In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sb
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn import metrics
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
df = pd.read_csv('WineQTdataset.csv')
```

In [3]:
```
df
```

Out[3]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56
...	...	...	...	...	...	...	...	...	...	...
1138	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75
1139	6.8	0.620	0.08	1.9	0.068	28.0	38.0	0.99651	3.42	0.82
1140	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58
1141	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76
1142	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71

1143 rows × 13 columns

In [4]:
```
df.head(5)
```

Out[4]:

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

In [5]:

```
df.tail(5)
```

Out[5]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates
1138	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75
1139	6.8	0.620	0.08	1.9	0.068	28.0	38.0	0.99651	3.42	0.82
1140	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58
1141	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76
1142	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71

In [6]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1143 entries, 0 to 1142
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
```

In [7]:

```
df.isnull()
```

Out[7]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates
0	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...
1138	False	False	False	False	False	False	False	False	False	False
1139	False	False	False	False	False	False	False	False	False	False
1140	False	False	False	False	False	False	False	False	False	False
1141	False	False	False	False	False	False	False	False	False	False
1142	False	False	False	False	False	False	False	False	False	False

1143 rows × 13 columns

In [8]:

```
df.sum()
```

Out[8]:

```
fixed acidity             9499.600000
volatile acidity           607.320000
citric acid                306.740000
residual sugar            2894.250000
chlorides                   99.364000
free sulfur dioxide      17848.500000
total sulfur dioxide     52480.500000
density                   1139.262860
pH                        3784.490000
sulphates                  751.760000
alcohol                  11935.333333
quality                   6466.000000
Id                      920080.000000
dtype: float64
```

In [9]:

```
df.isnull().sum()
```

Out[9]:

```
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
Id                      0
dtype: int64
```

In [10]:

```
df.describe()
```

Out[10]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total d
count	1143.000000	1143.000000	1143.000000	1143.000000	1143.000000	1143.000000	1143.0
mean	8.311111	0.531339	0.268364	2.532152	0.086933	15.615486	45.9
std	1.747595	0.179633	0.196686	1.355917	0.047267	10.250486	32.7
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.0
25%	7.100000	0.392500	0.090000	1.900000	0.070000	7.000000	21.0
50%	7.900000	0.520000	0.250000	2.200000	0.079000	13.000000	37.0
75%	9.100000	0.640000	0.420000	2.600000	0.090000	21.000000	61.0
max	15.900000	1.580000	1.000000	15.500000	0.611000	68.000000	289.0

In [11]:
```
df = pd.get_dummies(df,drop_first=True)
df
```

Out[11]:

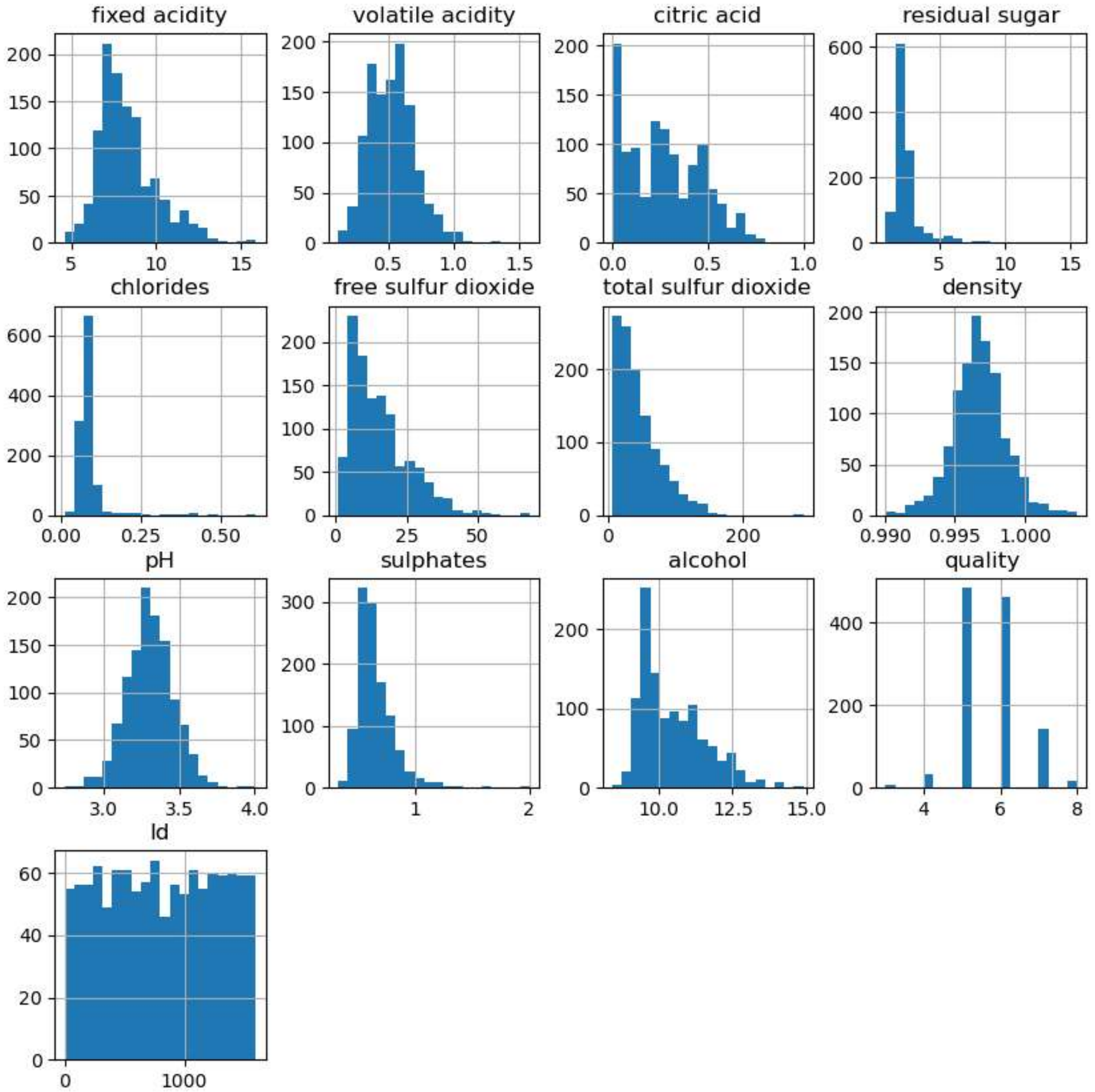
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56
...	...	...	...	...	...	...	...	...	...	...
1138	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75
1139	6.8	0.620	0.08	1.9	0.068	28.0	38.0	0.99651	3.42	0.82
1140	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58
1141	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76
1142	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71

1143 rows × 13 columns

In [13]:
```
for col in df.columns:
 if df[col].isnull().sum() > 0:
    df[col] = df[col].fillna(df[col].mean())
df.isnull().sum().sum()
```

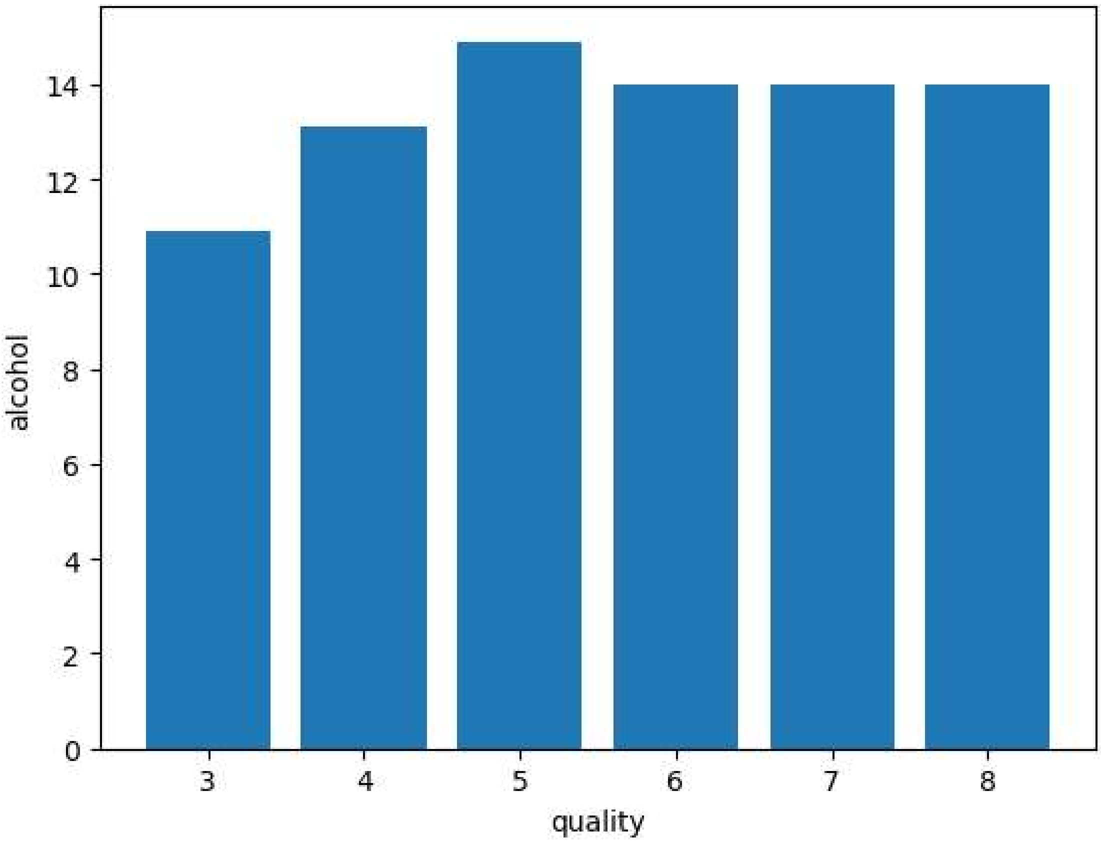
Out[13]:
```
0
```

In [14]:
```
df.hist(bins=20, figsize=(10, 10))
plt.show()
```

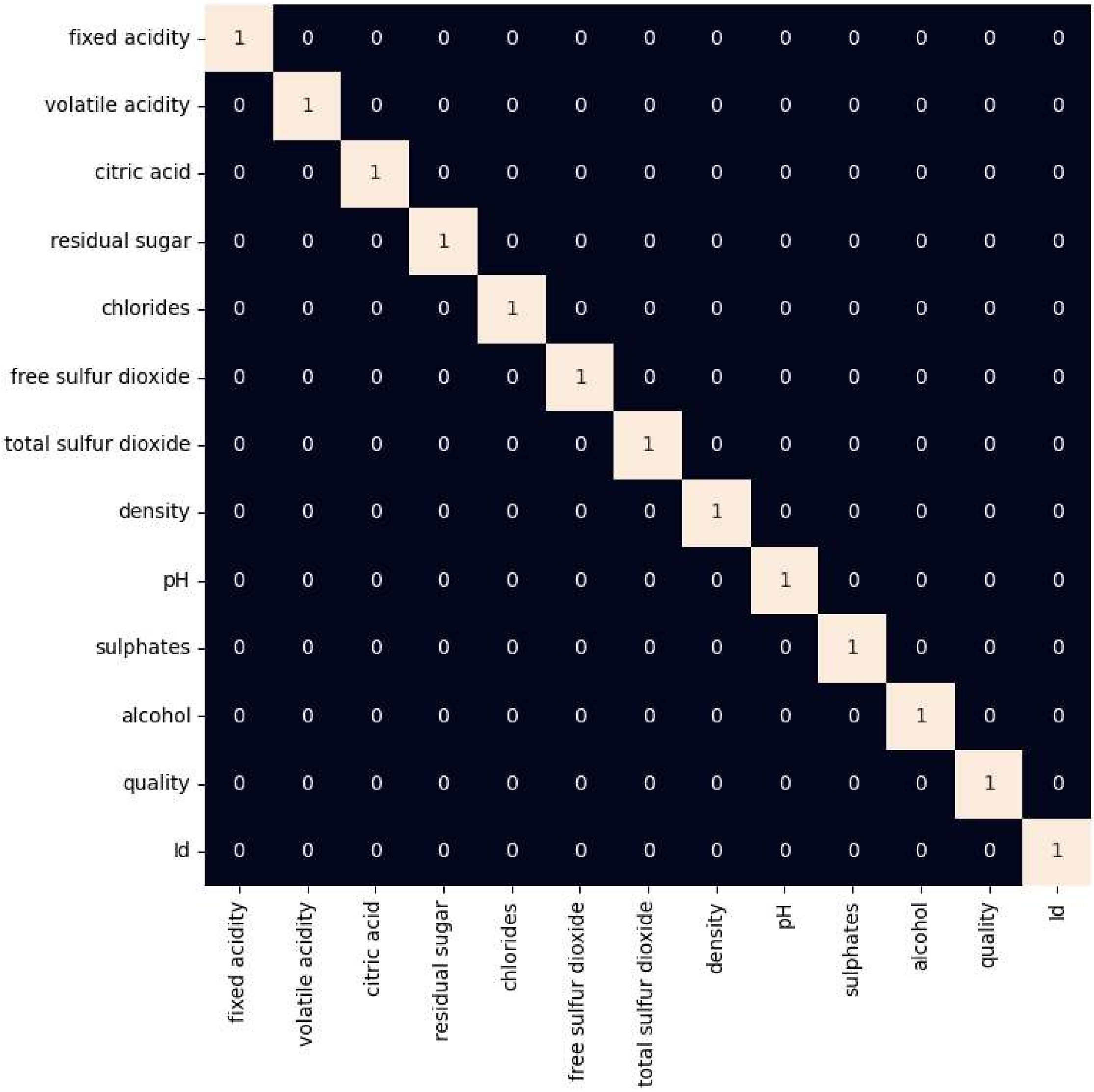


In [15]:

```
plt.bar(df['quality'], df['alcohol'])
plt.xlabel('quality')
plt.ylabel('alcohol')
plt.show()
```



In [18]:
```
plt.figure(figsize=(8, 8))
sb.heatmap(df.corr() > 0.7, annot=True, cbar=False)
plt.show()
```



In [19]:
```
df = df.drop('total sulfur dioxide', axis=1)
```

In [20]:
```
df['best quality'] = [1 if x > 5 else 0 for x in df.quality]
df.replace({'white': 1, 'red': 0}, inplace=True)
```

In [21]:
```
features = df.drop(['quality', 'best quality'], axis=1)
target = df['best quality']
xtrain, xtest, ytrain, ytest = train_test_split(
 features, target, test_size=0.2, random_state=40)
xtrain.shape, xtest.shape
```

Out[21]:
```
((914, 11), (229, 11))
```

In [22]:
```
xtrain
```

Out[22]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	density	pH	sulphates	alcohol
675	9.0	0.36	0.52	2.10	0.111	5.0	0.99568	3.31	0.62	11.3
653	8.6	0.47	0.27	2.30	0.055	14.0	0.99516	3.18	0.80	11.2
845	7.7	0.57	0.21	1.50	0.069	4.0	0.99458	3.16	0.54	9.8
1027	6.9	0.58	0.20	1.75	0.058	8.0	0.99322	3.38	0.49	11.7
1023	10.0	0.38	0.38	1.60	0.169	27.0	0.99914	3.15	0.65	8.5
...	...	...	...	...	...	...	...	...	...	...
626	6.8	0.57	0.00	2.50	0.072	32.0	0.99491	3.43	0.56	11.2
1016	7.6	0.41	0.33	2.50	0.078	6.0	0.99570	3.30	0.58	11.2
165	8.2	1.00	0.09	2.30	0.065	7.0	0.99685	3.32	0.55	9.0
7	7.3	0.65	0.00	1.20	0.065	15.0	0.99460	3.39	0.47	10.0
219	8.4	0.65	0.60	2.10	0.112	12.0	0.99730	3.20	0.52	9.2

914 rows × 11 columns

In [23]:
```
xtest
```

Out[23]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	density	pH	sulphates	alcohol
471	7.2	0.57	0.06	1.6	0.076	9.0	0.99720	3.36	0.70	9.6
192	11.5	0.18	0.51	4.0	0.104	4.0	0.99960	3.28	0.97	10.1
1035	6.5	0.90	0.00	1.6	0.052	9.0	0.99467	3.50	0.63	10.9
476	8.2	0.73	0.21	1.7	0.074	5.0	0.99680	3.20	0.52	9.5
512	8.4	0.56	0.04	2.0	0.082	10.0	0.99760	3.22	0.44	9.6
...	...	...	...	...	...	...	...	...	...	...
281	7.7	0.69	0.05	2.7	0.075	15.0	0.99740	3.26	0.61	9.1
176	7.1	0.60	0.00	1.8	0.074	16.0	0.99720	3.47	0.70	9.9
537	8.3	0.65	0.10	2.9	0.089	17.0	0.99803	3.29	0.55	9.5
1043	7.3	0.48	0.32	2.1	0.062	31.0	0.99728	3.30	0.65	10.0
801	8.5	0.28	0.35	1.7	0.061	6.0	0.99524	3.30	0.74	11.8

229 rows × 11 columns

In [24]:

```
ytrain
```

Out[24]:

```
675     1
653     0
845     1
1027    0
1023    0
       ..
626     1
1016    0
165     1
7       1
219     0
Name: best quality, Length: 914, dtype: int64
```

In [25]:

```
ytest
```

Out[25]:

```
471     1
192     1
1035    1
476     0
512     0
       ..
281     0
176     1
537     0
1043    1
801     1
Name: best quality, Length: 229, dtype: int64
```

In [26]:

```
xtrain.shape
```

Out[26]:

```
(914, 11)
```

In [27]:

```
xtest.shape
```

Out[27]:

```
(229, 11)
```

In [30]:

```
ytrain.shape
```

Out[30]:

```
(914,)
```

In [29]:

```
ytest.shape
```

Out[29]:

```
(229,)
```

In [31]:

```
norm = MinMaxScaler()
xtrain = norm.fit_transform(xtrain)
xtest = norm.transform(xtest)
```

In [46]:

```
models = [LogisticRegression(),SVC(kernel='rbf')]
for i in range(2):
 models[i].fit(xtrain, ytrain)
 print(f'models[i] : ')
 print('Training Accuracy : ', metrics.roc_auc_score(ytrain, models[i].pred
 print('Validation Accuracy : ', metrics.roc_auc_score(
 ytest, models[i].predict(xtest)))
 print()
```

```
models[i] :
Training Accuracy :  0.7546950559364851
Validation Accuracy :  0.7255154639175256

models[i] :
Training Accuracy :  0.7648213641284736
Validation Accuracy :  0.7358247422680412
```

In [36]:

```
from sklearn.metrics import confusion_matrix
import matplotlib.pyplot as plt

y_pred = models[1].predict(xtest)
cm = confusion_matrix(ytest, y_pred)
print(cm)
```

```
[[70 27]
 [33 99]]
```

In [38]:

```
print(metrics.classification_report(ytest,
 models[1].predict(xtest)))
```

```
              precision    recall  f1-score   support

           0       0.68      0.72      0.70        97
           1       0.79      0.75      0.77       132

    accuracy                           0.74       229
   macro avg       0.73      0.74      0.73       229
weighted avg       0.74      0.74      0.74       229
```

In [43]:

```
for a in range(len(df.corr().columns)):
    for b in range(a):
        if abs(df.corr().iloc[a,b]) >0.7:
            name = df.corr().columns[a]
            print(name)
```

```
best quality
```

In [ ]: