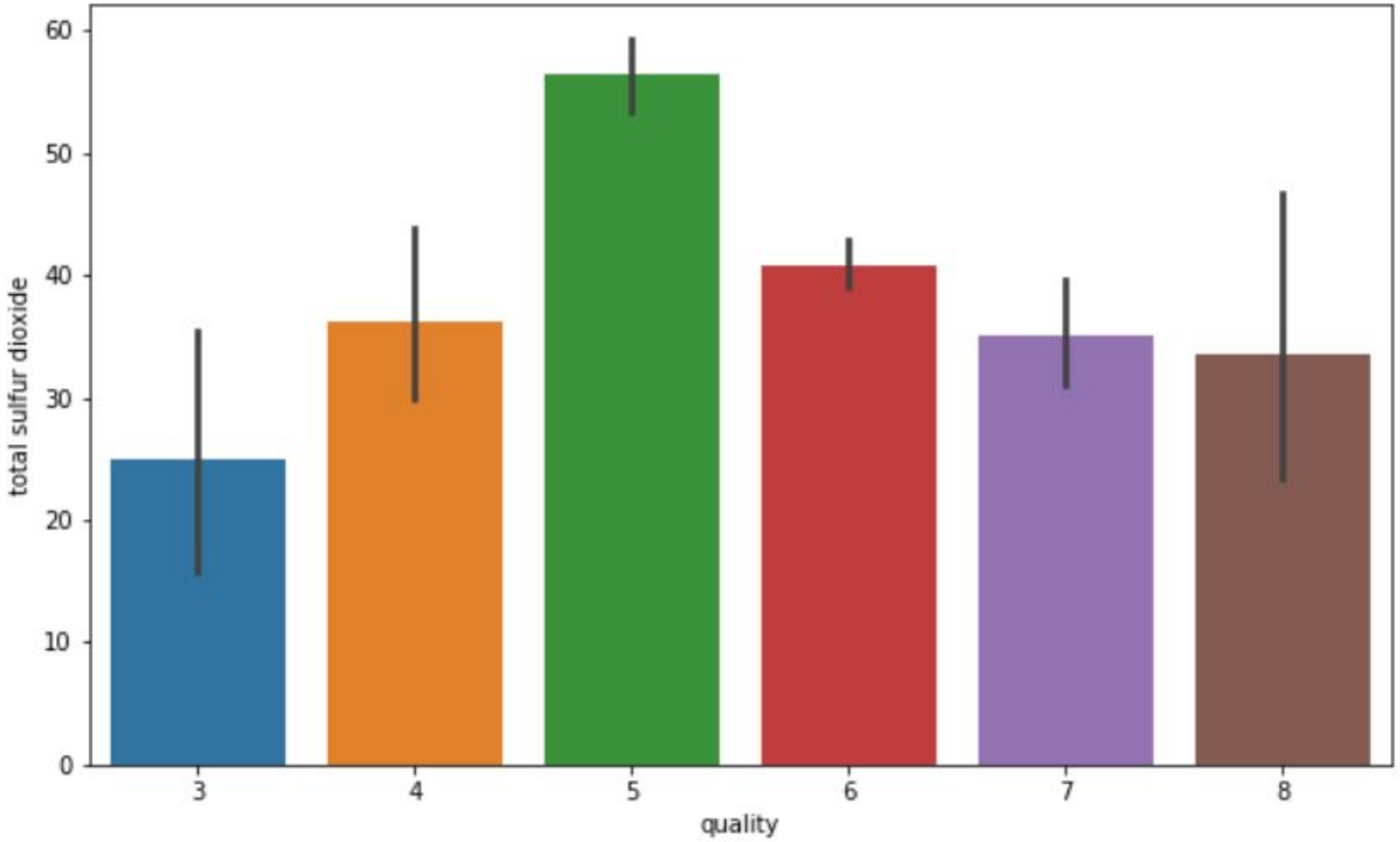
EDA



EDA



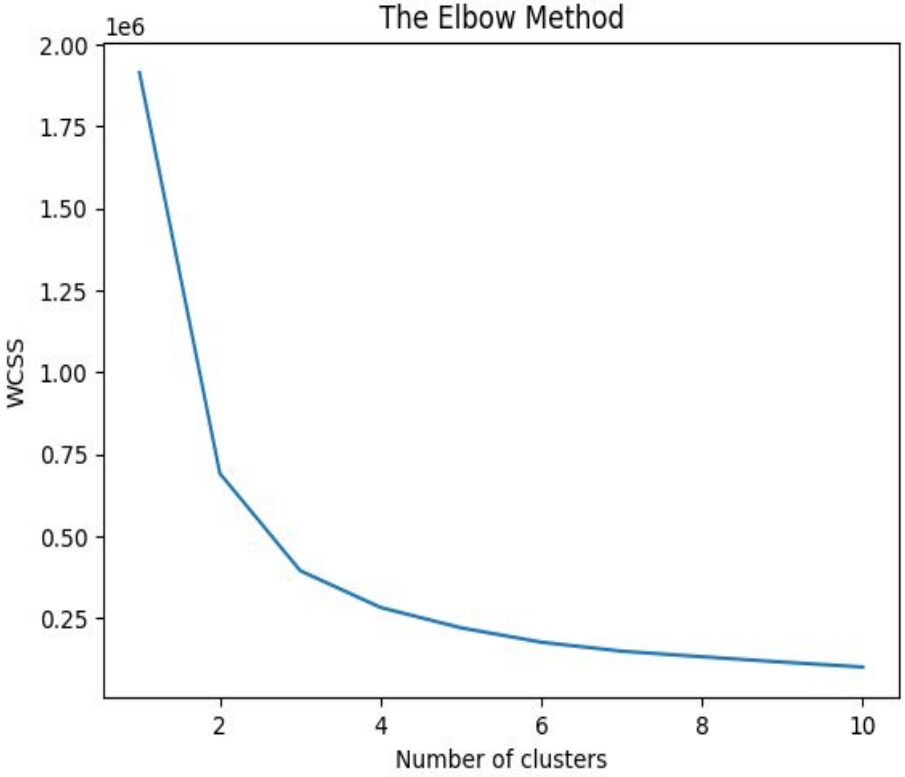
EDA

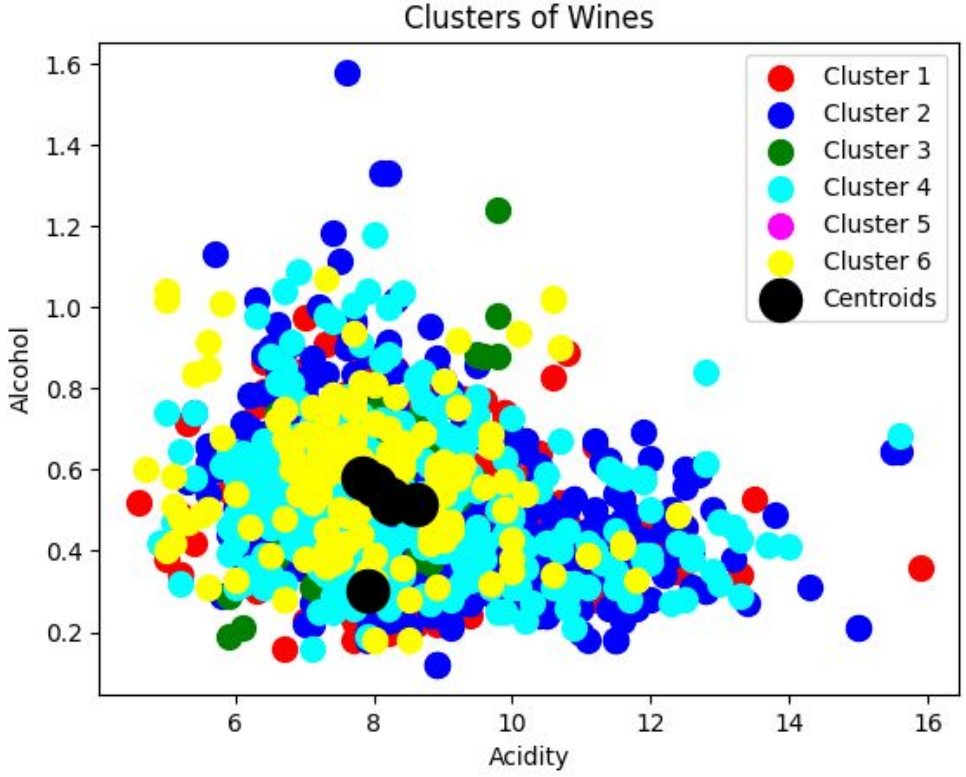


K-Means Clustering

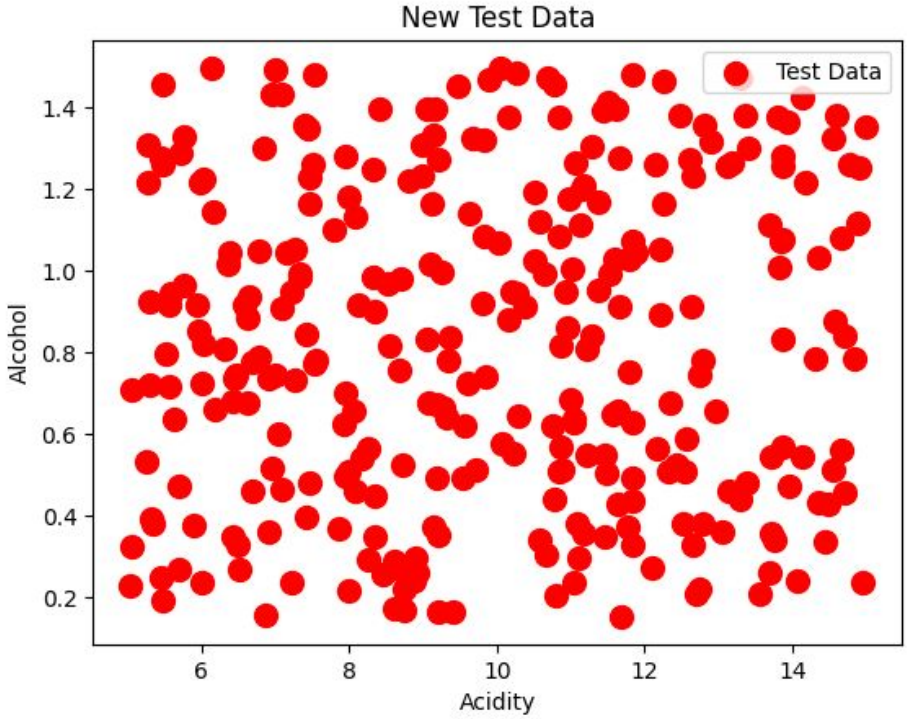
- We used the elbow method to find the optimal numbers of clusters
- Found the optimal number being 6
- Used the acidity and alcohol features of the data to categorize types of wine
- Initially clusters were tightly coupled
- The centroids were very close to each other

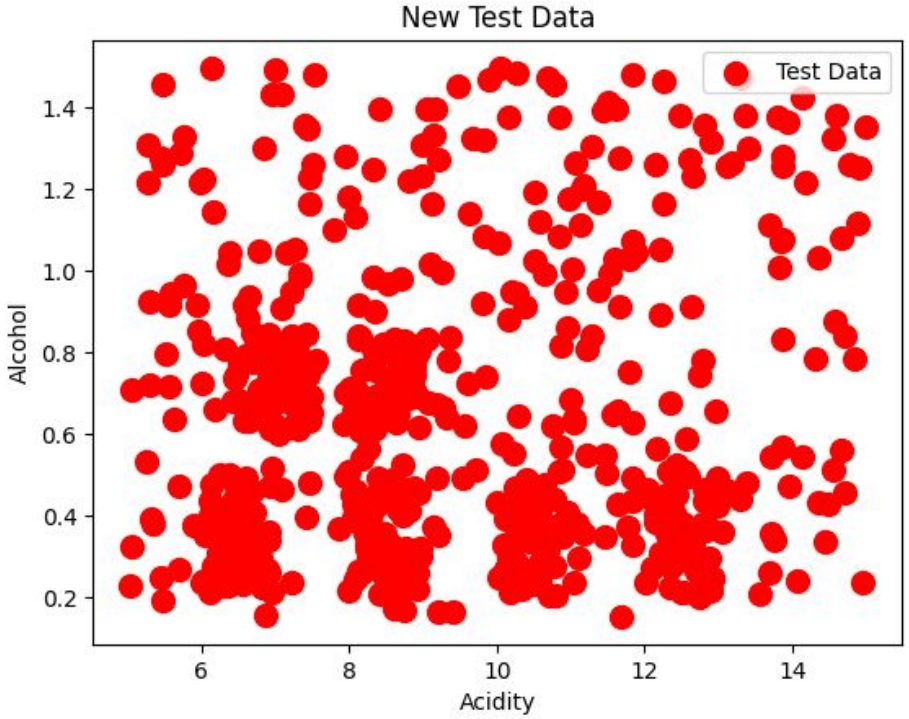
Initial clusters



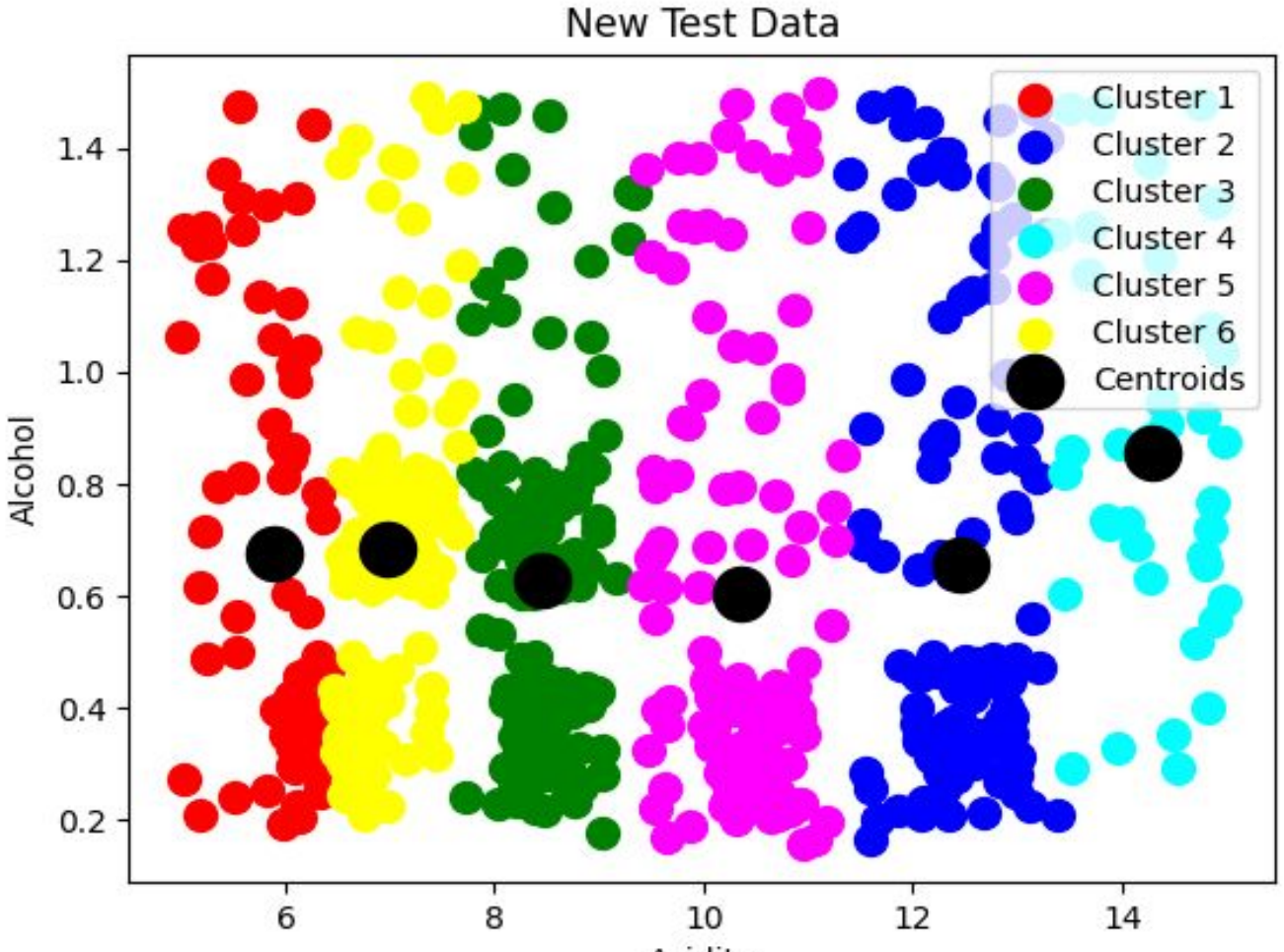


Adding extra data for training the model





Final Clusters – they are distinct



Random Forest Classification

- Initially we used fixed acidity,volatile acidity,citric acid,total sulfur dioxide,pH,alcohol for features
- Set the quality of the wine as the target value then we split the data for testing
- We applied standard scaling for the features
- We trained the model using entropy as the criterion and 100 estimators (trees)
- We got the confusion matrix, accuracy score, and classification report

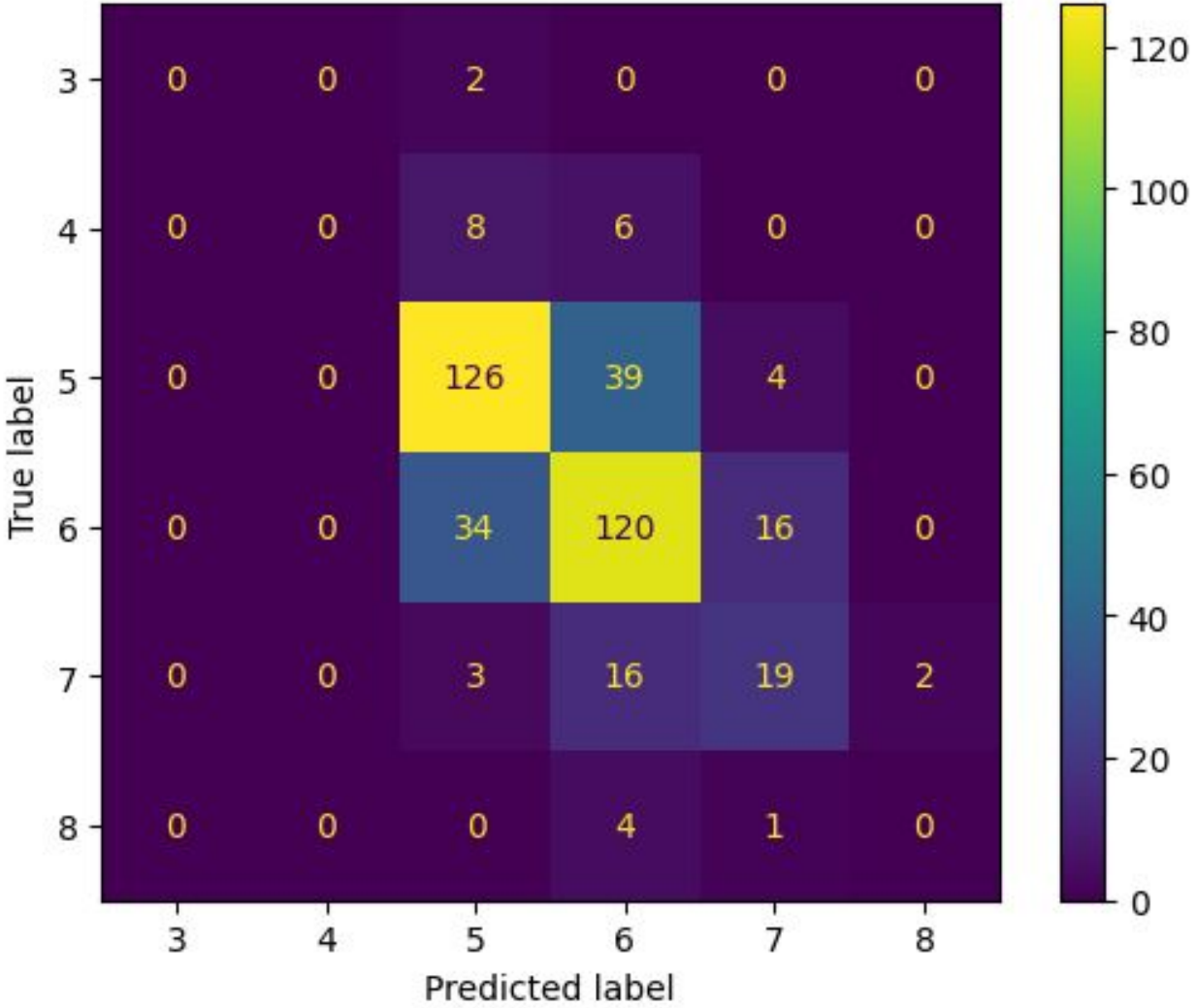
```
[[  0   0   2   0   0   0]
 [  0   0   8   6   0   0]
 [  0   0 126  39   4   0]
 [  0   0  34 120  16   0]
 [  0   0   3  16  19   2]
 [  0   0   0   4   1   0]]

Accuracy score: 0.6625

              precision    recall  f1-score   support

           3       0.00      0.00      0.00         2
           4       0.00      0.00      0.00        14
           5       0.73      0.75      0.74       169
           6       0.65      0.71      0.68       170
           7       0.47      0.47      0.48        40
           8       0.00      0.00      0.00         5

    accuracy                           0.66       400
   macro avg       0.31      0.32      0.31       400
weighted avg       0.63      0.66      0.65       400
```

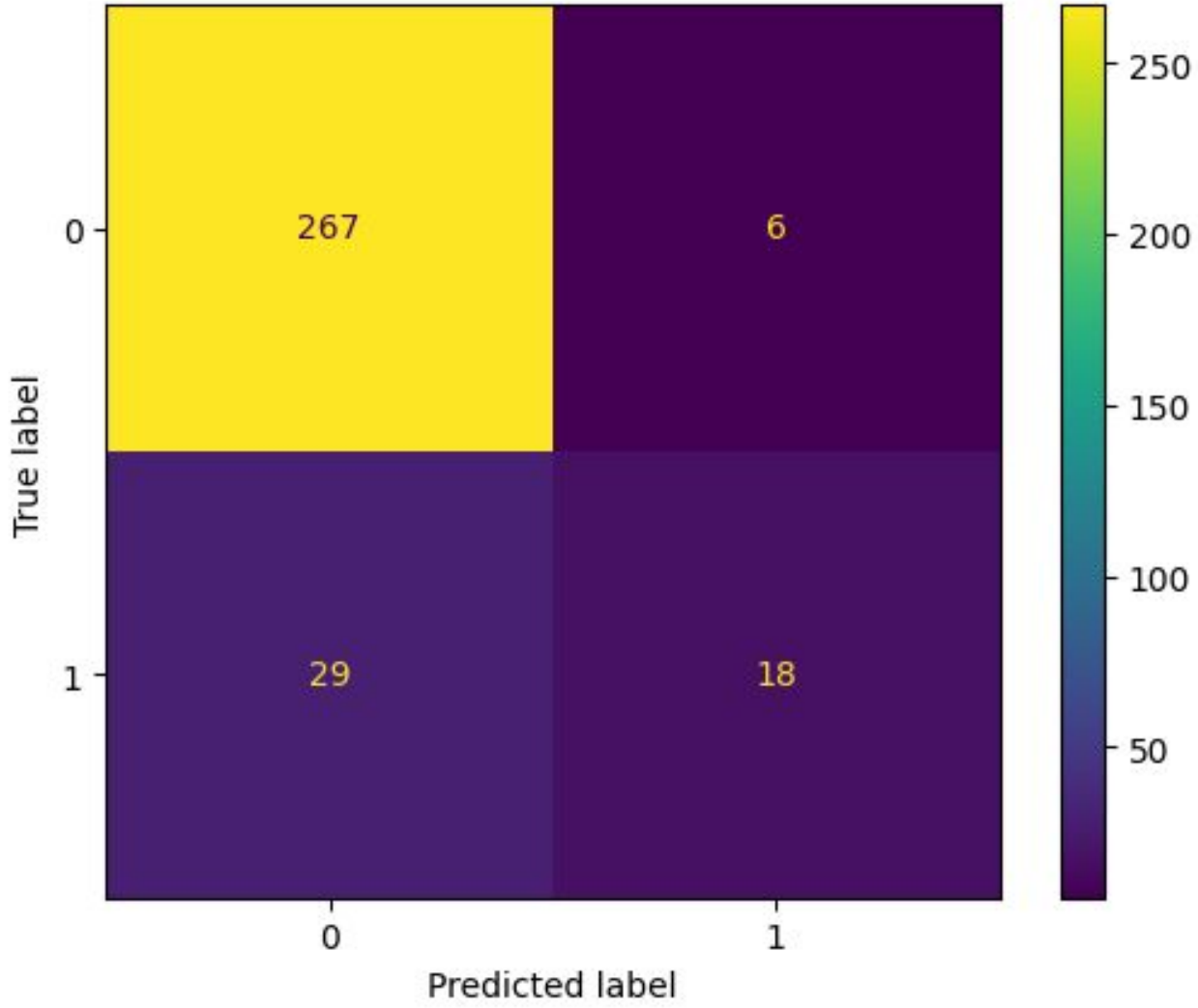

3
4
5
6
7
8
True label
0 0 2 0 0 0
0 0 8 6 0 0
0 0 126 39 4 0
0 0 34 120 16 0
0 0 3 16 19 2
0 0 0 4 1 0
3
4
5
6
7
8
Predicted label
120
100
80
60
40
20
0

Random Forest Classification

- Used a label and encoder to divide wine quality into good and bad
- Then we used all features and we set the quality as the target value and then we split the data for testing again
- Applied standard scaling
- Trained the model using entropy as the criterion and 200 estimators (trees)
- Got the confusion matrix, classification report, and accuracy score

	precision	recall	f1-score	support
0	0.91	0.97	0.94	273
1	0.74	0.43	0.54	47
accuracy			0.89	320
macro avg	0.82	0.70	0.74	320
weighted avg	0.88	0.89	0.88	320

```
[[266   7]
 [ 27  20]]
Accuracy score: 0.89375
```


267
6
29
18
0
1
True label
0
1
Predicted label
250
200
150
100
50

Logistic Regression

- We started by used all features
- We set the quality of the wine as the target value then we split the data for testing
- We applied standard scaling for the features
- We trained the model using the LBFGS algorithm, multinomial class and 1000 iterations
- We got the confusion matrix, accuracy score, and classification report and F1 score

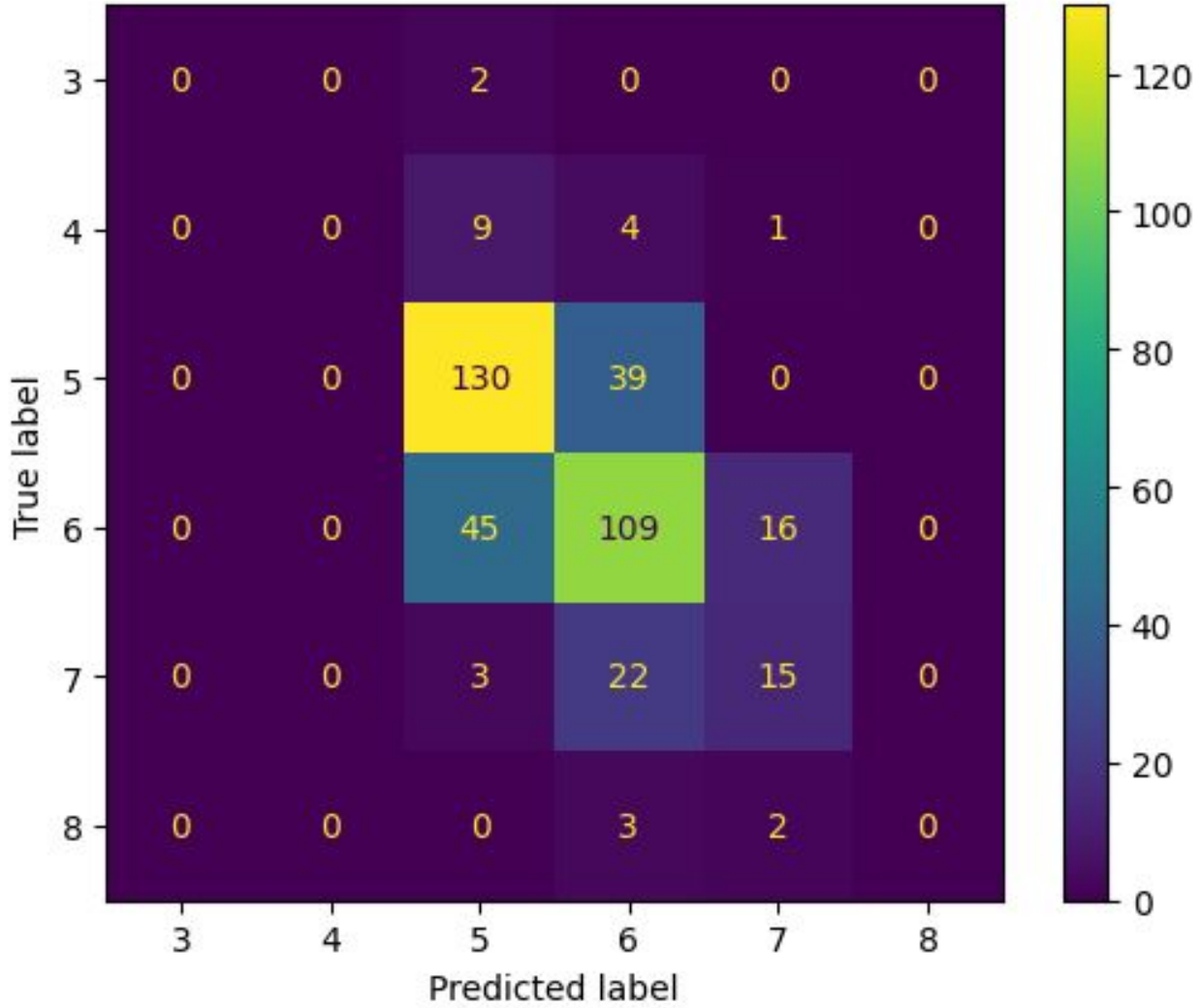
```
[[  0   0   2   0   0   0]
 [  0   0   9   4   1   0]
 [  0   0 130  39   0   0]
 [  0   0  45 109  16   0]
 [  0   0   3  22  15   0]
 [  0   0   0   3   2   0]]
Accuracy score: 0.635

              precision    recall  f1-score   support

           3       0.00      0.00      0.00         2
           4       0.00      0.00      0.00        14
           5       0.69      0.77      0.73       169
           6       0.62      0.64      0.63       170
           7       0.44      0.38      0.41        40
           8       0.00      0.00      0.00         5

    accuracy                           0.64       400
   macro avg       0.29      0.30      0.29       400
weighted avg       0.60      0.64      0.61       400

F1 score: 0.6143869978039154
```


3
4
5
6
7
8
True label
0 0 2 0 0 0
0 0 9 4 1 0
0 0 130 39 0 0
0 0 45 109 16 0
0 0 3 22 15 0
0 0 0 3 2 0
3
4
5
6
7
8
Predicted label
120
100
80
60
40
20
0

Logistic Regression

- We used a label and encoder to divide wine quality into good and bad
- Then we used all features and we set the quality as the target value and then we split the data for testing again
- Applied standard scaling
- Trained the model using the LBFGS algorithm, multinomial class and 1000 iterations again
- Got the confusion matrix, classification report, accuracy score and F1 score

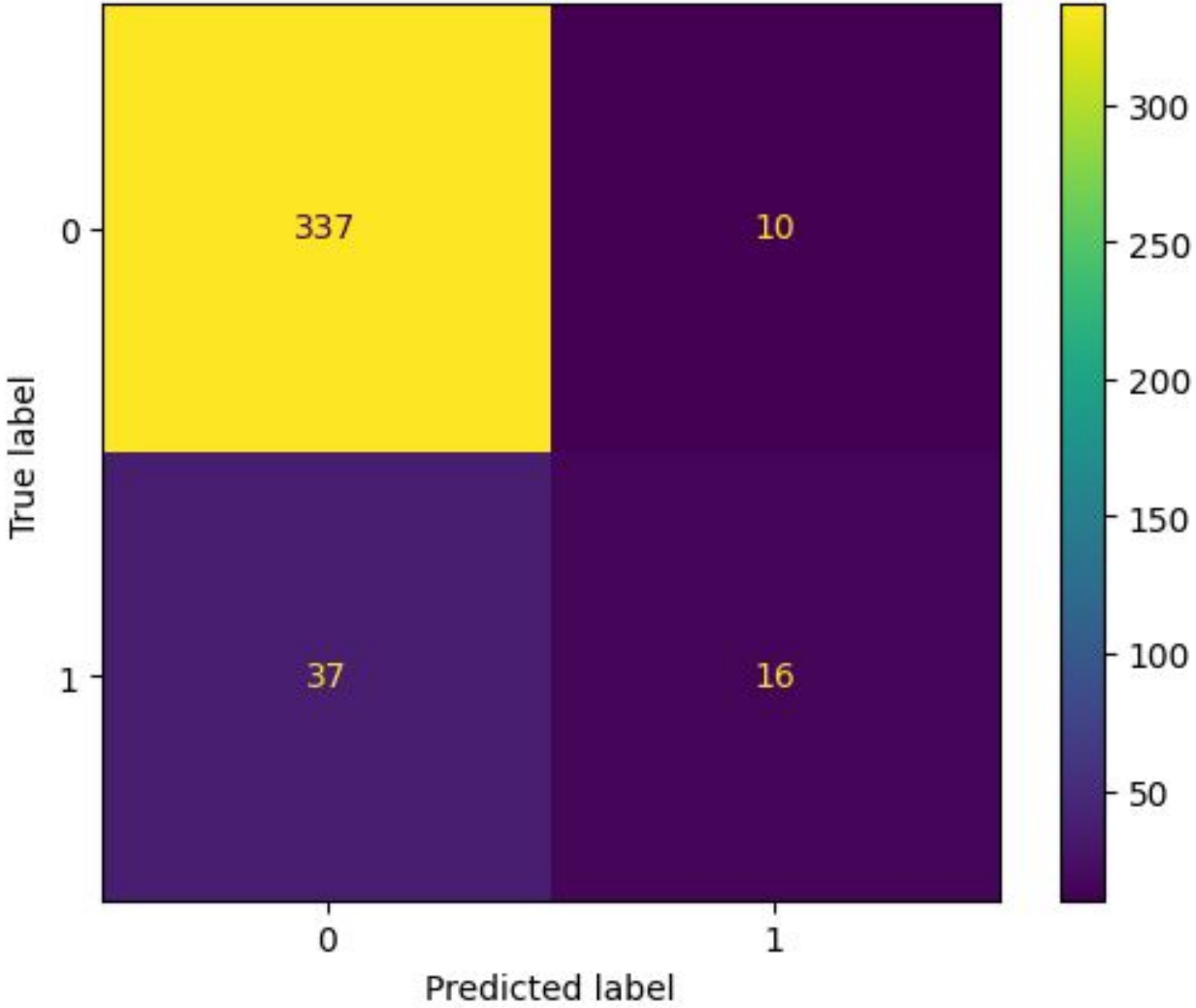
```
[[268   5]
 [ 35  12]]
Accuracy score: 0.875

              precision    recall  f1-score   support

           0       0.88      0.98      0.93       273
           1       0.71      0.26      0.37        47

    accuracy                           0.88       320
   macro avg       0.80      0.62      0.65       320
weighted avg       0.86      0.88      0.85       320

F1 score: 0.8489583333333333
```


337
10
37
16
0
1
True label
0
1
Predicted label
300
250
200
150
100
50

Insights and Conclusion

- We found out that wine types differ by acidity level and not alcohol percentage
- With our dataset, Random Forest Classification performs better than Logistic Regression because of the nature of the target value (i.e. discrete classes)
- Even with less features to work with Random Forest Classification outperforms Logistic Regression for the target values (6 features vs 10 features)
- We get more accurate results and higher precision when we labeled the target value into “good” and “bad” wine quality.
- Even when using labeling, Random Forest Classification performs better than Logistic Regression
- We could use the results from K-Means clustering for different types of wine to determine which type of wine has the highest percentage of “good” wines against “bad” wines

Thanks for watching

Questions