Predicting Wine Type and Quality using Machine Learning Models

16.03.2023

Agenda

Topics Covered

1 About Dataset (Recap)

2 Used Mchine Learning Models

3 Sumarry

Red wine Dataset

Rows and Columns

1599, 12

Were collected

2004-2007

Missing Values

Zero

Outliers

Yes

White wine Dataset

Rows and Columns

4898, 12

Were collected

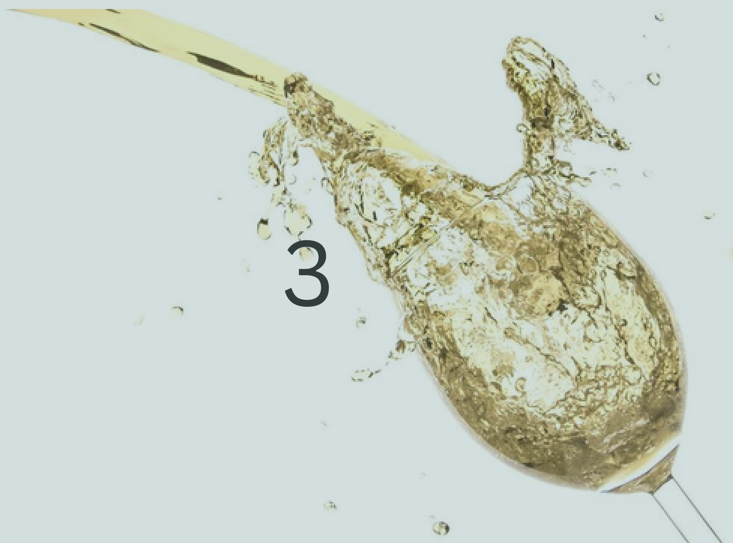
2004-2007

Missing Values

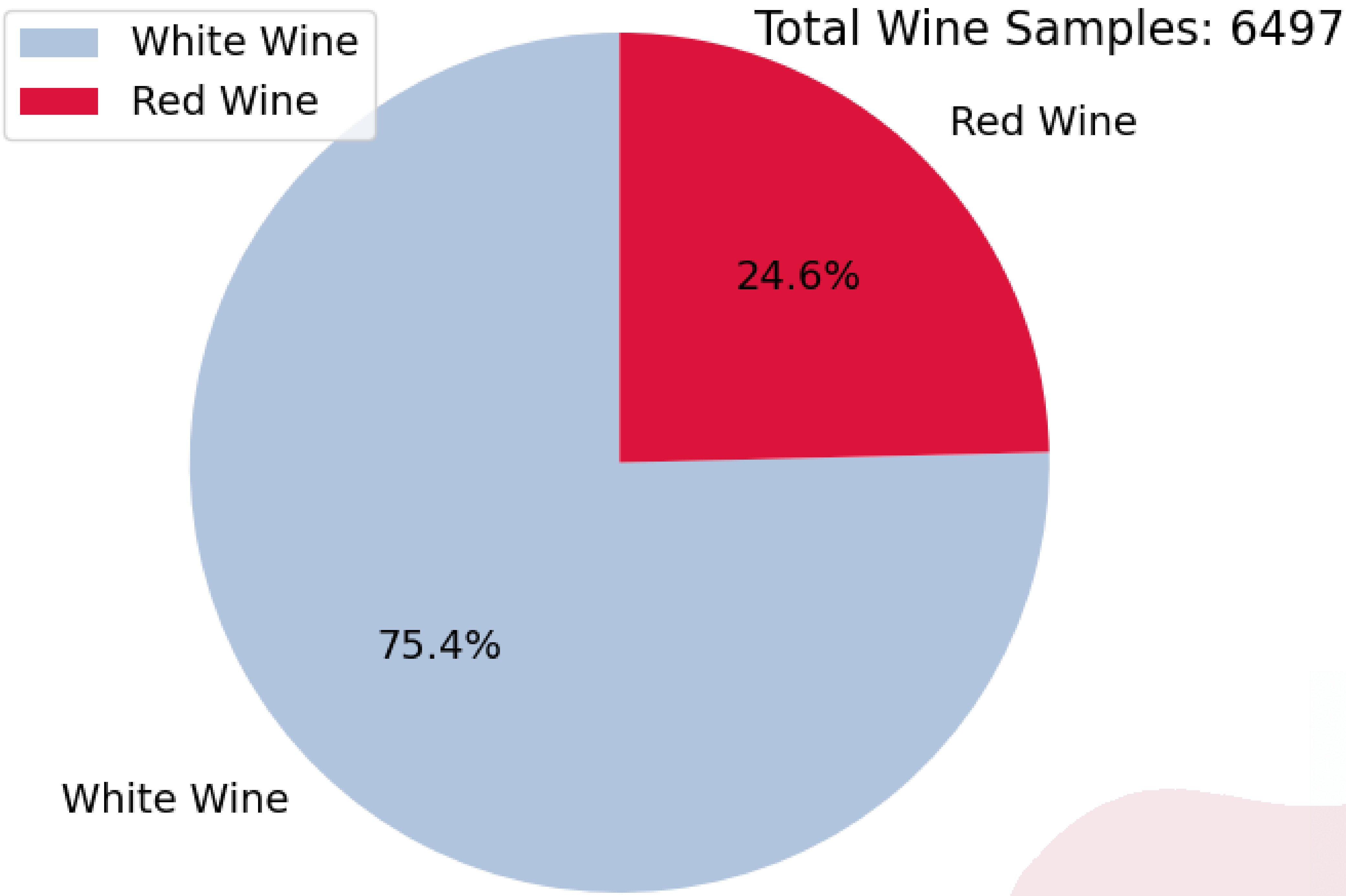
Zero

Outliers

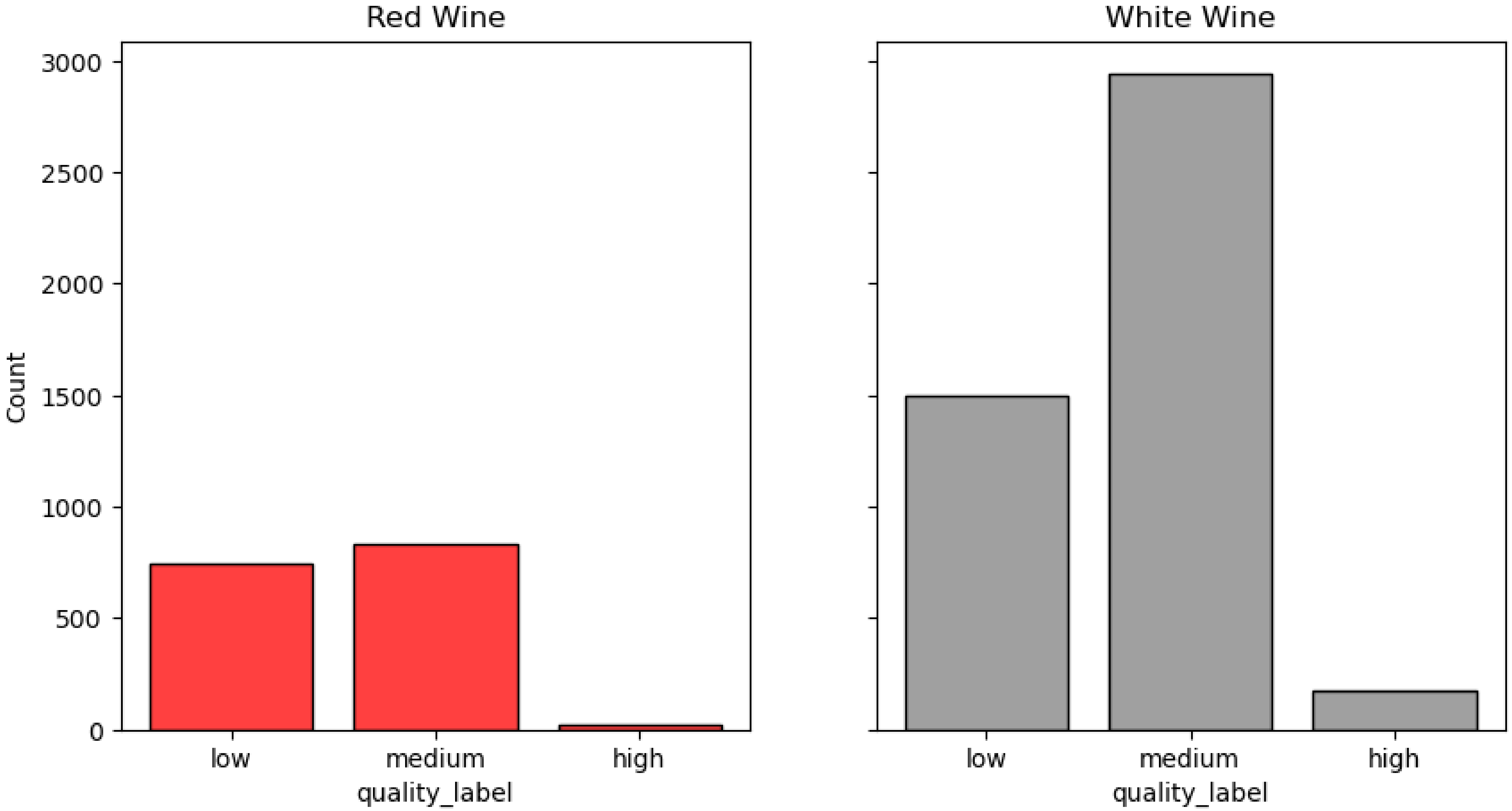
Yes



Wine Types

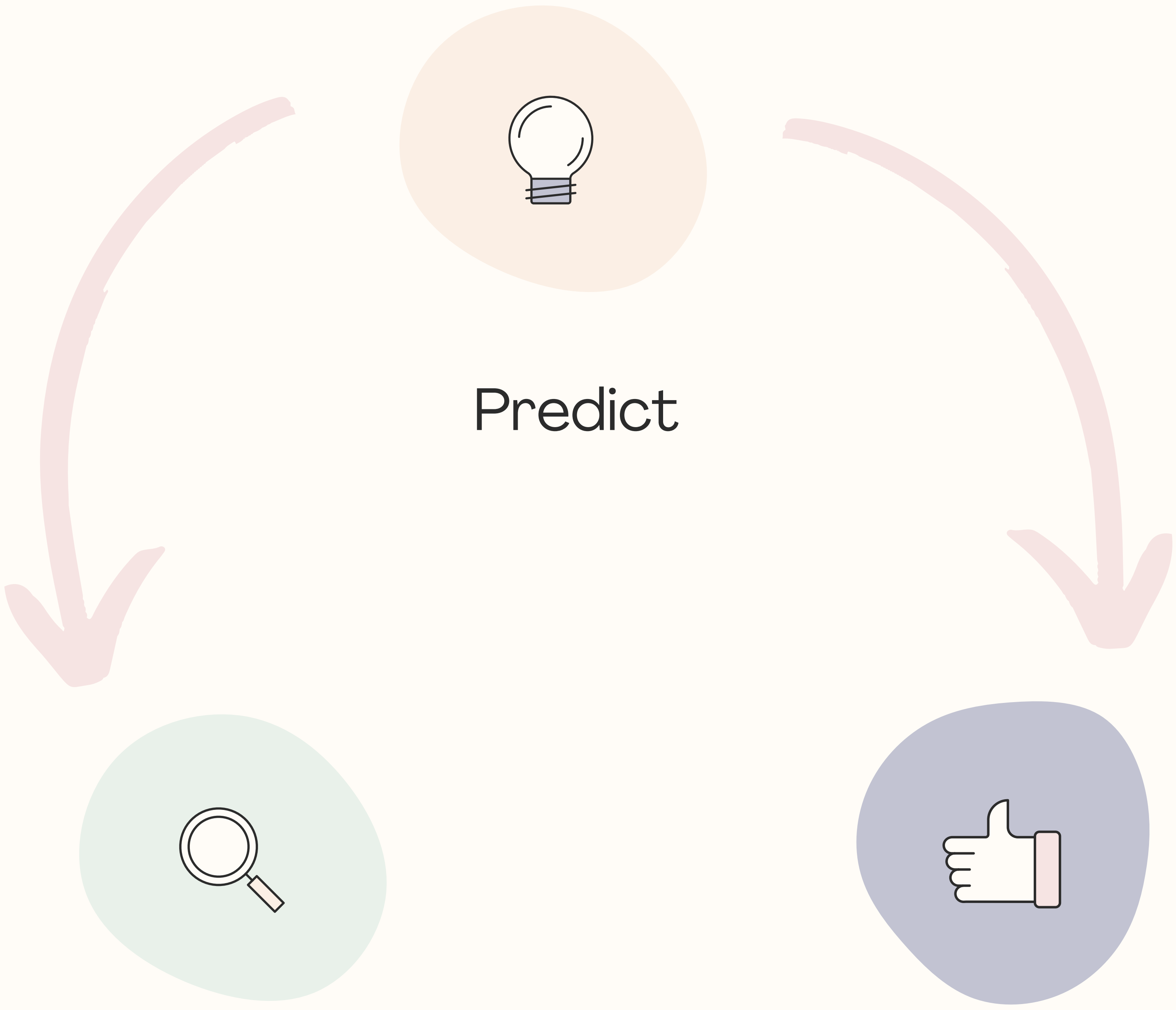

White Wine
Red Wine
Total Wine Samples: 6497
Red Wine
24.6%
75.4%
White Wine

Wine Quality Distribution By Type



Used Machine Learning Models

What My Model Should Do?



Predict

Type of Wine

Quality of Wine

Classification Models

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Naive Bayes
5. Support Vector
6. AdaBoostClassifier
7. Neural Network

Prepare The Data

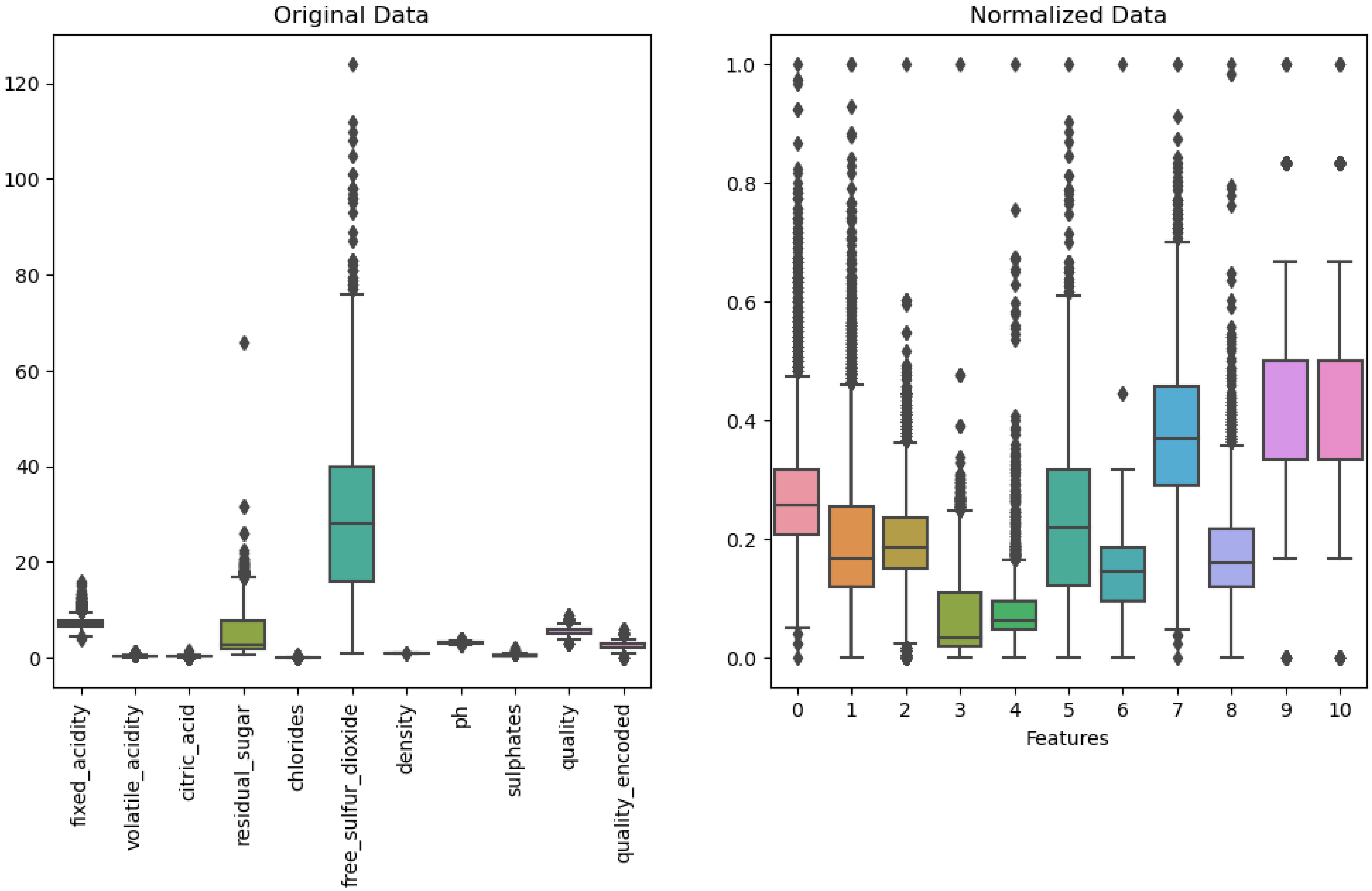
A Feature Selection

B Split the Dataset

C Cross Validation

D Normalization

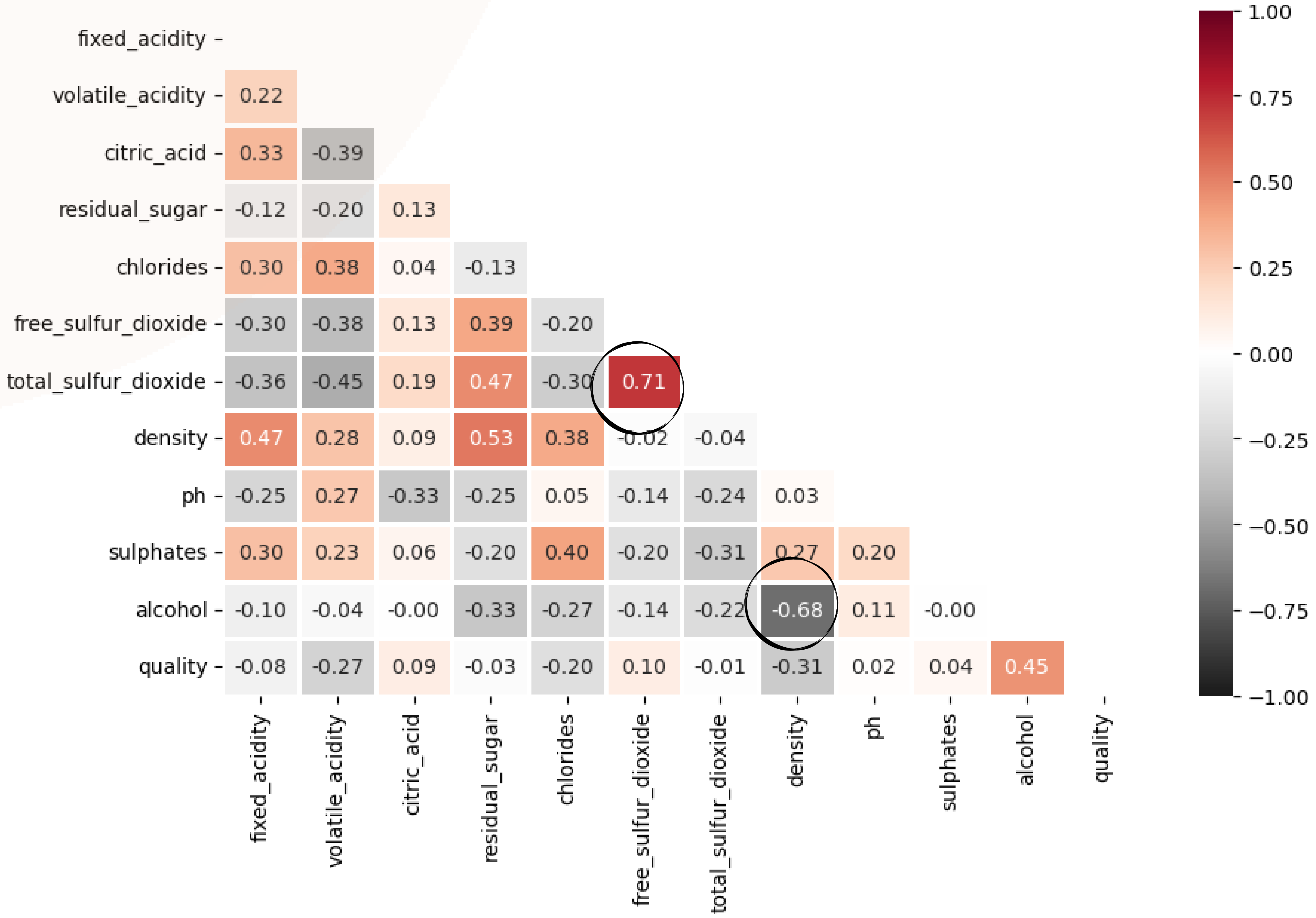
Normalization



Wine Type Prediction
(Joined both Datasets)

Feature Selection

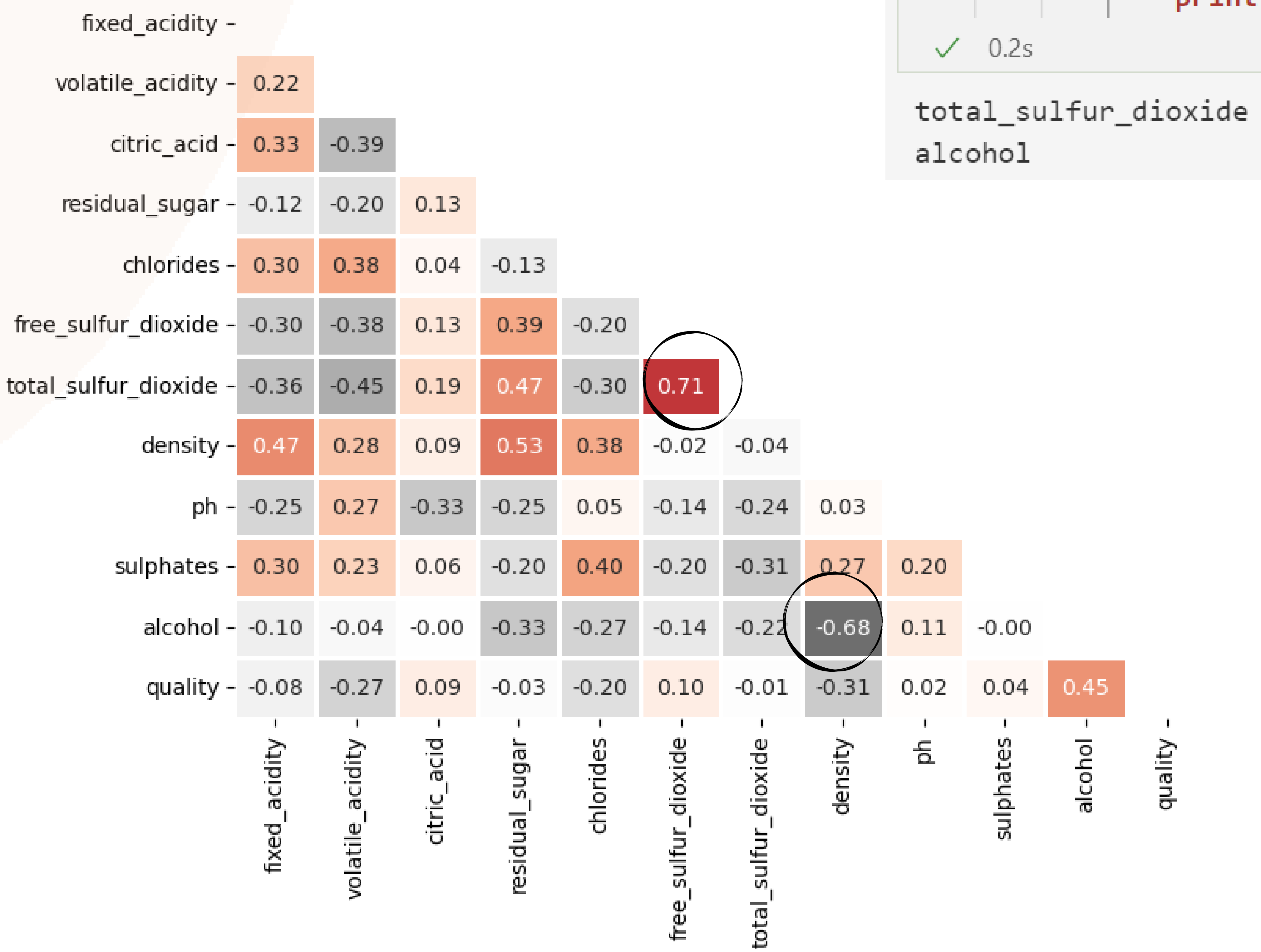
Using Correlations



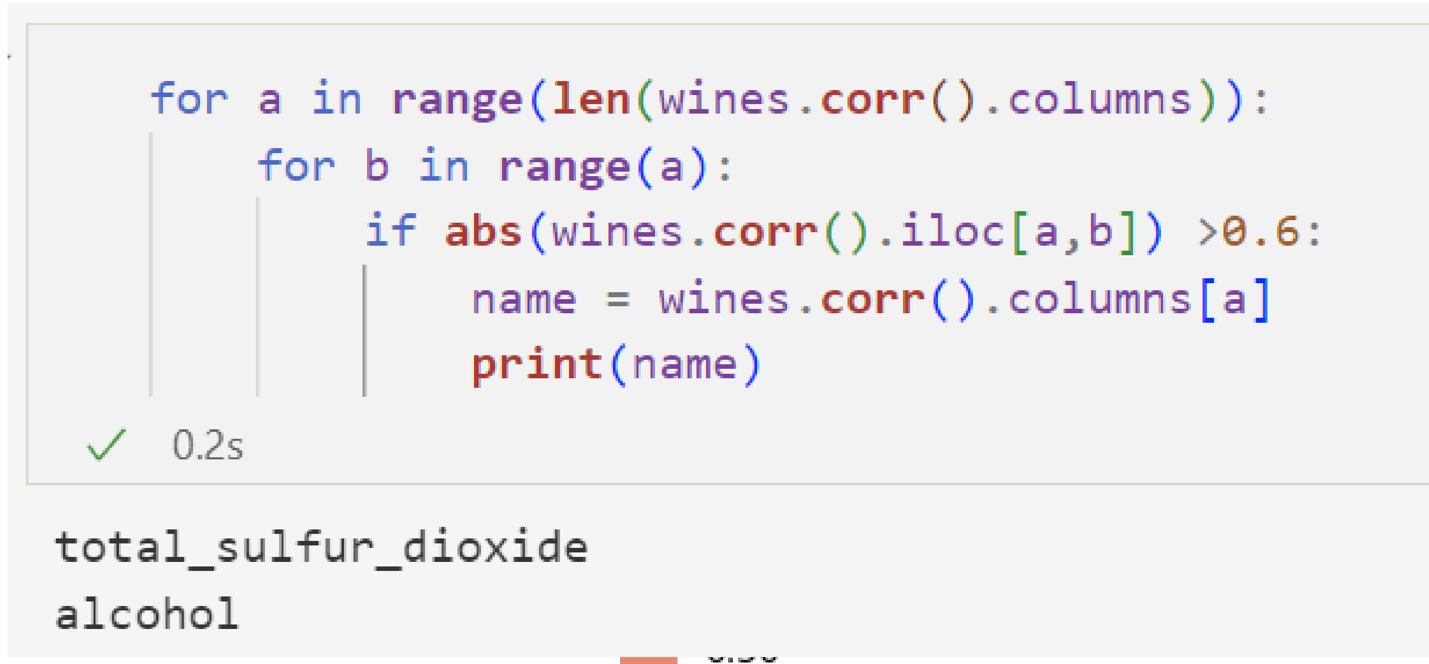
	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	ph	sulphates	alcohol	quality
fixed_acidity												
volatile_acidity	0.22											
citric_acid	0.33	-0.39										
residual_sugar	-0.12	-0.20	0.13									
chlorides	0.30	0.38	0.04	-0.13								
free_sulfur_dioxide	-0.30	-0.38	0.13	0.39	-0.20							
total_sulfur_dioxide	-0.36	-0.45	0.19	0.47	-0.30	0.71						
density	0.47	0.28	0.09	0.53	0.38	-0.02	-0.04					
ph	-0.25	0.27	-0.33	-0.25	0.05	-0.14	-0.24	0.03				
sulphates	0.30	0.23	0.06	-0.20	0.40	-0.20	-0.31	0.27	0.20			
alcohol	-0.10	-0.04	-0.00	-0.33	-0.27	-0.14	-0.22	-0.68	0.11	-0.00		
quality	-0.08	-0.27	0.09	-0.03	-0.20	0.10	-0.01	-0.31	0.02	0.04	0.45	

Feature Selection

Using Correlations



	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	ph	sulphates	alcohol	quality
fixed_acidity												
volatile_acidity	0.22											
citric_acid	0.33	-0.39										
residual_sugar	-0.12	-0.20	0.13									
chlorides	0.30	0.38	0.04	-0.13								
free_sulfur_dioxide	-0.30	-0.38	0.13	0.39	-0.20							
total_sulfur_dioxide	-0.36	-0.45	0.19	0.47	-0.30	0.71						
density	0.47	0.28	0.09	0.53	0.38	-0.02	-0.04					
ph	-0.25	0.27	-0.33	-0.25	0.05	-0.14	-0.24	0.03				
sulphates	0.30	0.23	0.06	-0.20	0.40	-0.20	-0.31	0.27	0.20			
alcohol	-0.10	-0.04	-0.00	-0.33	-0.27	-0.14	-0.22	-0.68	0.11	-0.00		
quality	-0.08	-0.27	0.09	-0.03	-0.20	0.10	-0.01	-0.31	0.02	0.04	0.45	



```
for a in range(len(wines.corr().columns)):
    for b in range(a):
        if abs(wines.corr().iloc[a,b]) >0.6:
            name = wines.corr().columns[a]
            print(name)
```

```
total_sulfur_dioxide
alcohol
```

Accuracy Measure Table (Wine Type) 1242

Actual Number of White Wine Samples = 900

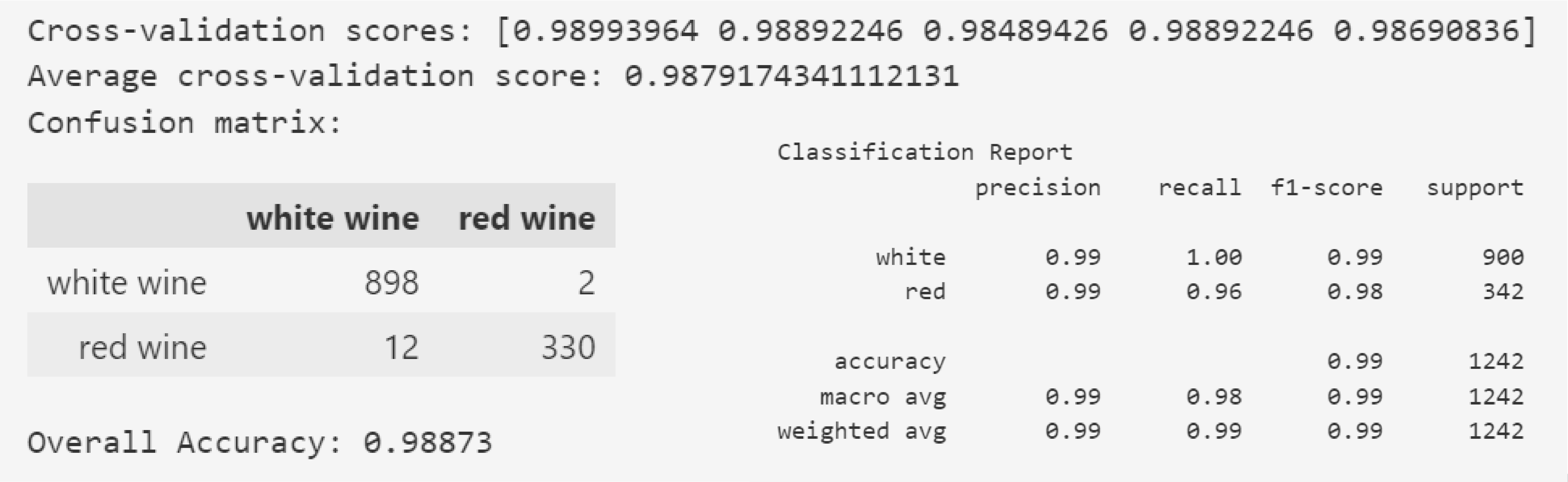
Actual Number of Red Wine Samples = 342

Model Name	Mean CV Score	Pred_White Wine	Pred_Red Wine	Overall Accuracy
Logistic Regression	0.96	887	317	0.96
Decision Tree	0.97	884	320	0.96
Random Forest	0.98	898	330	0.98
Naive Bayes	0.95	875	323	0.96
Support Vector	0.86	855	208	0.85
AdaboostClassifier	0.97	885	322	0.97
Neural Network	0.97	886	323	0.97

Accuracy Measure Table (Wine Type) 1242

Model Name	Mean CV Score	Pred_White Wine	Pred_Red Wine	Overall Accuracy (Test Accuracy)
Logistic Regression	0.96	887	317	0.96
Decision Tree	0.97	884	320	0.96
Random Forest	0.98	898	330	0.98
Naive Bayes	0.95	875	323	0.96
Support Vector	0.86	855	208	0.85
AdaboostClassifier	0.97	885	322	0.97
Neural Network	0.97	886	323	0.97

Random Forest (Wine Type)



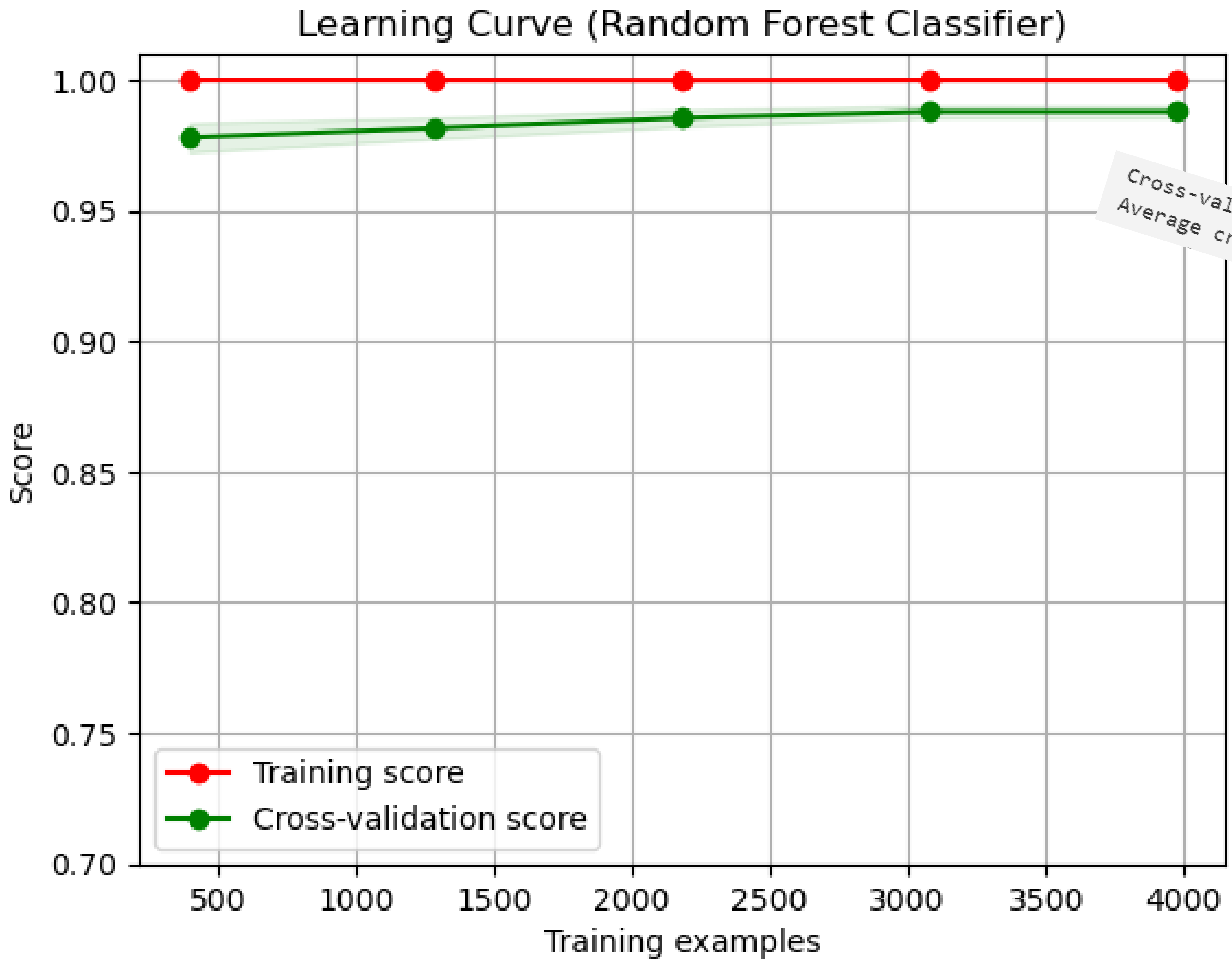
```
Cross-validation scores: [0.98993964 0.98892246 0.98489426 0.98892246 0.98690836]
Average cross-validation score: 0.9879174341112131
Confusion matrix:
```

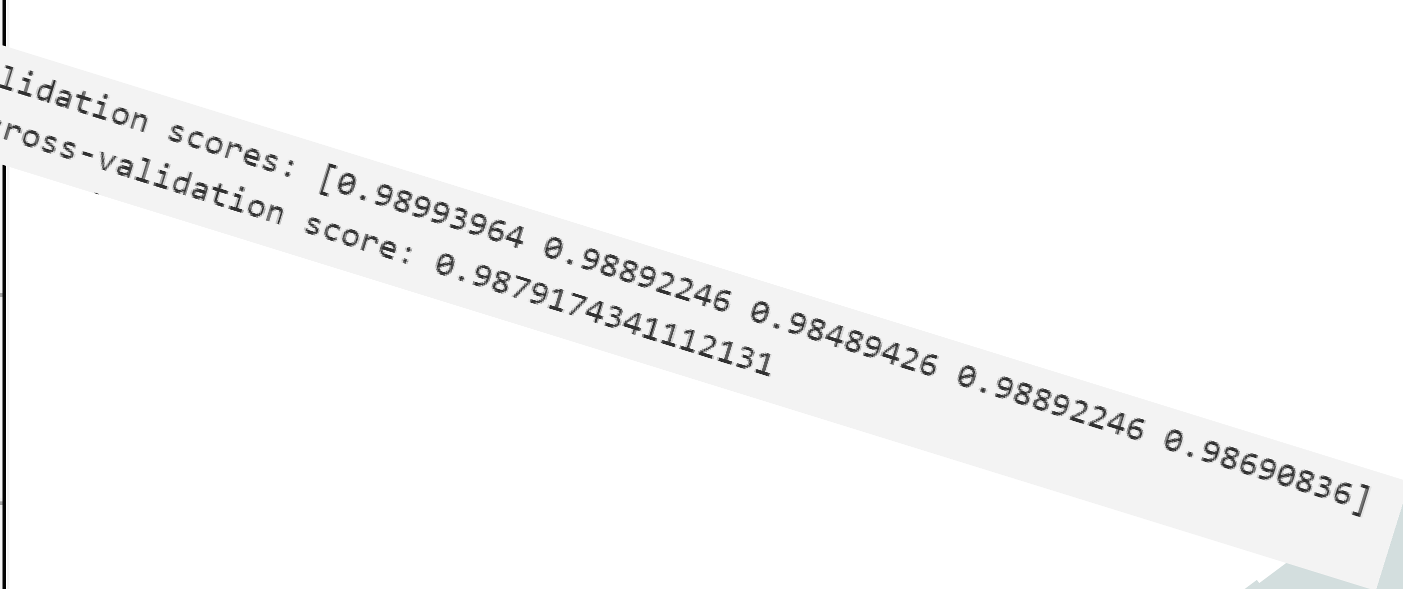
	white wine	red wine
white wine	898	2
red wine	12	330

```
Overall Accuracy: 0.98873
```

```
Classification Report
              precision    recall  f1-score   support

       white       0.99      1.00      0.99       900
         red       0.99      0.96      0.98       342

    accuracy                           0.99      1242
   macro avg       0.99      0.98      0.99      1242
weighted avg       0.99      0.99      0.99      1242
```

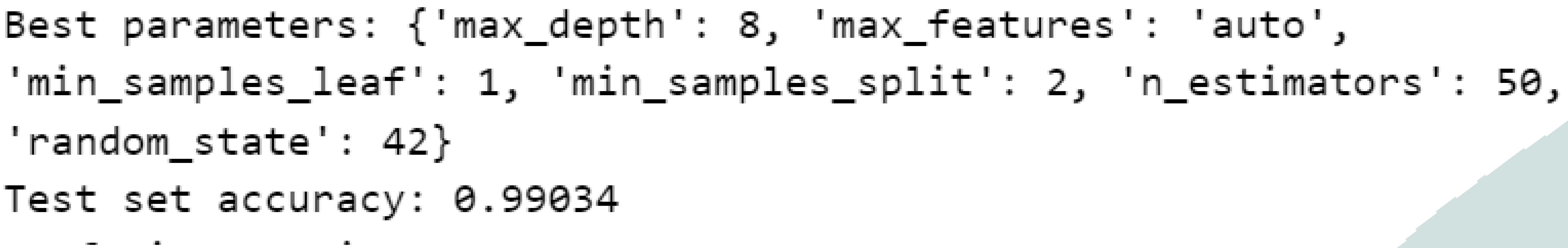
Learning Curve (Whine Type)



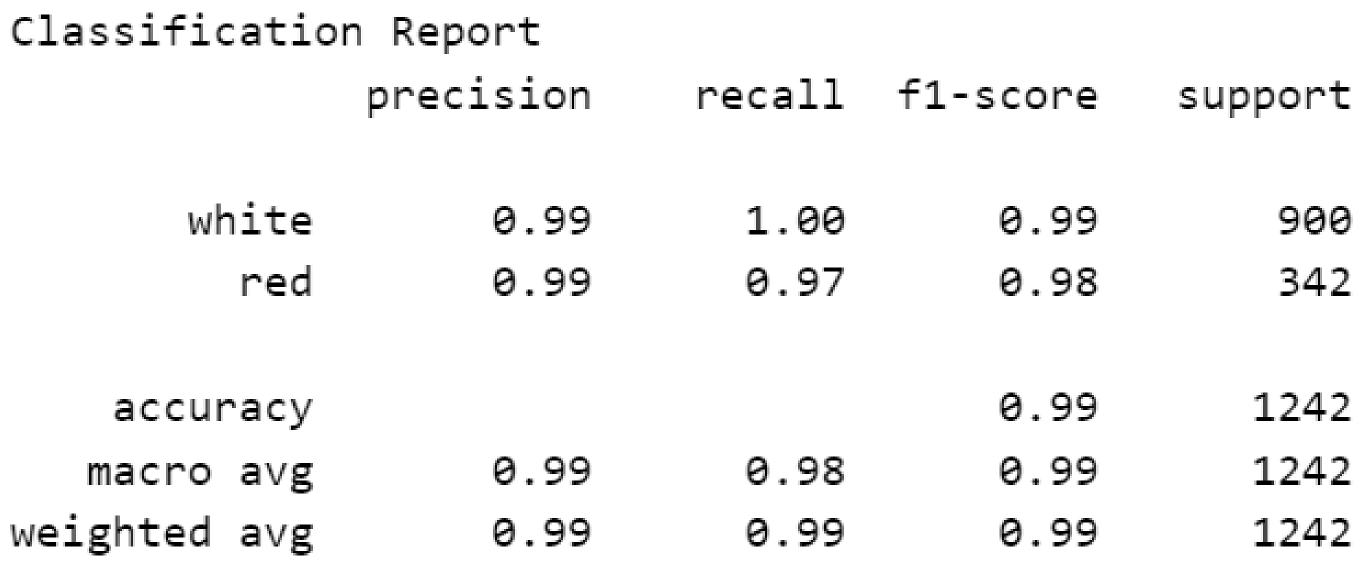


```
Cross-validation scores: [0.98993964 0.98892246 0.98489426 0.98892246 0.98690836]
Average cross-validation score: 0.9879174341112131
```

Improved Model



```
Best parameters: {'max_depth': 8, 'max_features': 'auto',
'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50,
'random_state': 42}
Test set accuracy: 0.99034
```



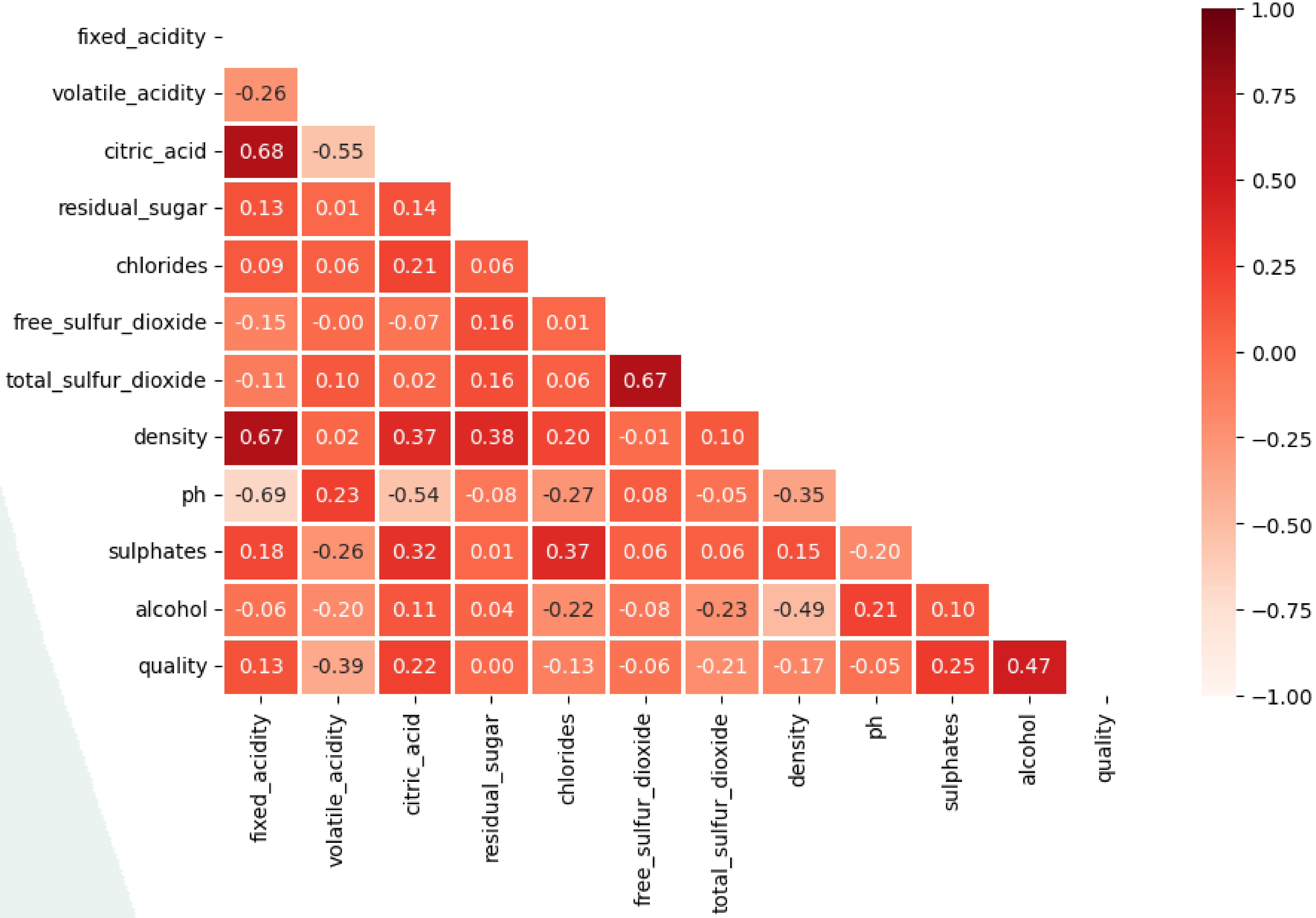
```
Classification Report
              precision    recall  f1-score   support

       white       0.99      1.00      0.99       900
         red       0.99      0.97      0.98       342

    accuracy                           0.99      1242
   macro avg       0.99      0.98      0.99      1242
weighted avg       0.99      0.99      0.99      1242
```

Wine Quality Prediction

Red Wine Quality Prediction

Feature Selection

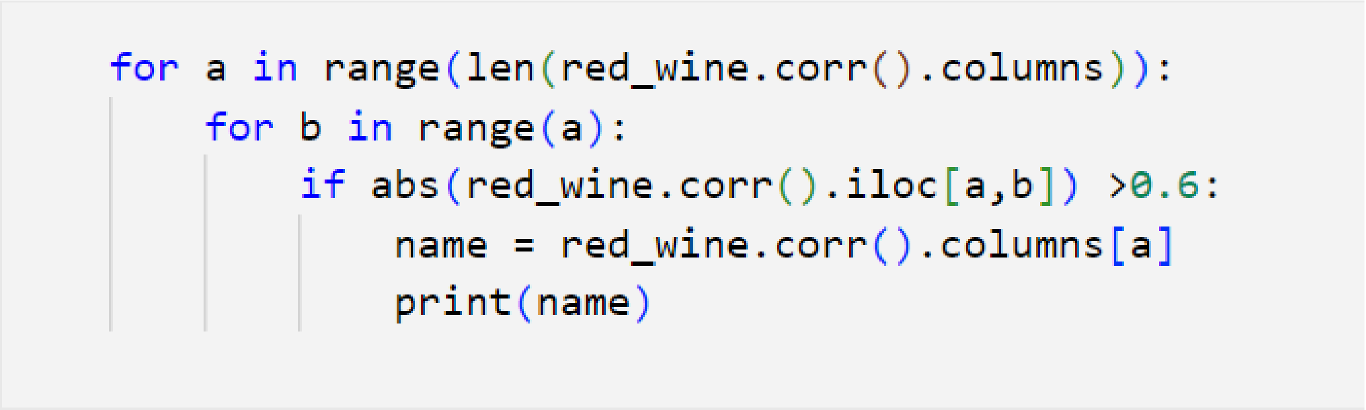
Using Correlations



	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	ph	sulphates	alcohol	quality
fixed_acidity												
volatile_acidity	-0.26											
citric_acid	0.68	-0.55										
residual_sugar	0.13	0.01	0.14									
chlorides	0.09	0.06	0.21	0.06								
free_sulfur_dioxide	-0.15	-0.00	-0.07	0.16	0.01							
total_sulfur_dioxide	-0.11	0.10	0.02	0.16	0.06	0.67						
density	0.67	0.02	0.37	0.38	0.20	-0.01	0.10					
ph	-0.69	0.23	-0.54	-0.08	-0.27	0.08	-0.05	-0.35				
sulphates	0.18	-0.26	0.32	0.01	0.37	0.06	0.06	0.15	-0.20			
alcohol	-0.06	-0.20	0.11	0.04	-0.22	-0.08	-0.23	-0.49	0.21	0.10		
quality	0.13	-0.39	0.22	0.00	-0.13	-0.06	-0.21	-0.17	-0.05	0.25	0.47	

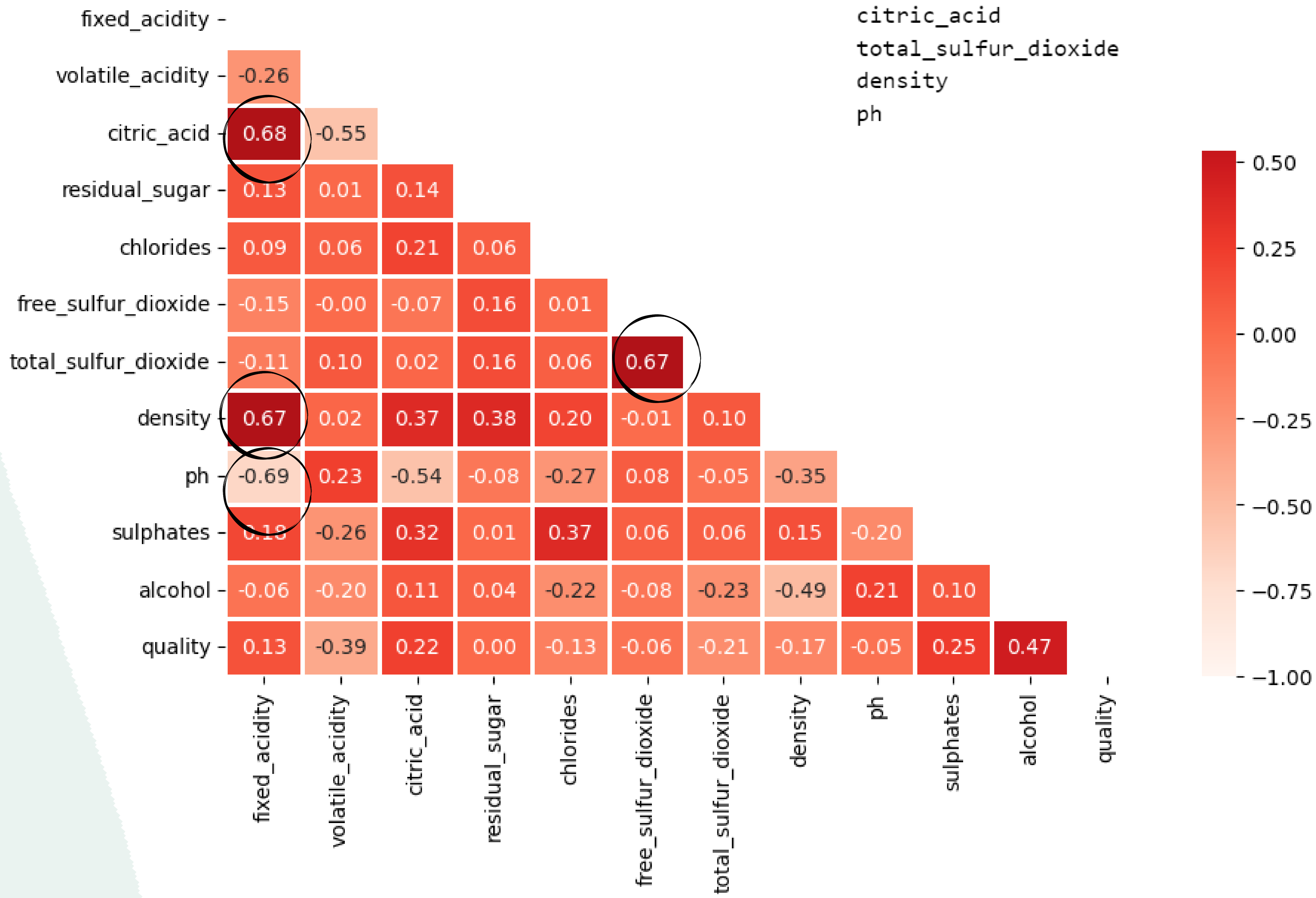
Feature Selection

Using Correlations



```
for a in range(len(red_wine.corr().columns)):
    for b in range(a):
        if abs(red_wine.corr().iloc[a,b]) >0.6:
            name = red_wine.corr().columns[a]
            print(name)
```

```
citric_acid
total_sulfur_dioxide
density
ph
```



Accuracy Measure Table (Red Wine) 319

Low = 150, Medium = 129 and High = 40

Model Name	Mean CV Score	Low	Medium	High	Overall Accuracy
Logistic Regression	0.60	125	71	6	0.63
Decision Tree	0.64	102	80	16	0.62
Random Forest	0.72	115	80	17	0.66
Naive Bayes	0.60	113	61	25	0.62
Support Vector	0.62	122	74	13	0.65
AdaboostClassifier	0.65	98	78	17	0.60
Neural Network	0.61	121	70	13	0.63

Accuracy Measure Table (Red Wine) 319

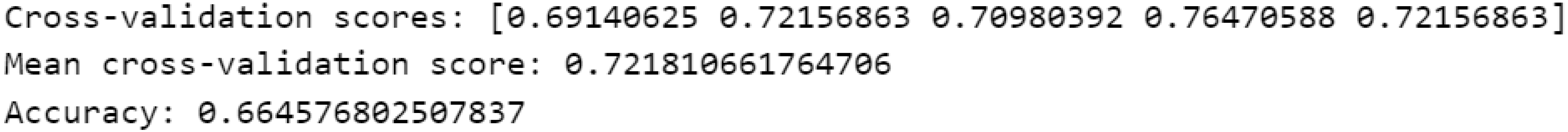
Low = 150, Medium = 129 and High = 40

Model Name	Mean CV Score	Low	Medium	High	Overall Accuracy
Logistic Regression	0.60	125	71	6	0.63
Decision Tree	0.64	102	80	16	0.62
Random Forest	0.72	115	80	17	0.66
Naive Bayes	0.60	113	61	25	0.62
Support Vector	0.62	122	74	13	0.65
AdaboostClassifier	0.65	98	78	17	0.60
Neural Network	0.61	121	70	13	0.63

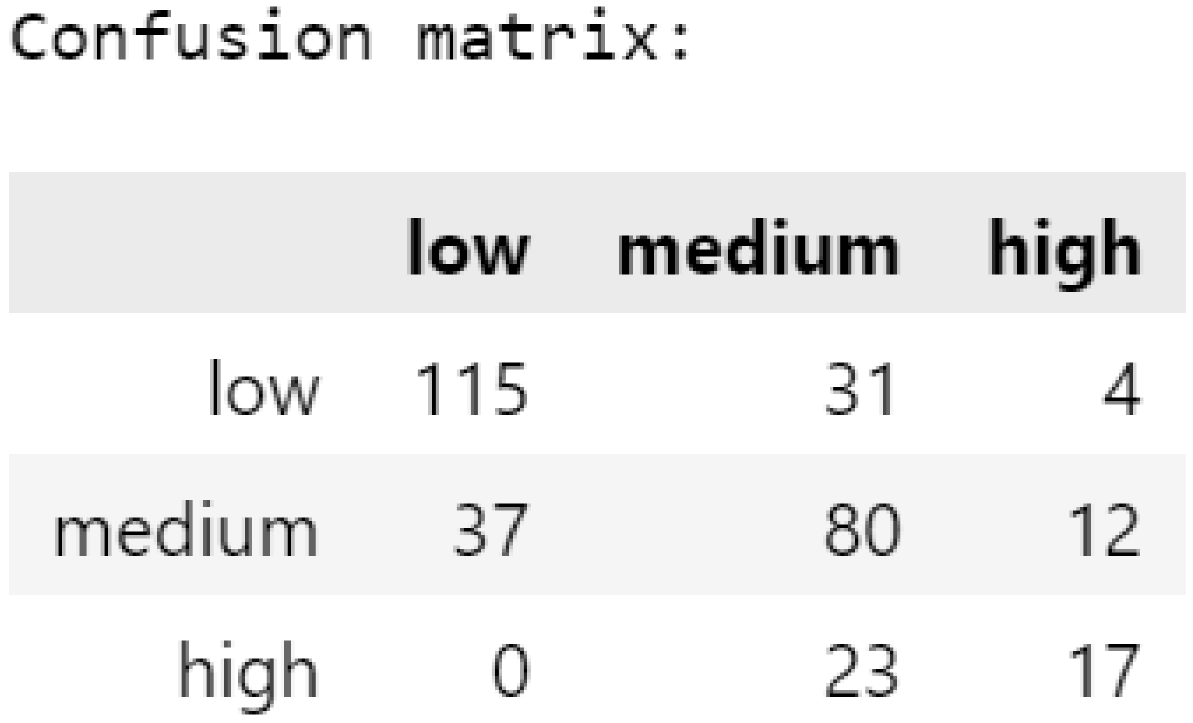
Accuracy Measure Table (Red Wine) 319

Model Name	Mean CV Score	Low	Medium	High	Overall Accuracy
Logistic Regression	0.60	125	71	6	0.63
Decision Tree	0.64	102	80	16	0.62
Random Forest	0.72	115	80	17	0.66
Naive Bayes	0.60	113	61	25	0.62
Support Vector	0.62	122	74	13	0.65
AdaboostClassifier	0.65	98	78	17	0.60
Neural Network	0.61	121	70	13	0.63

Random Forest (Red Wine)

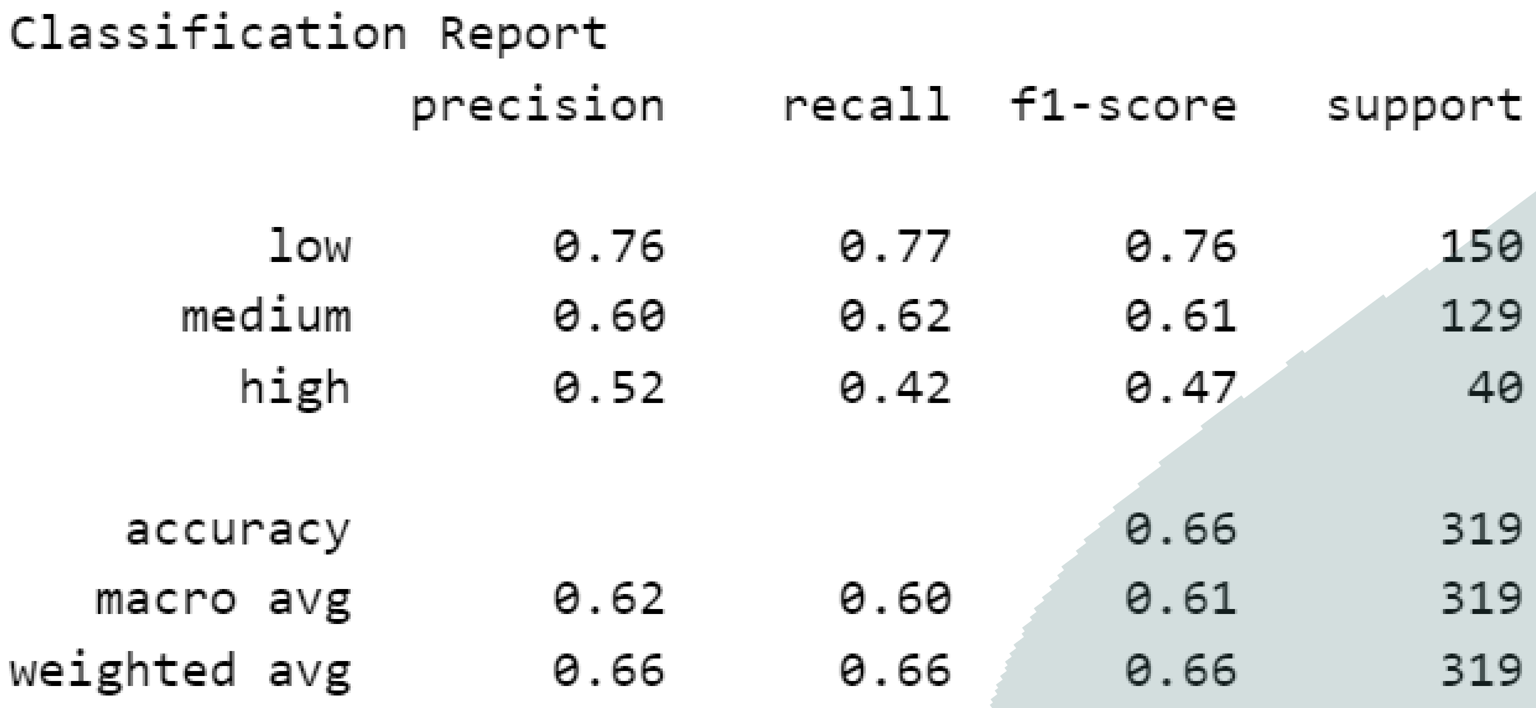


```
Cross-validation scores: [0.69140625 0.72156863 0.70980392 0.76470588 0.72156863]
Mean cross-validation score: 0.721810661764706
Accuracy: 0.664576802507837
```



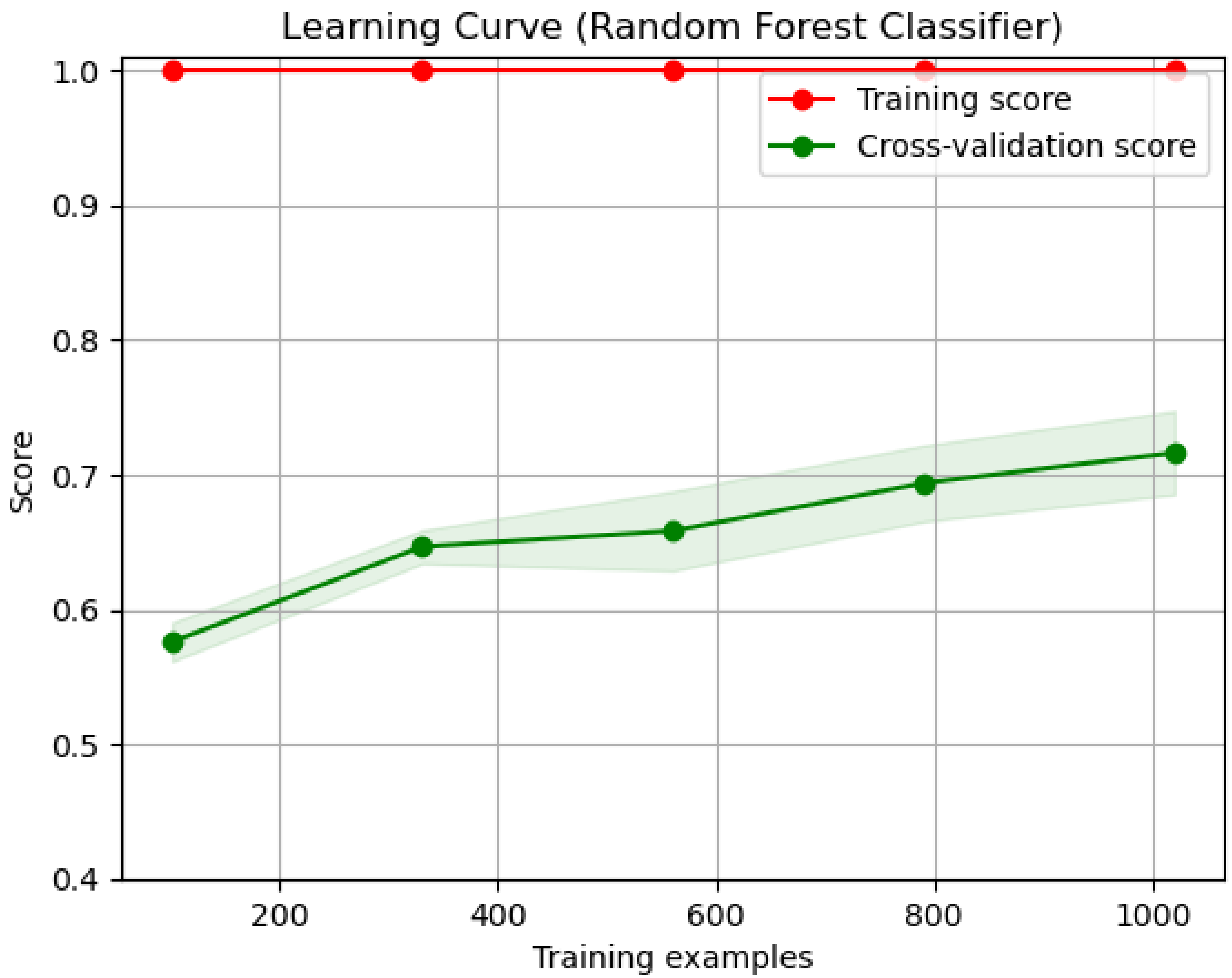
```
Confusion matrix:
```

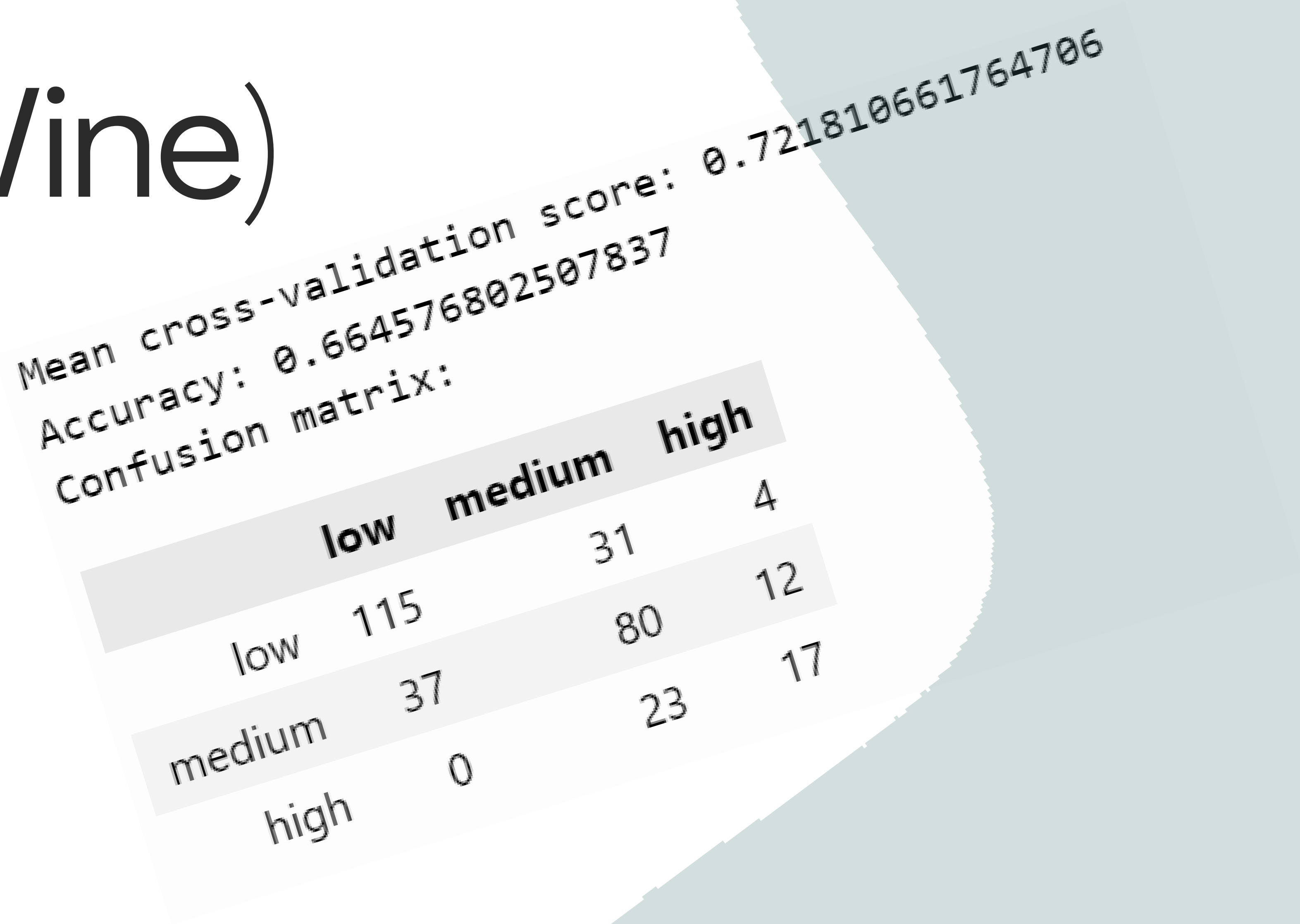
	low	medium	high
low	115	31	4
medium	37	80	12
high	0	23	17



```
Classification Report
              precision    recall  f1-score   support

         low       0.76      0.77      0.76       150
      medium       0.60      0.62      0.61       129
        high       0.52      0.42      0.47        40

    accuracy                           0.66       319
   macro avg       0.62      0.60      0.61       319
weighted avg       0.66      0.66      0.66       319
```

Learning Curve (Red Wine)

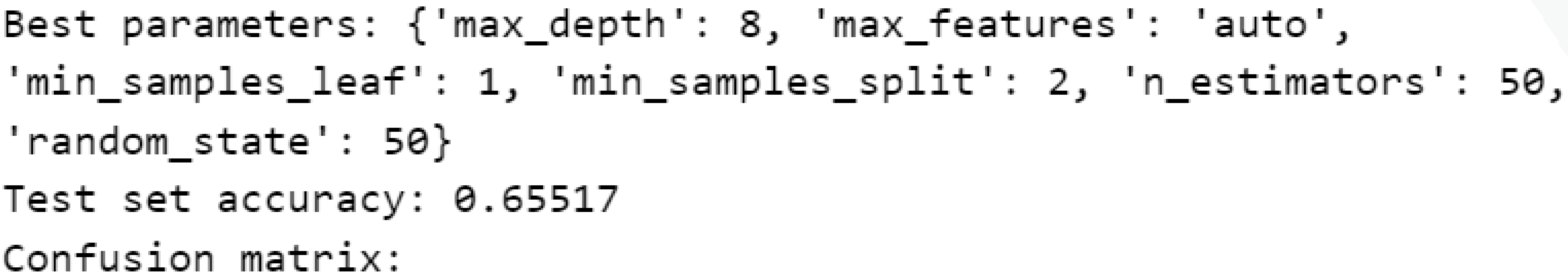




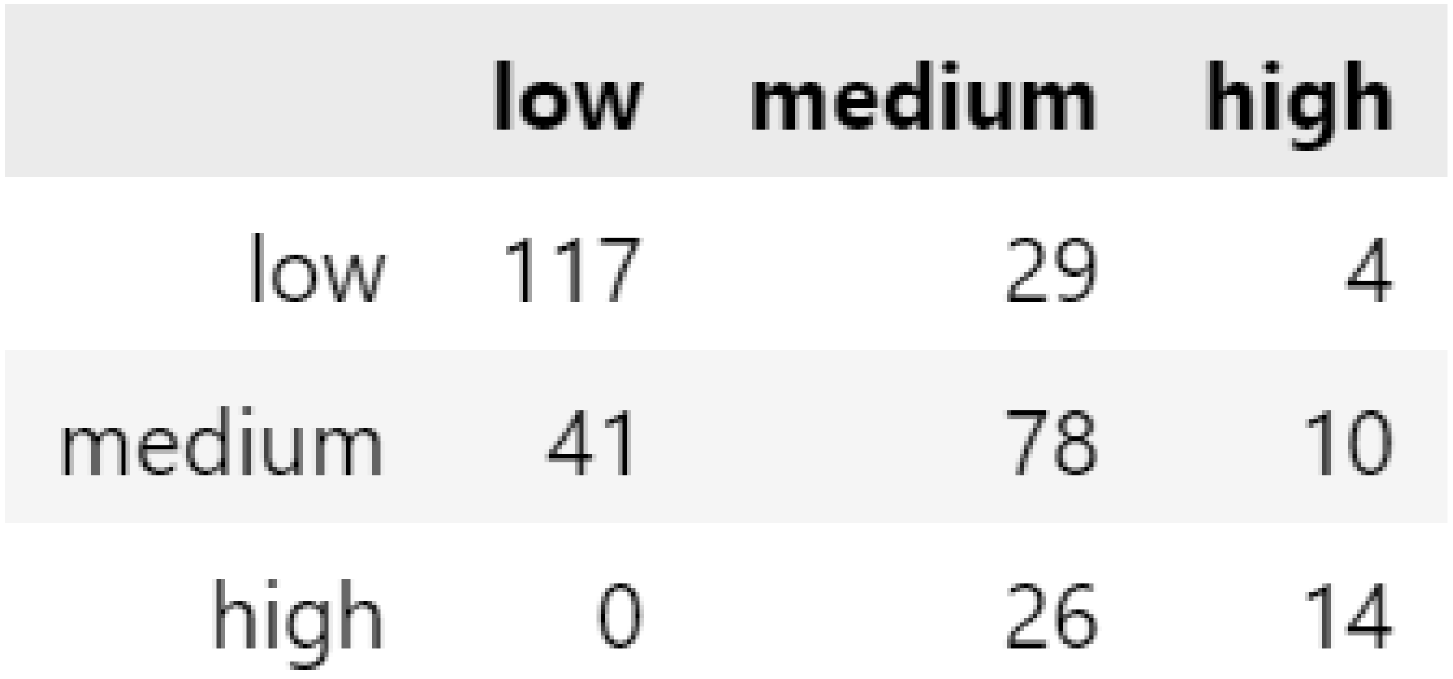
```
Mean cross-validation score: 0.721810661764706
Accuracy: 0.664576802507837
Confusion matrix:
```

	low	medium	high
low	115	31	4
medium	37	80	12
high	0	23	17

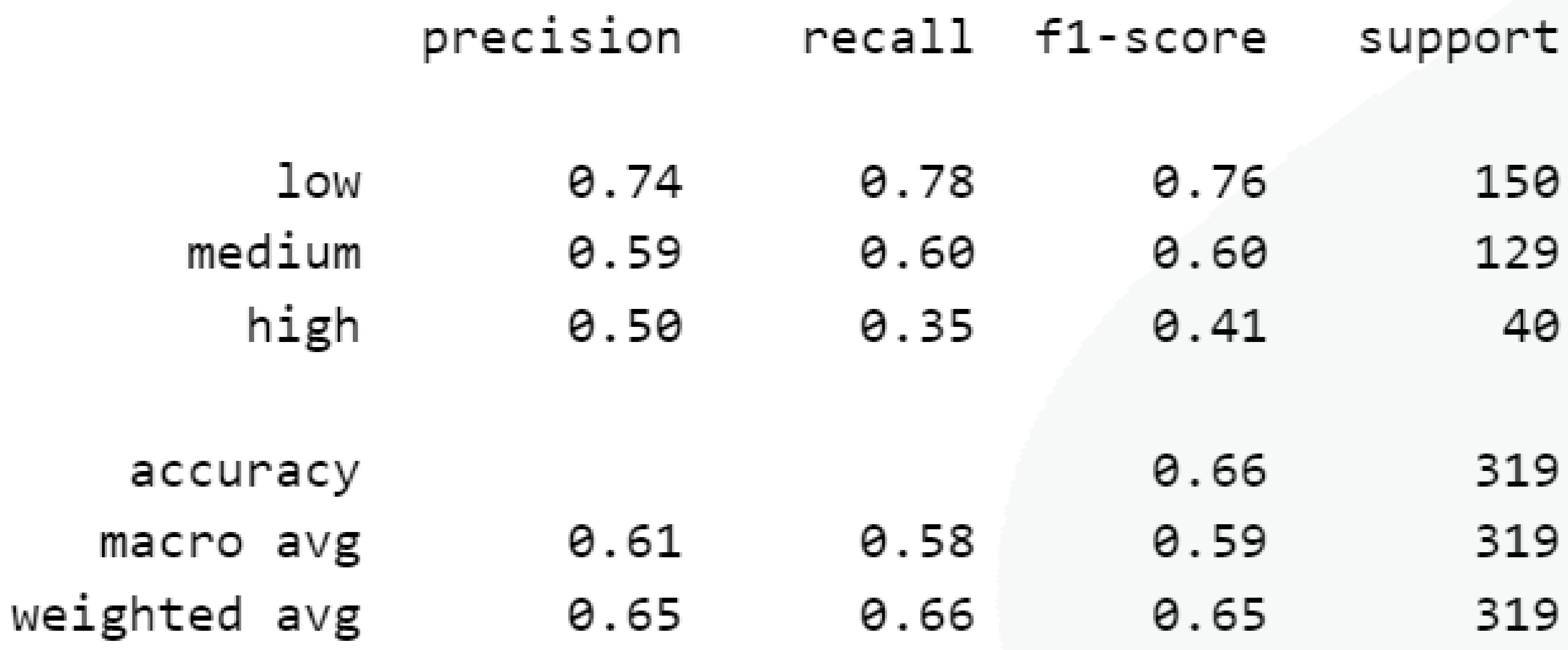
Improved Model



```
Best parameters: {'max_depth': 8, 'max_features': 'auto',
'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50,
'random_state': 50}
Test set accuracy: 0.65517
Confusion matrix:
```



	low	medium	high
low	117	29	4
medium	41	78	10
high	0	26	14

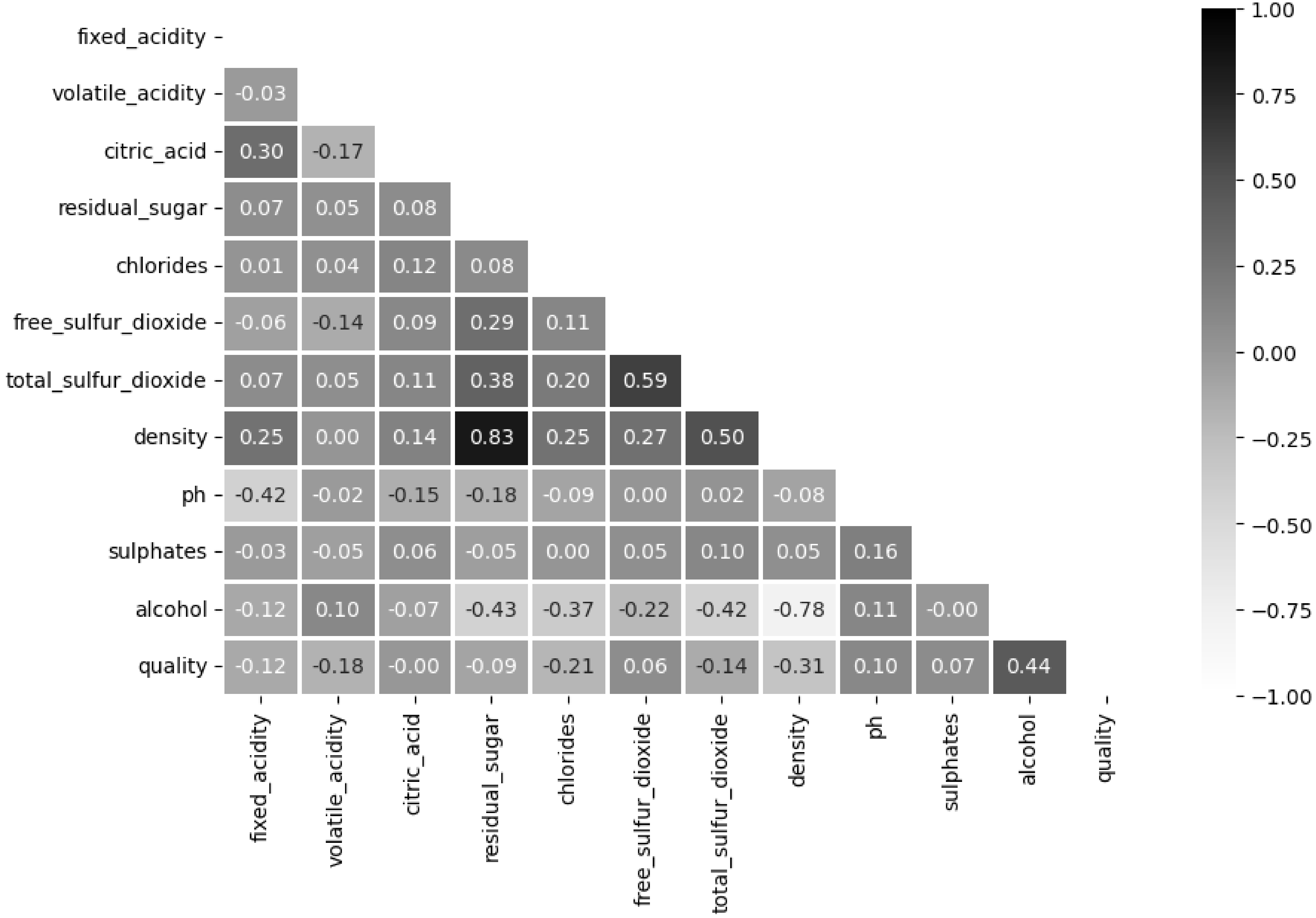


```
              precision    recall  f1-score   support

         low       0.74      0.78      0.76       150
      medium       0.59      0.60      0.60       129
        high       0.50      0.35      0.41        40

    accuracy                           0.66       319
   macro avg       0.61      0.58      0.59       319
weighted avg       0.65      0.66      0.65       319
```

White Wine Quality Prediction

Feature Selection

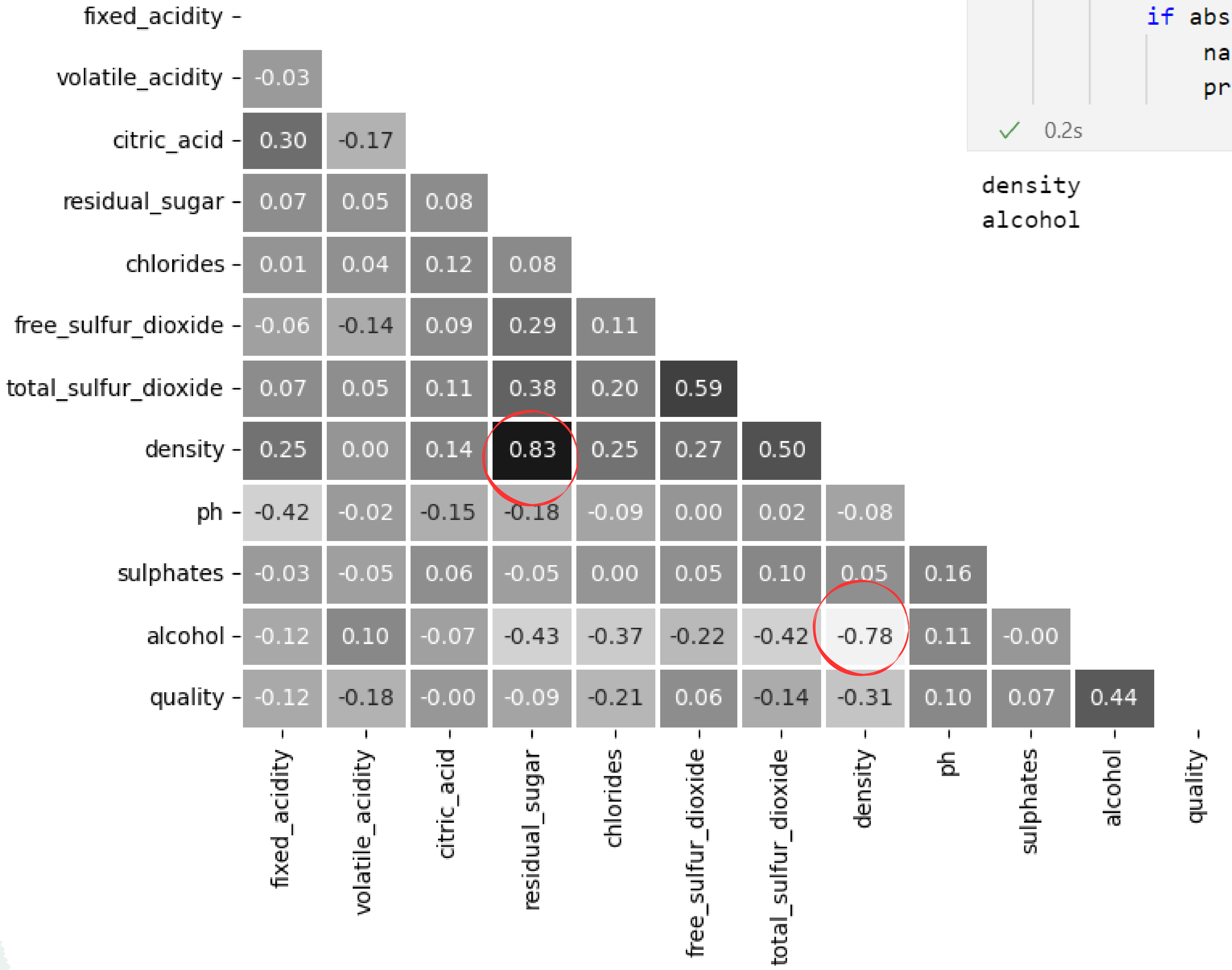
Using Correlations



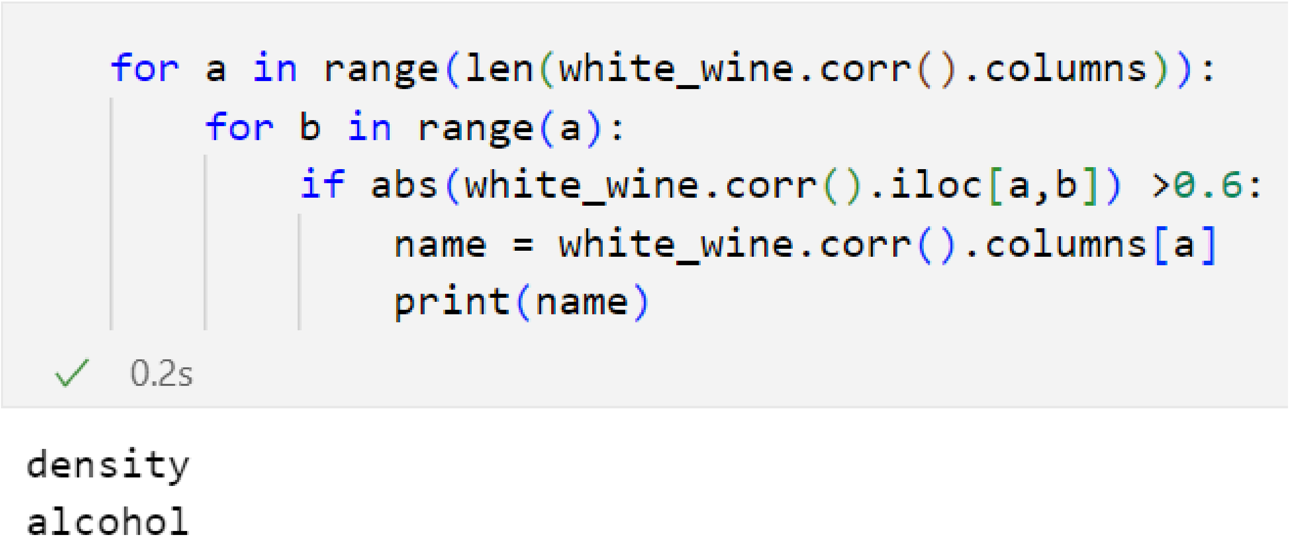
	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	ph	sulphates	alcohol	quality
fixed_acidity												
volatile_acidity	-0.03											
citric_acid	0.30	-0.17										
residual_sugar	0.07	0.05	0.08									
chlorides	0.01	0.04	0.12	0.08								
free_sulfur_dioxide	-0.06	-0.14	0.09	0.29	0.11							
total_sulfur_dioxide	0.07	0.05	0.11	0.38	0.20	0.59						
density	0.25	0.00	0.14	0.83	0.25	0.27	0.50					
ph	-0.42	-0.02	-0.15	-0.18	-0.09	0.00	0.02	-0.08				
sulphates	-0.03	-0.05	0.06	-0.05	0.00	0.05	0.10	0.05	0.16			
alcohol	-0.12	0.10	-0.07	-0.43	-0.37	-0.22	-0.42	-0.78	0.11	-0.00		
quality	-0.12	-0.18	-0.00	-0.09	-0.21	0.06	-0.14	-0.31	0.10	0.07	0.44	

Feature Selection

Using Correlations



	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	ph	sulphates	alcohol	quality
fixed_acidity												
volatile_acidity	-0.03											
citric_acid	0.30	-0.17										
residual_sugar	0.07	0.05	0.08									
chlorides	0.01	0.04	0.12	0.08								
free_sulfur_dioxide	-0.06	-0.14	0.09	0.29	0.11							
total_sulfur_dioxide	0.07	0.05	0.11	0.38	0.20	0.59						
density	0.25	0.00	0.14	0.83	0.25	0.27	0.50					
ph	-0.42	-0.02	-0.15	-0.18	-0.09	0.00	0.02	-0.08				
sulphates	-0.03	-0.05	0.06	-0.05	0.00	0.05	0.10	0.05	0.16			
alcohol	-0.12	0.10	-0.07	-0.43	-0.37	-0.22	-0.42	-0.78	0.11	-0.00		
quality	-0.12	-0.18	-0.00	-0.09	-0.21	0.06	-0.14	-0.31	0.10	0.07	0.44	



```
for a in range(len(white_wine.corr().columns)):
    for b in range(a):
        if abs(white_wine.corr().iloc[a,b]) >0.6:
            name = white_wine.corr().columns[a]
            print(name)
```

```
density
alcohol
```

Accuracy Measure Table (White Wine) 923

Low = 297, Medium = 421, High = 205

Model Name	Mean CV Score	Low	Medium	High	Overall Accuracy
Logistic Regression	0.49	125	352	12	0.52
Decision Tree	0.59	195	272	119	0.63
Random Forest	0.67	216	321	115	0.70
Naive Bayes	0.43	112	183	149	0.48
Support Vector	0.53	168	338	25	0.56
AdaboostClassifier	0.60	198	283	118	0.64
Neural Network	0.55	177	273	70	0.56

Accuracy Measure Table (White Wine) 923

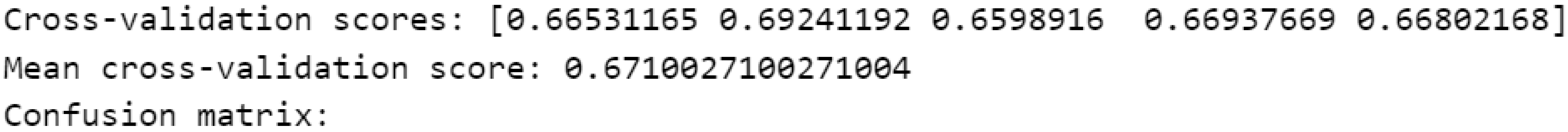
Low = 297, Medium = 421, High = 205

Model Name	Mean CV Score	Low	Medium	High	Overall Accuracy
Logistic Regression	0.49	125	352	12	0.52
Decision Tree	0.59	195	272	119	0.63
Random Forest	0.67	216	321	115	0.70
Naive Bayes	0.43	112	183	149	0.48
Support Vector	0.53	168	338	25	0.56
AdaboostClassifier	0.60	198	283	118	0.64
Neural Network	0.55	177	273	70	0.56

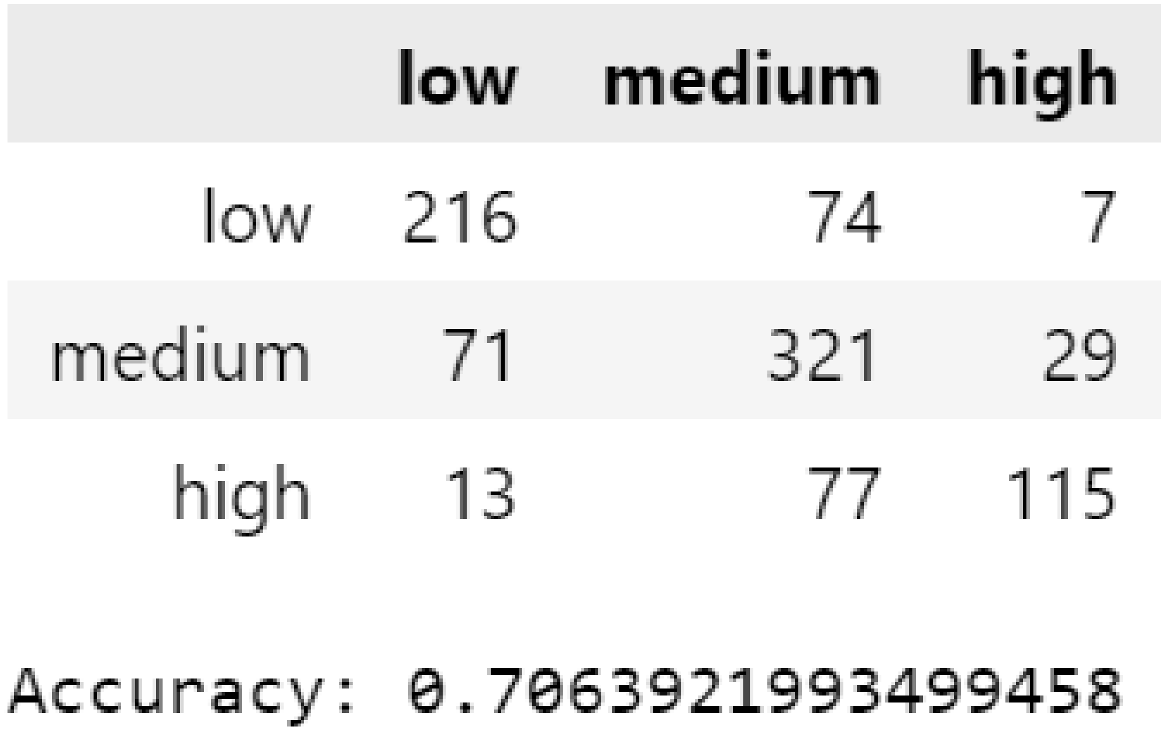
Accuracy Measure Table (White Wine) 923

Model Name	Mean CV Score	Low	Medium	High	Overall Accuracy
Logistic Regression	0.49	125	352	12	0.52
Decision Tree	0.59	195	272	119	0.63
Random Forest	0.67	216	321	115	0.70
Naive Bayes	0.43	112	183	149	0.48
Support Vector	0.53	168	338	25	0.56
AdaboostClassifier	0.60	198	283	118	0.64
Neural Network	0.55	177	273	70	0.56

Random Forest (White Wine)

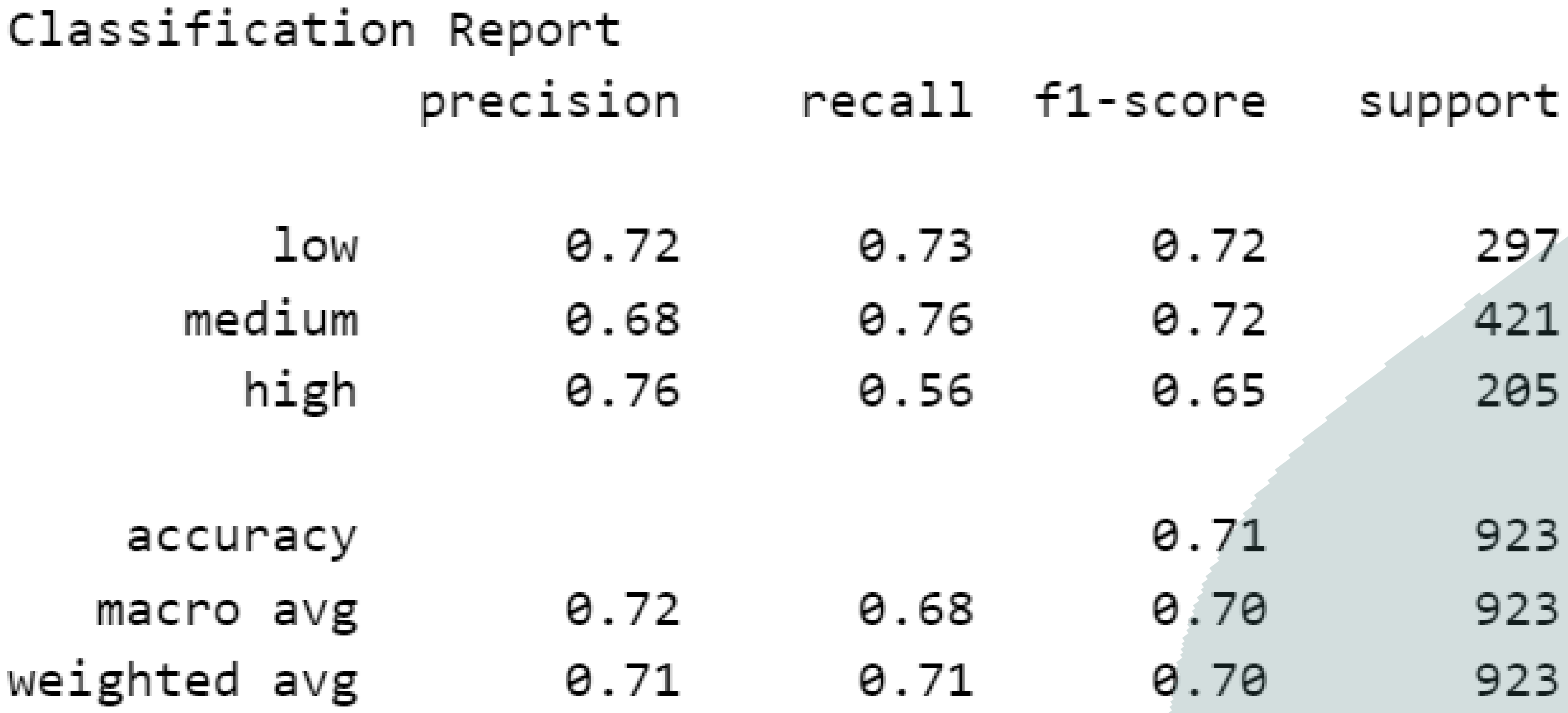


```
Cross-validation scores: [0.66531165 0.69241192 0.6598916  0.66937669 0.66802168]
Mean cross-validation score: 0.6710027100271004
Confusion matrix:
```



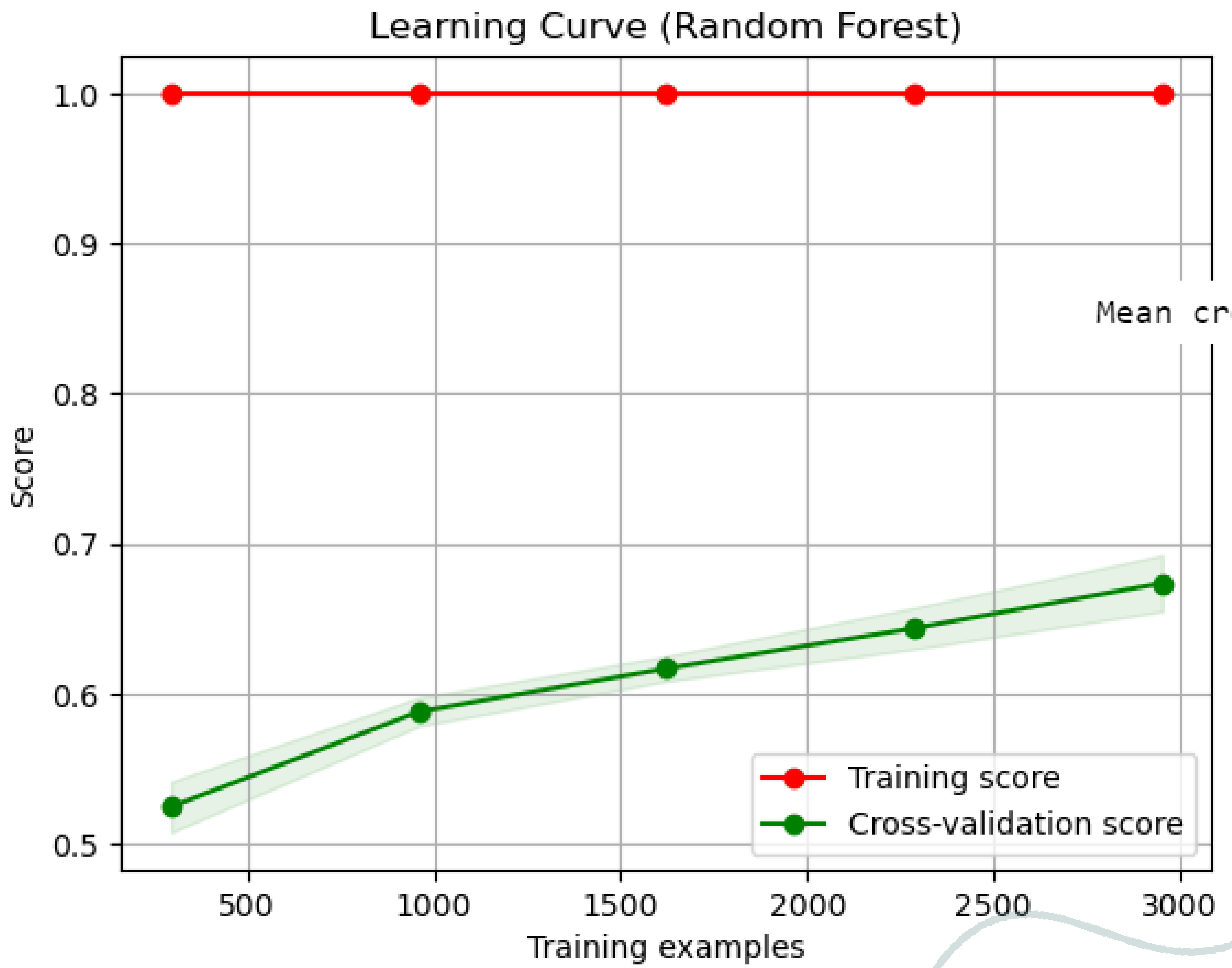
	low	medium	high
low	216	74	7
medium	71	321	29
high	13	77	115

```
Accuracy: 0.7063921993499458
```



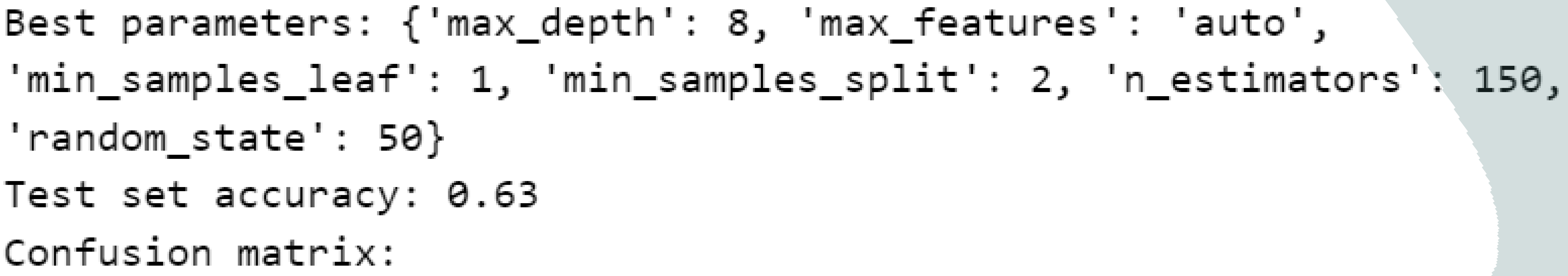
```
Classification Report
              precision    recall  f1-score   support

         low       0.72      0.73      0.72       297
      medium       0.68      0.76      0.72       421
        high       0.76      0.56      0.65       205

    accuracy                           0.71       923
   macro avg       0.72      0.68      0.70       923
weighted avg       0.71      0.71      0.70       923
```

Learning Curve (White Wine)

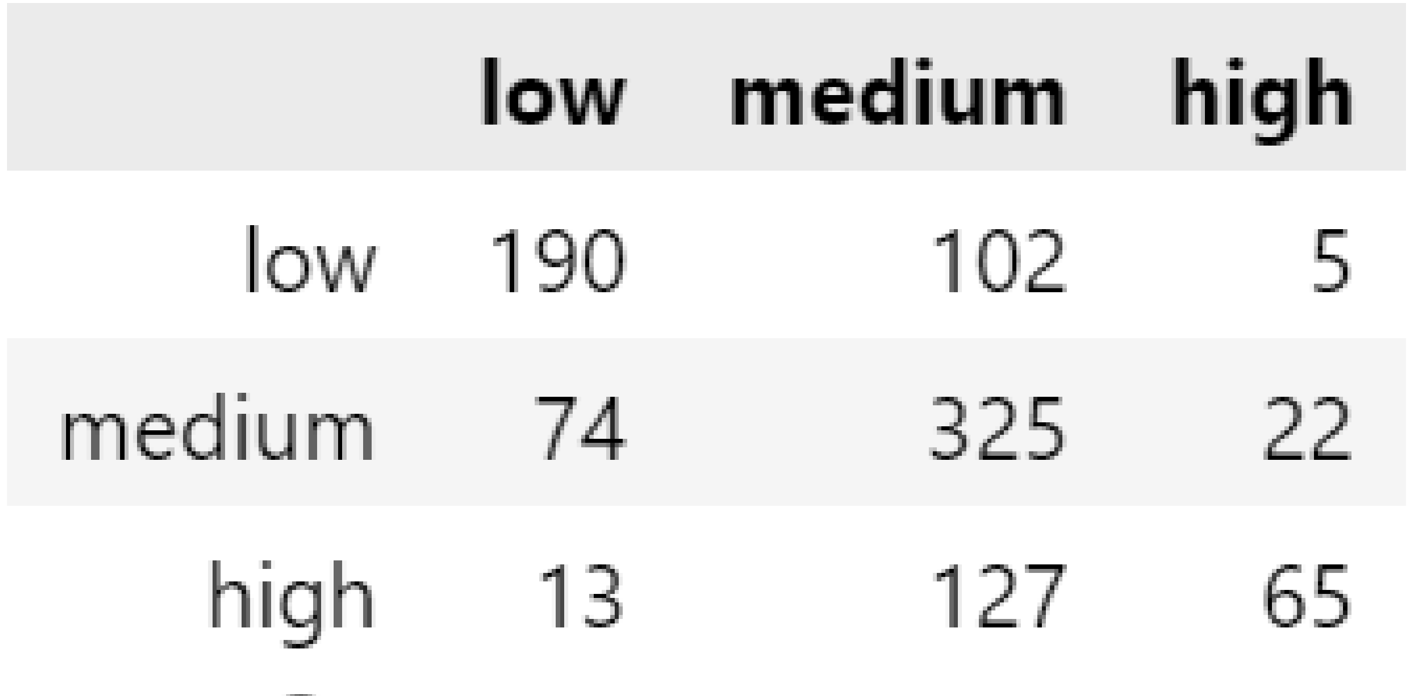


```
Mean cross-validation score: 0.6710027100271004
```

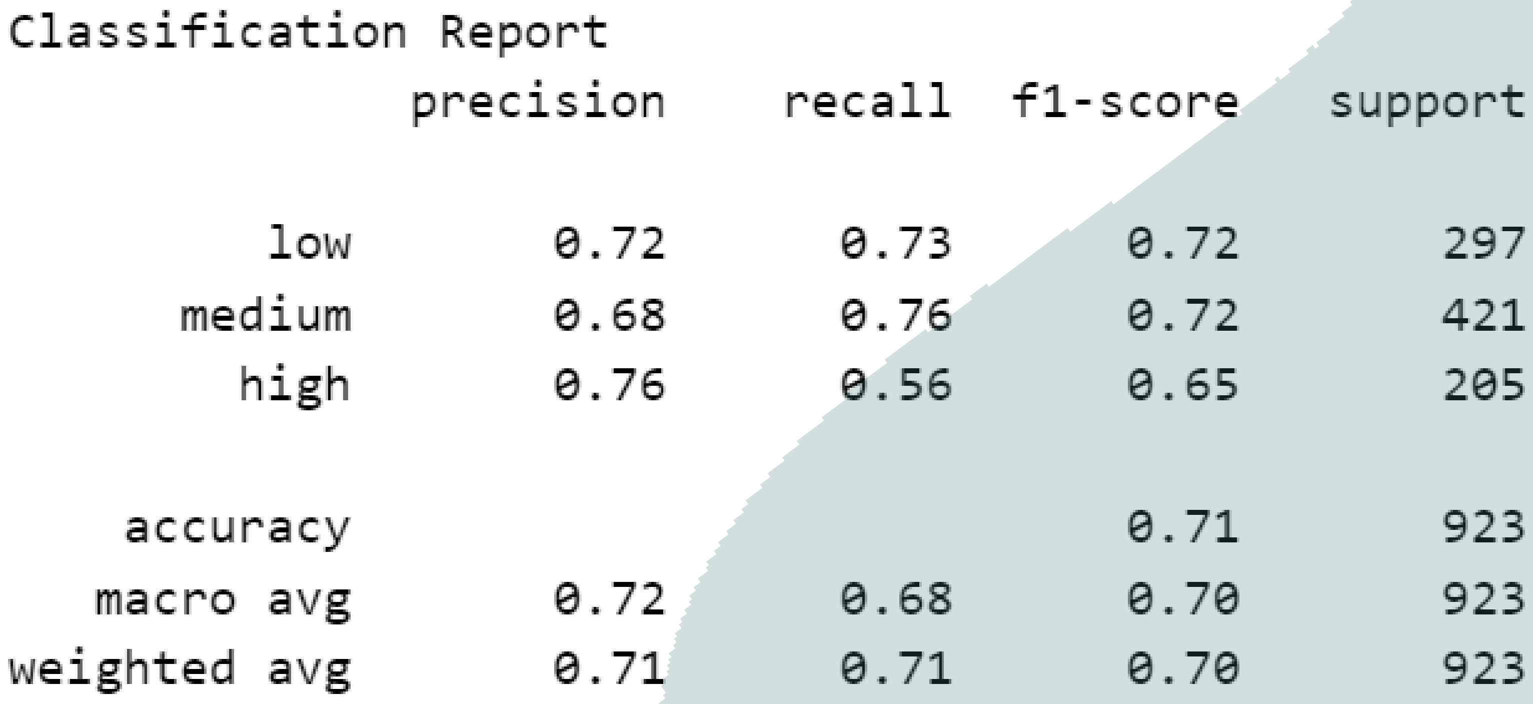
Improved Model



```
Best parameters: {'max_depth': 8, 'max_features': 'auto',
'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 150,
'random_state': 50}
Test set accuracy: 0.63
Confusion matrix:
```



	low	medium	high
low	190	102	5
medium	74	325	22
high	13	127	65



```
Classification Report
              precision    recall  f1-score   support

         low       0.72      0.73      0.72       297
      medium       0.68      0.76      0.72       421
        high       0.76      0.56      0.65       205

    accuracy                           0.71       923
   macro avg       0.72      0.68      0.70       923
weighted avg       0.71      0.71      0.70       923
```

Summary

Cross Validation
(Used Technique)

Random Forest
(Best Model)

Hyperparameter
Tuning
(To Improve)

References

1 https://www.winesofportugal.com/en/portuguese-wines/wine-styles/

2 https://winefolly.com/tips/red-wine-vs-white-wine-the-real-differences/

3 https://rpubs.com/nimit/Report

Thank you!