WINE QUALITY PREDICTION

- GROUP-9
- GROUP MEMBERS:
- B.DHARANI
- N.SHAILAJA
- MADHRAJU SURESH
- BORADO LAXMIKANTH
- B.HARSHAVARDHAN REDDY

- AURORA'S DEGREE AND PG COLLEGE
- B.SC COMP.SCIENCE
- B.SC DATA SCIENCE

AGENDA

- OBJECTIVE
- DATA DESCRIPTION
- DATA ANALYSIS
- INSIGHTS FROM ANALYSIS

OBJECTIVE

- The data set contains information about red wine and white wine
- Wine quality is measured on 0(low)-10(high)scale
- Management want

  - To understand the characteristics of these wines

  - How different ingredients affect the quality

  - To implement various types of ML Algorithms to the given data

DATA DESCRIPTION

The Attributes are as follows:

- ✓ Fixed acidity
- ✓ Volatile acidity
- ✓ Citric acid
- ✓ Residual sugar
- ✓ Chlorides
- ✓ Total sulphur dioxide
- ✓ Density
- ✓ PH
- ✓ Sulphates
- ✓ Alcohol
- ✓ Quality
- ✓ Wine type

DATA ANALYSIS

- To understand the characteristics of wine and find ingredients contributing more for quality,we have done the following:-

  - Exploratory Data Analysis(EDA)

  - Data cleaning and Data processing

  - Different types of plots like Bar plot,Scatter plot etc

  - Logistic Regression

  - Decision Tree

  - Random forest model

DATA SET

```
[ ] wine_df.head()
```

	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

DATA CLEANING

- Here we find null values by using code.Finding and dealing with null values is a crucial step in data analysis and data preprocessing.
- There are no null values in the given data set
- Then we find unique values which helps in data preprocessing

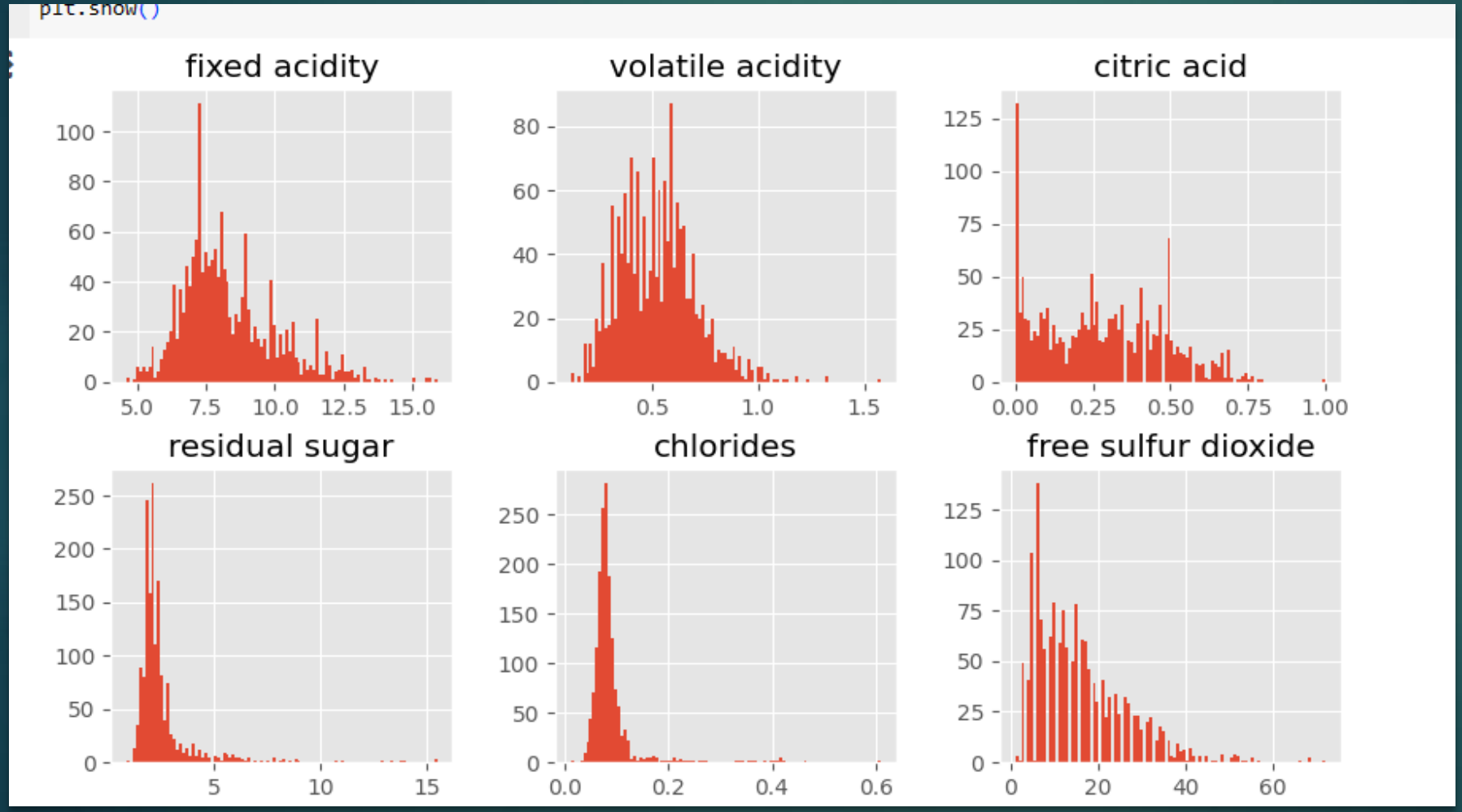
```
[ ] wine_df.isnull().sum()
```

```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
dtype: int64
```

```
[ ] wine_df.nunique()
```

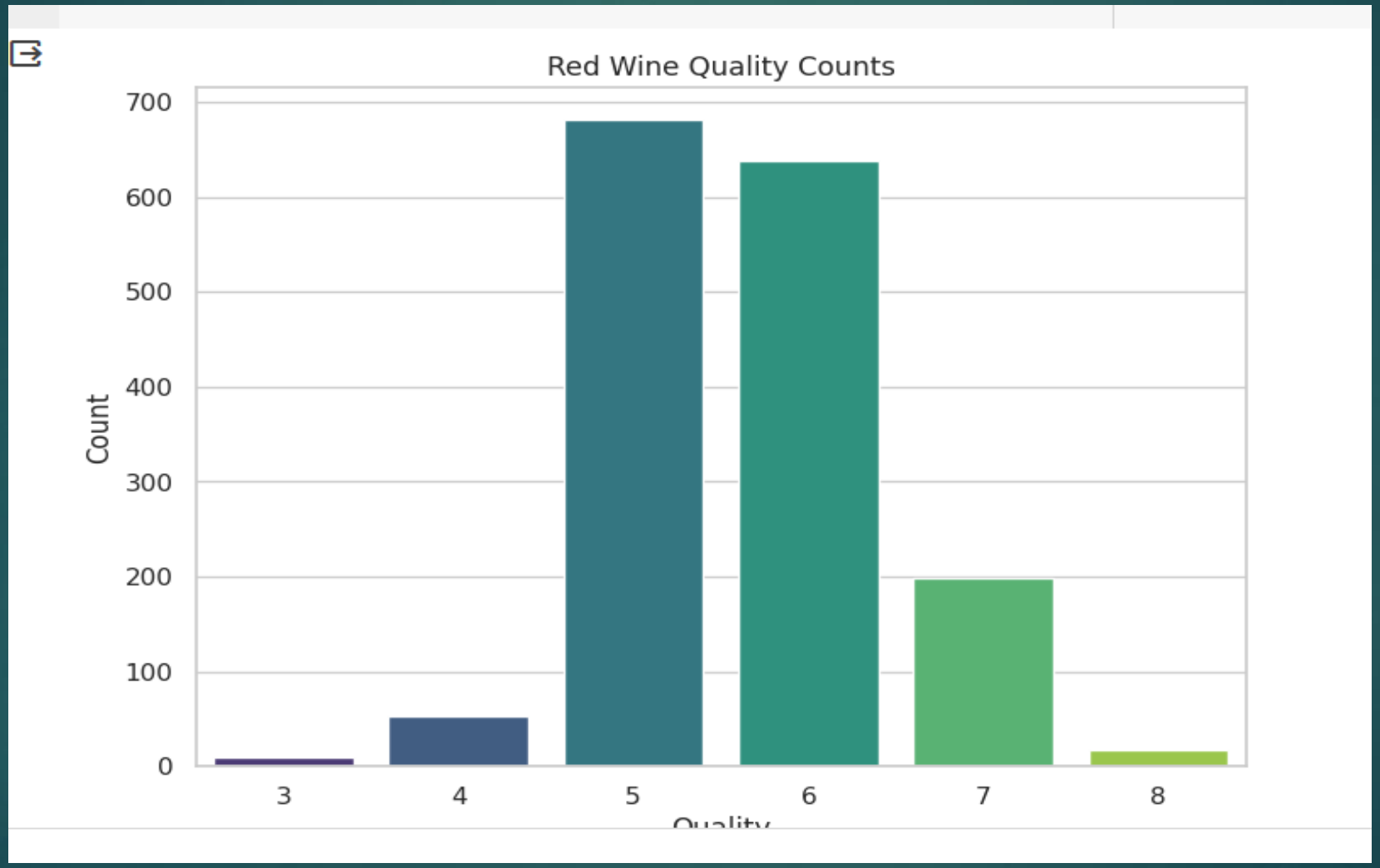
```
fixed acidity           96
volatile acidity       143
citric acid             80
residual sugar          91
chlorides              153
free sulfur dioxide     60
total sulfur dioxide   144
density                436
pH                      89
sulphates               96
alcohol                 65
quality                  6
dtype: int64
```

Exploratory Data Analysis(EDA)

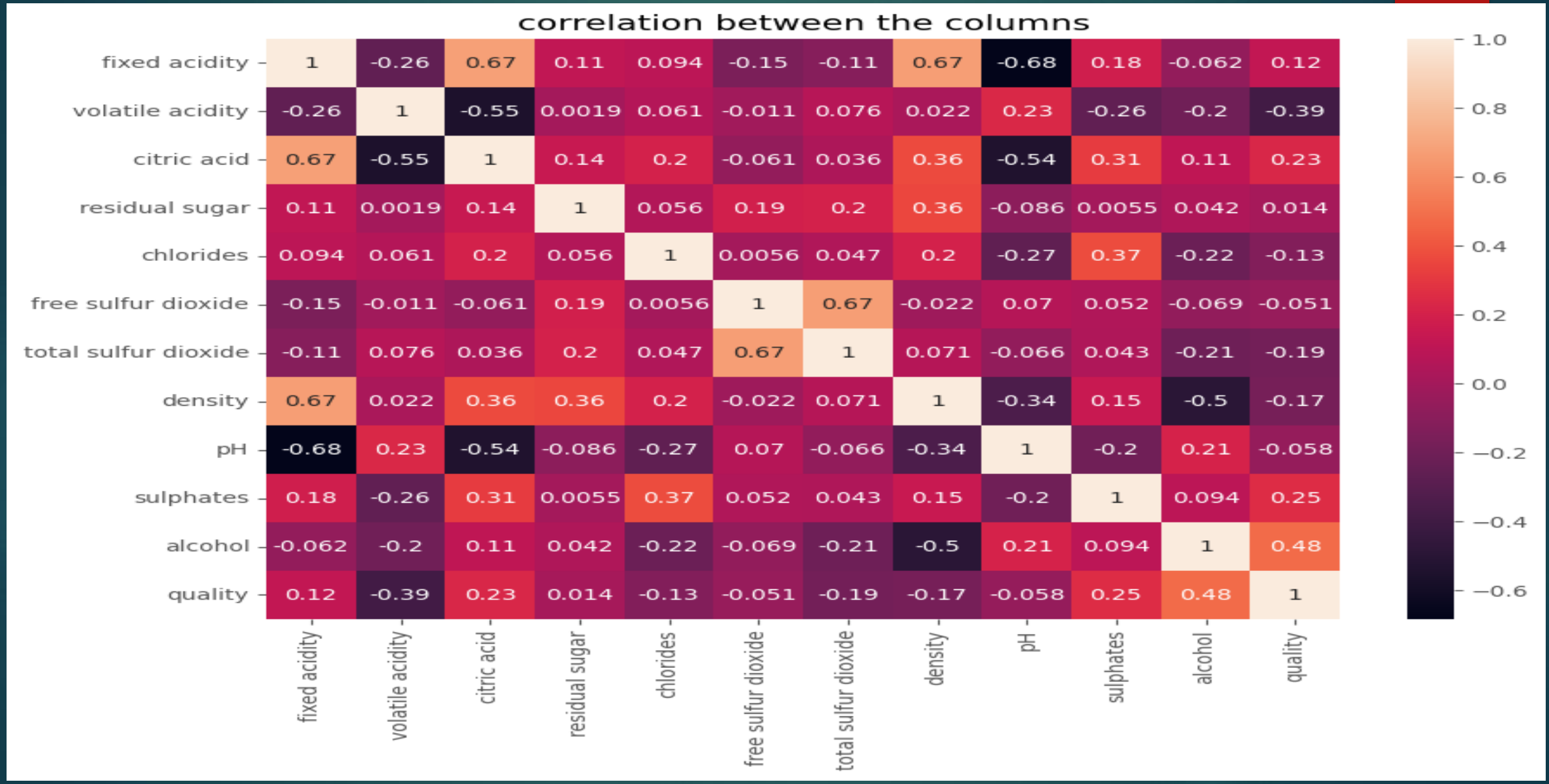


The below histogram plot shows the information about various qulities like fixed acidity,volatile acidity,citric acid,residual sugar,chlorides etc

BAR PLOT:


Red Wine Quality Counts
Count
700
600
500
400
300
200
100
0
3
4
5
6
7
8

Heat Map


correlation between the columns
fixed acidity 1 -0.26 0.67 0.11 0.094 -0.15 -0.11 0.67 -0.68 0.18 -0.062 0.12
volatile acidity -0.26 1 -0.55 0.0019 0.061 -0.011 0.076 0.022 0.23 -0.26 -0.2 -0.39
citric acid 0.67 -0.55 1 0.14 0.2 -0.061 0.036 0.36 -0.54 0.31 0.11 0.23
residual sugar 0.11 0.0019 0.14 1 0.056 0.19 0.2 0.36 -0.086 0.0055 0.042 0.014
chlorides 0.094 0.061 0.2 0.056 1 0.0056 0.047 0.2 -0.27 0.37 -0.22 -0.13
free sulfur dioxide -0.15 -0.011 -0.061 0.19 0.0056 1 0.67 -0.022 0.07 0.052 -0.069 -0.051
total sulfur dioxide -0.11 0.076 0.036 0.2 0.047 0.67 1 0.071 -0.066 0.043 -0.21 -0.19
density 0.67 0.022 0.36 0.36 0.2 -0.022 0.071 1 -0.34 0.15 -0.5 -0.17
pH -0.68 0.23 -0.54 -0.086 -0.27 0.07 -0.066 -0.34 1 -0.2 0.21 -0.058
sulphates 0.18 -0.26 0.31 0.0055 0.37 0.052 0.043 0.15 -0.2 1 0.094 0.25
alcohol -0.062 -0.2 0.11 0.042 -0.22 -0.069 -0.21 -0.5 0.21 0.094 1 0.48
quality 0.12 -0.39 0.23 0.014 -0.13 -0.051 -0.19 -0.17 -0.058 0.25 0.48 1
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

LOGISTIC REGRESSION MODEL

```
[ ]  logreg = LogisticRegression()
     logreg.fit(X_train, y_train)
     logreg_pred = logreg.predict(X_test)
     logreg_acc = accuracy_score(logreg_pred, y_test)
     print("test accuracy is:{:.2f}%".format(logreg_acc*100))
```

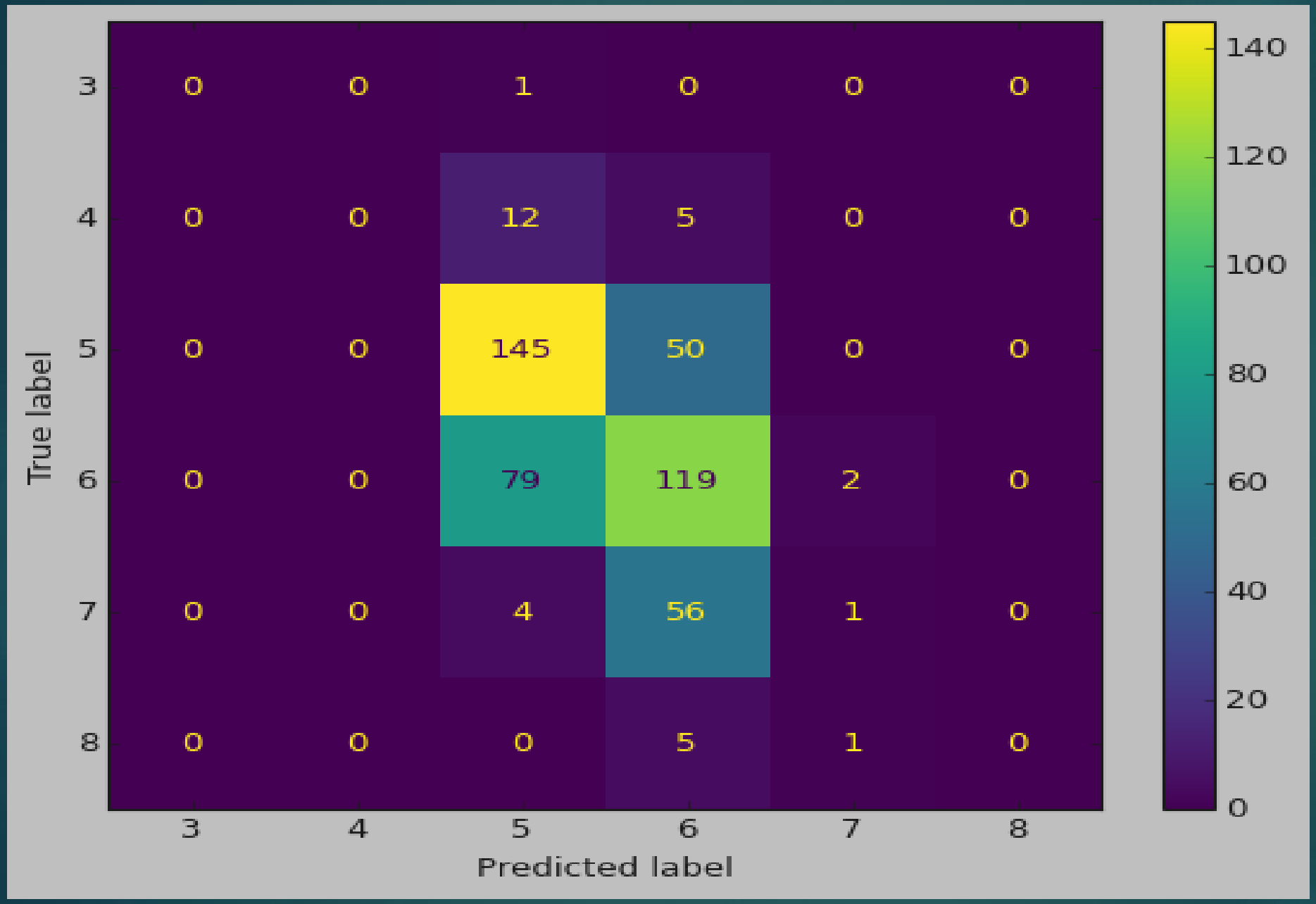
```
test accuracy is:55.21%
```

```
print(classification_report(y_test, logreg_pred))
```

```
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         1
           4       0.00      0.00      0.00        17
           5       0.60      0.74      0.67       195
           6       0.51      0.59      0.55       200
           7       0.25      0.02      0.03        61
           8       0.00      0.00      0.00         6

    accuracy                           0.55       480
   macro avg       0.23      0.23      0.21       480
weighted avg       0.49      0.55      0.50       480
```



Confusion matrix display for logistic regression model

DECISION TREE

```
[ ]  dtree = DecisionTreeClassifier()
     dtree.fit(X_train, y_train)
     dtree_pred = dtree.predict(X_test)
     dtree_acc = accuracy_score(dtree_pred, y_test)
     print("Test accuracy:{:.2f}%".format(dtree_acc*100))
```

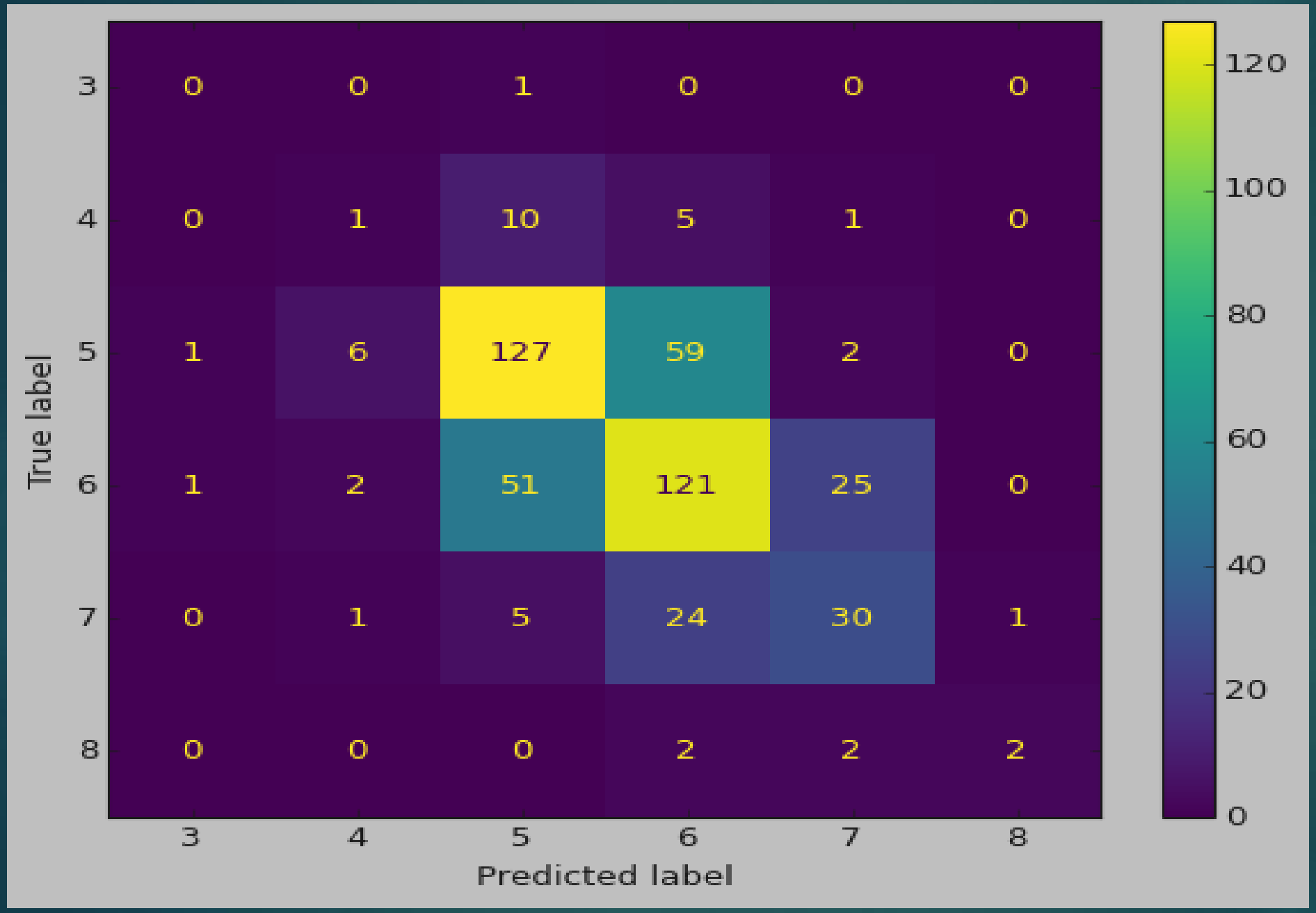
```
Test accuracy:58.54%
```

```
print(classification_report(y_test, dtree_pred))
```

```
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         1
           4       0.10      0.06      0.07        17
           5       0.65      0.65      0.65       195
           6       0.57      0.60      0.59       200
           7       0.50      0.49      0.50        61
           8       0.67      0.33      0.44         6

    accuracy                           0.59       480
   macro avg       0.42      0.36      0.38       480
weighted avg       0.58      0.59      0.58       480
```



Confusion matrix display for Decision Tree

RANDOM FOREST

```
[ ]  rforest = RandomForestClassifier()
     rforest.fit(X_train, y_train)
     rforest_pred = rforest.predict(X_test)
     rforest_acc = accuracy_score(rforest_pred, y_test)
     print("Test accuracy: {:.2f}%".format(rforest_acc*100))
```

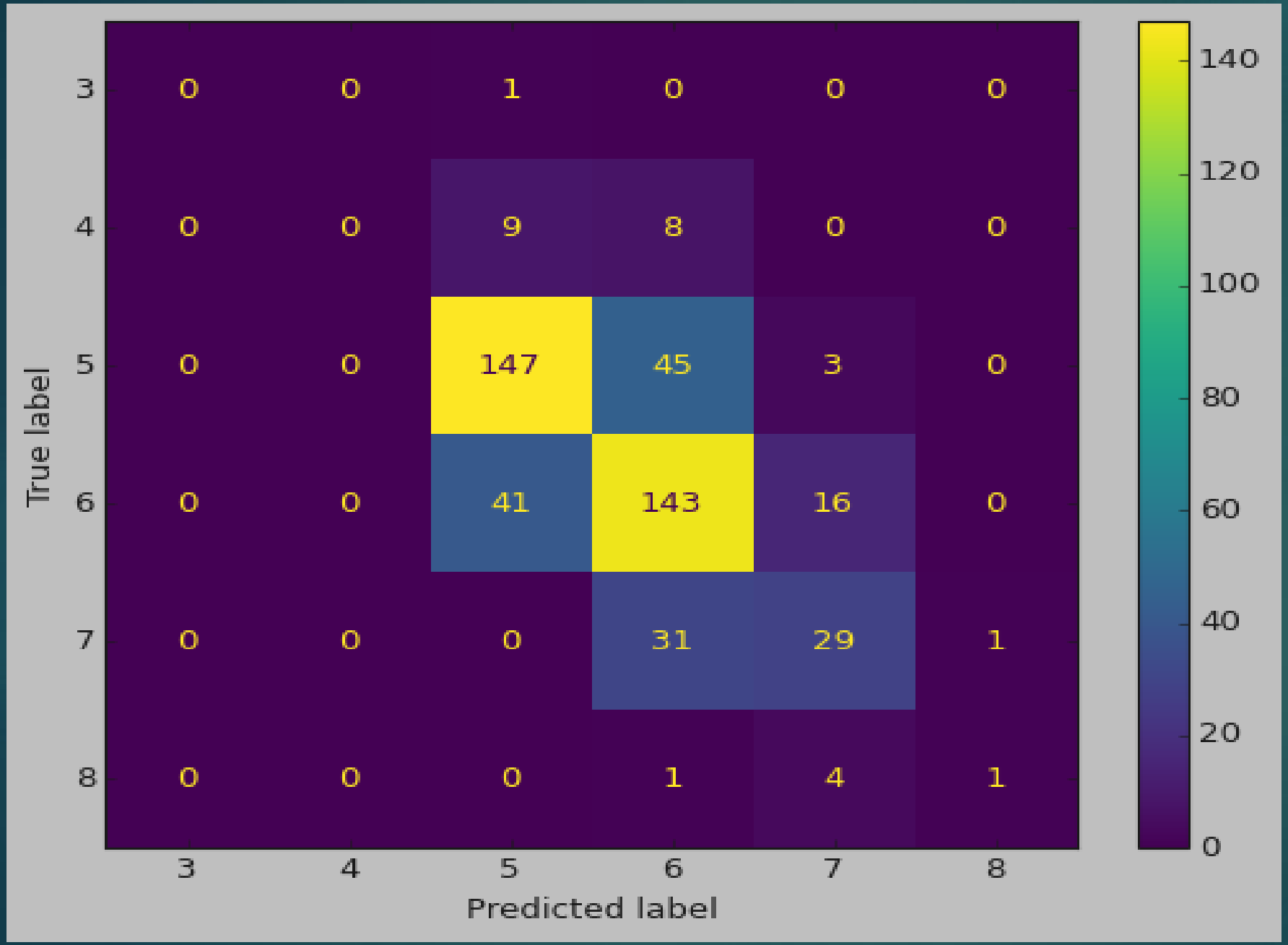
```
Test accuracy: 66.67%
```

```
[ ]  print(classification_report(y_test, rforest_pred))
```

```
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         1
           4       0.00      0.00      0.00        17
           5       0.74      0.75      0.75       195
           6       0.63      0.71      0.67       200
           7       0.56      0.48      0.51        61
           8       0.50      0.17      0.25         6

    accuracy                           0.67       480
   macro avg       0.40      0.35      0.36       480
weighted avg       0.64      0.67      0.65       480
```



Confusion matrix display for Random Forest Classifier

OBSERVATIONS

➢ The dataset was thoroughly explored, with visualizations and statistical analyses conducted to understand the data's distribution and relationships.

➢ Feature correlations were calculated, and it was observed that certain features, such as alcohol content and volatile acidity, had significant correlations with wine quality.

➢ The relationship between wine quality and various chemical attributes was explored. Box plots and violin plots showed how feature values varied across different quality levels.

➢ Classification or regression models were built to predict wine quality based on the input features. Model performance was assessed using metrics such as accuracy, precision, recall, and mean squared error

CONCLUSION

- Quality Distribution: •We can conclude that the distribution of red wine quality ratings in the dataset. The conclusion may be that the dataset contains wines of varying qualities, typically rated on a scale of 3 to 8. This distribution provides insight into the prevalence of different quality levels.
- Correlation Analysis: •Correlation plots can help identify which features have the strongest correlations with wine quality. The conclusion may be that specific chemical attributes significantly influence wine quality, while others have a weaker impact. We can conclude that the models are capable of predicting wine quality with a certain level of accuracy or precision, based on the input features

- PRESENTED BY:
- B.DHARANI
- N.SHAILAJA
- M.SURESH
- B.LAXMIKANTH
- B.HARSHAVARDHAN REDDY

Thank You

https://colab.research.google.com/drive/1RG6pCWQouZBc63uZ-7nCsGlaX3d7vs4M