

# Final Team Project

Andrew Kim, Luis Perez, Renetta Nelson

October 17, 2022

## Problem Statement

The purpose of this project is to automate the wine selection process in order to increase profit and build on the business's reputation. This will be done by implementing a model that predicts the quality of the wine. The profit margin of restaurants is approximately 70%. This means that over half the profit of these business types come from wine. On the other hand, there are also major expenses that pertain to wine as well. From vendors to sommeliers, there are dozens of additional expenses when it comes to finding and purchasing good quality wine. The profits of the business can no longer support the expenses of the wine selection process. Within a few months, the expenses will exceed the profits of the business and the business will have to close down. The automation of the wine selection process will reduce the expenses by approximately 25%, allowing the business to build its finances and stay in business.

In [129...]

```
#Import Libraries

import pandas as pd
import numpy as np
import random
from sklearn import preprocessing
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, confusion_matrix, roc_auc_score, roc_curve
from sklearn.neighbors import NearestNeighbors, KNeighborsClassifier
from sklearn.linear_model import LogisticRegression, LogisticRegressionCV
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier

from dmba import classificationSummary, gainsChart, liftChart
import scikitplot as skplt
import matplotlib.pyplot as plt
from dmba.metric import AIC_score
import statsmodels.api as sm
import warnings
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.neural_network import MLPClassifier
from dmba import classificationSummary
from sklearn.metrics import r2_score, plot_confusion_matrix, multilabel_confusion_matri
```

```

import seaborn as sns

warnings.filterwarnings("ignore")

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import datasets, linear_model
import matplotlib.pyplot as plt
import dmba
from dmba import regressionSummary
from dmba import adjusted_r2_score, AIC_score, BIC_score
from sklearn.cluster import KMeans

```

In [24]:

```

#Load dataset and put into a data frame

redwine_data = pd.read_csv("winequality-red.csv")

redwine_df = pd.DataFrame(redwine_data)

#Display first five rows of dataframe to confirm

redwine_df.head()

```

Out[24]:

	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	
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	

◀ ▶

## Data Preprocessing

Data Preprocessing Explanation for this section -> Can do a quick summary here, talking about each of the processes that are being done for data preprocessing

In [7]:

```

# Check type of variables
redwine_df.info()

```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   fixed acidity    1599 non-null   float64 
 1   volatile acidity 1599 non-null   float64 
 2   citric acid     1599 non-null   float64 

```

```
3    residual sugar          1599 non-null   float64
4    chlorides                1599 non-null   float64
5    free sulfur dioxide     1599 non-null   float64
6    total sulfur dioxide    1599 non-null   float64
7    density                  1599 non-null   float64
8    pH                       1599 non-null   float64
9    sulphates                1599 non-null   float64
10   alcohol                  1599 non-null   float64
11   quality                  1599 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

In [8]: `redwine_df.describe()`

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density
<b>count</b>	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
<b>mean</b>	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.997050
<b>std</b>	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.852589
<b>min</b>	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.94
<b>25%</b>	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.96
<b>50%</b>	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.98
<b>75%</b>	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	1.00
<b>max</b>	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.00

In [9]: `# Check for data size  
redwine_df.shape`

Out[9]: `(1599, 12)`

In [10]: `redwine_df['quality'].unique()`

Out[10]: `array([5, 6, 7, 4, 8, 3], dtype=int64)`

In [11]: `redwine_df.isna().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
```

```
In [12]: redwine_df.dtypes
```

```
fixed acidity      float64  
volatile acidity   float64  
citric acid        float64  
residual sugar     float64  
chlorides          float64  
free sulfur dioxide float64  
total sulfur dioxide float64  
density            float64  
pH                 float64  
sulphates          float64  
alcohol             float64  
quality             int64  
dtype: object
```

```
In [13]: # Removing Outliers
```

```
#based on the boxplots total sulfur dioxide has many outliers  
d1= redwine_df['total sulfur dioxide']  
mean = np.mean(redwine_df['total sulfur dioxide'])  
std = np.std(redwine_df['total sulfur dioxide'])  
print('mean of the dataset is', mean)  
print('std. deviation is', std)
```

```
#z method  
#total sulfur dioxide  
out=[]  
def Zscore_outlier(df):  
    m = np.mean(df)  
    sd = np.std(df)  
    for i in df:  
        z = (i-m)/sd  
        if np.abs(z) > 3:  
            out.append(i)  
    print("Outliers:",out)  
Zscore_outlier(redwine_df['total sulfur dioxide'])
```

```
#z method  
#free sulfur dioxide  
out=[]  
def Zscore_outlier(df):  
    m = np.mean(df)  
    sd = np.std(df)  
    for i in df:  
        z = (i-m)/sd  
        if np.abs(z) > 3:  
            out.append(i)  
    print("Outliers:",out)  
Zscore_outlier(redwine_df['free sulfur dioxide'])
```

```
mean of the dataset is 46.46779237023139  
std. deviation is 32.88503665178367
```

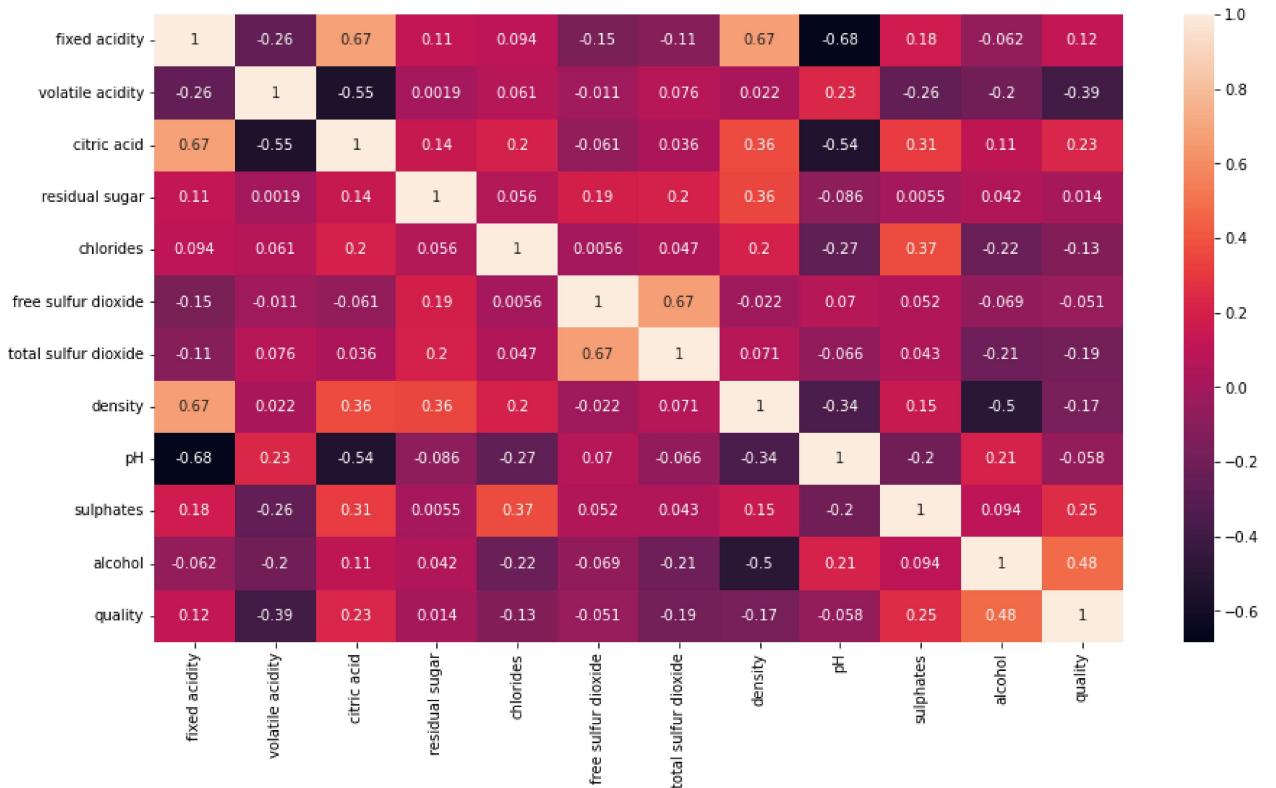
```
Outliers: [148.0, 153.0, 165.0, 151.0, 149.0, 147.0, 148.0, 155.0, 151.0, 152.0, 278.0, 289.0, 160.0, 147.0, 147.0]
```

```
Outliers: [52.0, 51.0, 50.0, 68.0, 68.0, 54.0, 53.0, 52.0, 51.0, 57.0, 50.0, 48.0, 48.0, 72.0, 51.0, 51.0, 52.0, 55.0, 55.0, 48.0, 48.0, 66.0]
```

```
In [14]: plt.figure(figsize = (15, 8))
```

```
sns.heatmap(redwine_df.corr(), annot = True)
```

```
Out[14]: <AxesSubplot:>
```

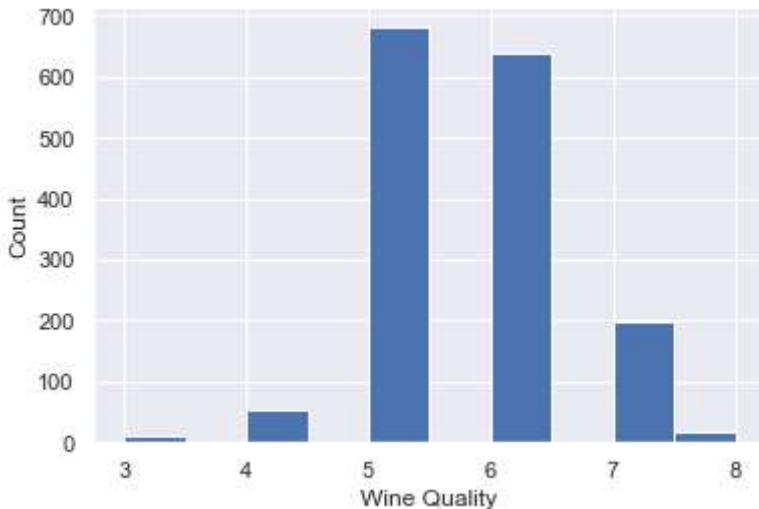


## Explanatory Data Analysis (EDA)

### Exploratory Data Analysis recap

```
In [16]: #Create histogram on 'quality' variable
sns.set()
redwine_df.quality.hist()
plt.xlabel('Wine Quality')
plt.ylabel('Count')
```

```
Out[16]: Text(0, 0.5, 'Count')
```



Explanation:

In [50]:

```
# Analyze the relationships between the predictors and the target variable ('quality').  
  
fig, axes = plt.subplots(4, 3, figsize = (20,20), sharey = True)  
  
sns.scatterplot(ax = axes[0,0], data = redwine_df, y = "quality", x = "alcohol", color = "blue")  
axes[0,0].set_title("Relationship Between Alcohol and Quality")  
  
sns.scatterplot(ax = axes[0, 1], data = redwine_df, y = "quality", x = "pH", color = "red")  
axes[0,1].set_title("Relationship Between pH and Quality")  
  
sns.scatterplot(ax = axes[0, 2], data = redwine_df, y = "quality", x = "sulphates", color = "green")  
axes[0,2].set_title("Relationship Between Sulphates and Quality")  
  
sns.scatterplot(ax = axes[1,0], data = redwine_df, y = "quality", x = "density", color = "purple")  
axes[1,0].set_title("Relationship Between Density and Quality")  
  
sns.scatterplot(ax = axes[1,1], data = redwine_df, y = "quality", x = "total sulfur dioxide", color = "orange")  
axes[1,1].set_title("Relationship Between Total Sulfur Dioxide and Quality")  
  
sns.scatterplot(ax = axes[1,2], data = redwine_df, y = "quality", x = "free sulfur dioxide", color = "brown")  
axes[1,2].set_title("Relationship Between Free Sulfur Dioxide and Quality")  
  
sns.scatterplot(ax = axes[2,0], data = redwine_df, y = "quality", x = "chlorides", color = "pink")  
axes[2,0].set_title("Relationship Between Chlorides and Quality")  
  
sns.scatterplot(ax = axes[2,1], data = redwine_df, y = "quality", x = "residual sugar", color = "teal")  
axes[2,1].set_title("Relationship Between Residual Sugar and Quality")  
  
sns.scatterplot(ax = axes[2,2], data = redwine_df, y = "quality", x = "citric acid", color = "yellow")  
axes[2,2].set_title("Relationship Between Citric Acid and Quality")  
  
sns.scatterplot(ax = axes[3,0], data = redwine_df, y = "quality", x = "volatile acidity", color = "lightblue")  
axes[3,0].set_title("Relationship Between Volatile Acidity and Quality")  
  
sns.scatterplot(ax = axes[3,1], data = redwine_df, y = "quality", x = "fixed acidity", color = "lightgreen")  
axes[3,1].set_title("Relationship Between Fixed Acidity and Quality")
```

Out[50]: Text(0.5, 1.0, 'Relationship Between Fixed Acidity and Quality')

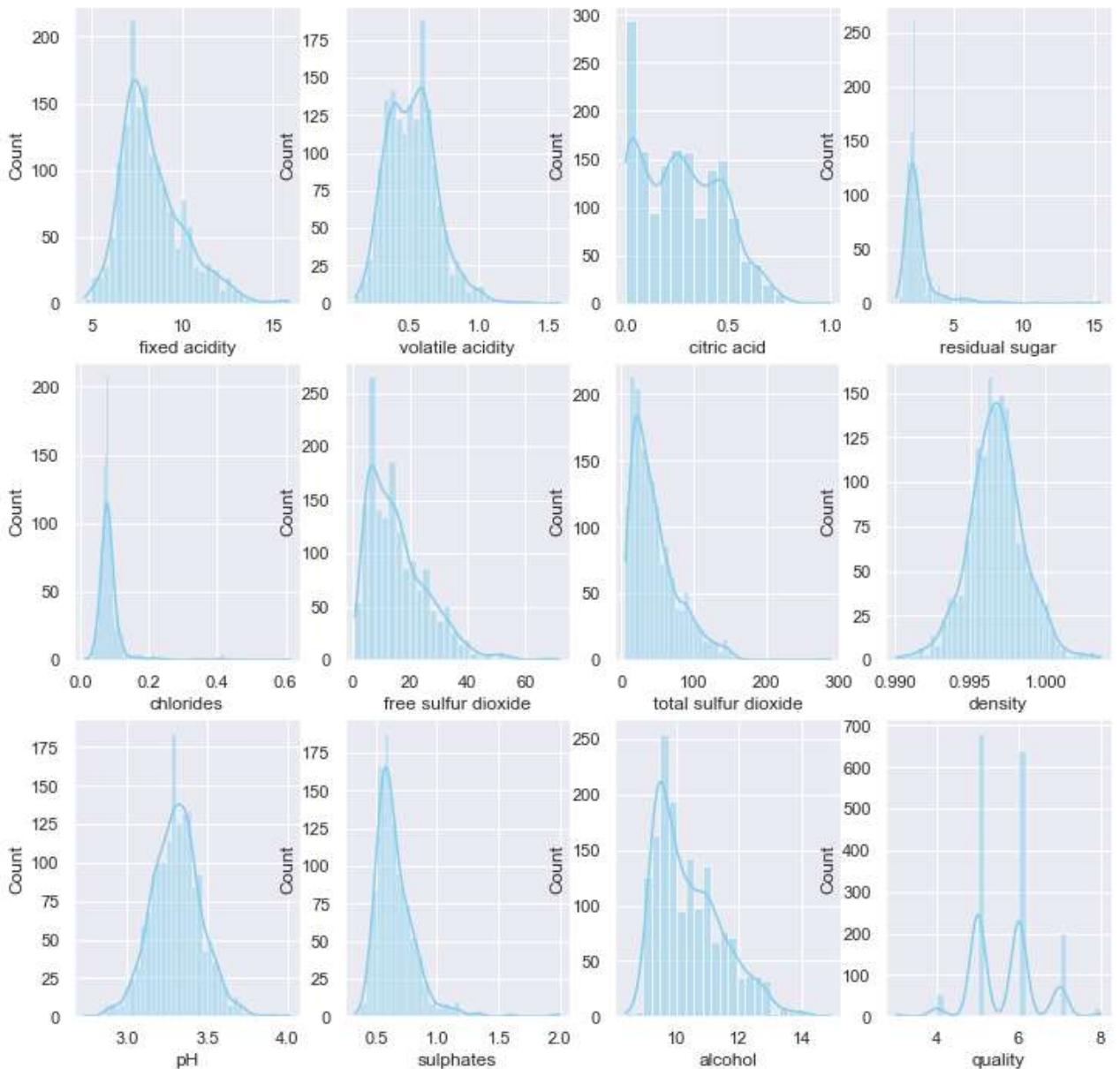


Explanation:

In [51]:

```
# Histogram

fig, axs = plt.subplots(3, 4, figsize=(12, 12))
columns = redwine_df.columns[:12]
k=0
sns.set(font_scale=1)
for i in range(3):
    for j in range(4):
        sns.histplot(data=redwine_df, x=columns[k], kde=True, color="skyblue", ax=axs[i][j])
        k+=1
```



Explanation:

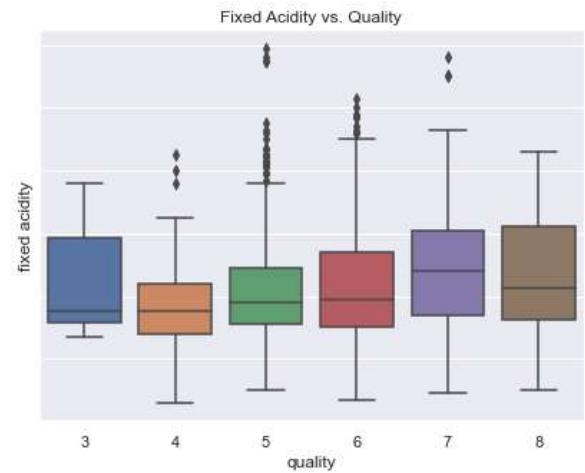
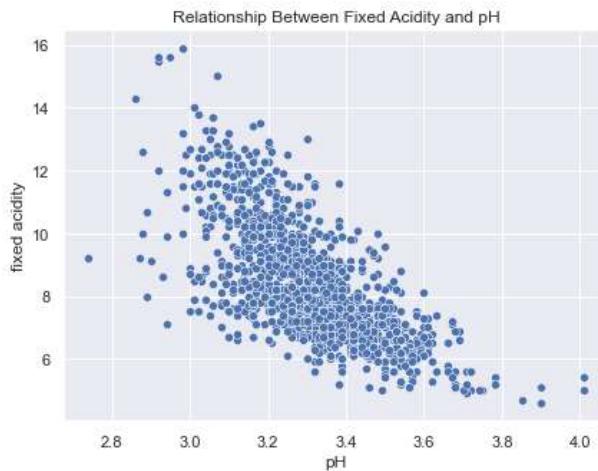
In [54]:

```
fig, axes = plt.subplots(1, 2, figsize = (15, 5), sharey = True)

sns.scatterplot(ax = axes[0], data = redwine_df, y = "fixed acidity", x = "pH")
axes[0].set_title("Relationship Between Fixed Acidity and pH")

sns.boxplot(ax = axes[1], data = redwine_df, y = "fixed acidity", x = "quality")
axes[1].set_title("Fixed Acidity vs. Quality")
```

Out[54]:



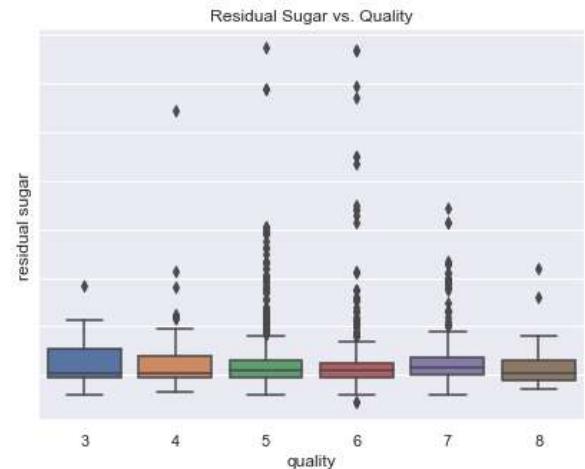
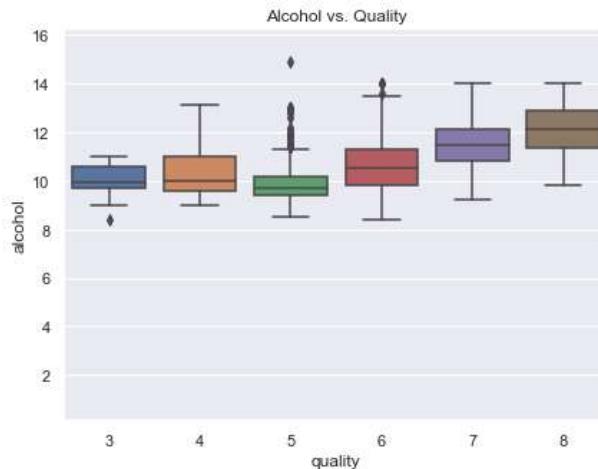
Explanation:

```
In [57]: fig, axes = plt.subplots(1, 2, figsize = (15, 5), sharey = True)

sns.boxplot(ax = axes[0], data = redwine_df, y = "alcohol", x = "quality")
axes[0].set_title("Alcohol vs. Quality")

sns.boxplot(ax = axes[1], data = redwine_df, y = "residual sugar", x = "quality")
axes[1].set_title("Residual Sugar vs. Quality")
```

Out[57]: Text(0.5, 1.0, 'Residual Sugar vs. Quality')



Explanation

## Data Splitting

```
In [58]: y = redwine_df['quality'].to_numpy()
x = redwine_df.drop(columns=['quality'])
```

```
In [59]: #Standardize the dataset
scaler = preprocessing.StandardScaler()
x_norm = scaler.fit_transform(x * 1.0)
```

```
In [60]: #Split the full dataframe into 60/40.
train_x, test_x, train_y, test_y = train_test_split(x, y, test_size=0.4, random_state=1
train_x.shape, test_x.shape

Out[60]: ((959, 11), (640, 11))
```

# Data Modeling

## LDA

```
In [ ]:
```

## Gradient Boosting

```
In [ ]:
```

## Logistic Regression

```
In [ ]:
```

## Random Forest

```
In [ ]:
```

## Decision Trees

```
In [149...]: decisiontree = DecisionTreeClassifier().fit(train_x, train_y)

y_predict = decisiontree.predict(test_x)

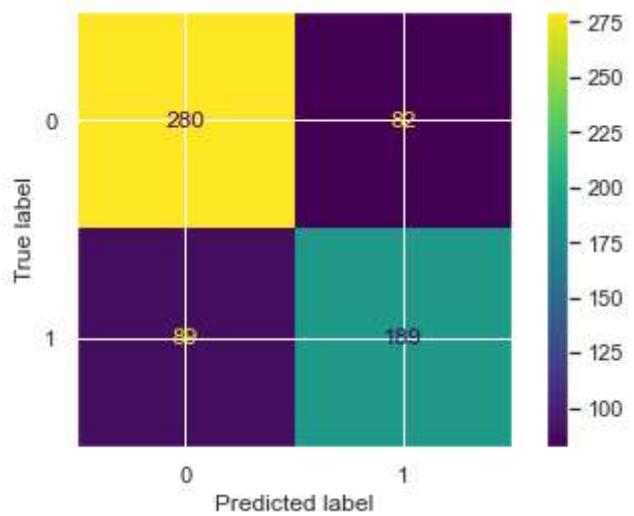
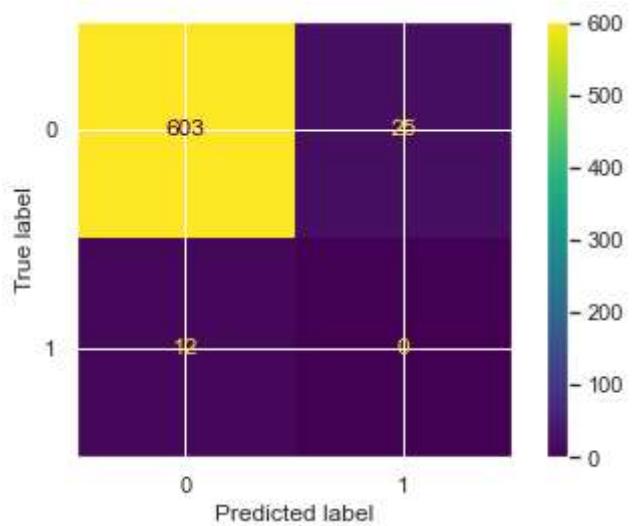
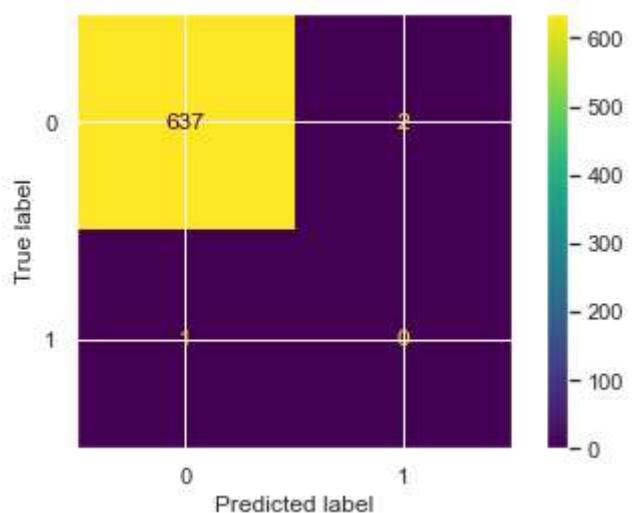
#plot_confusion_matrix(decisiontree, test_x, test_y)

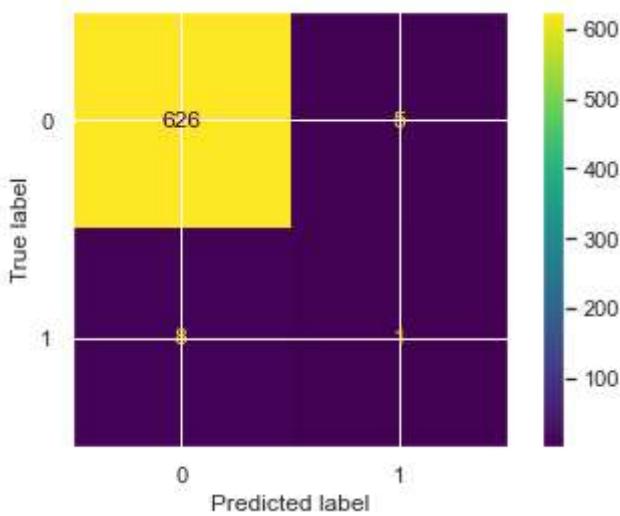
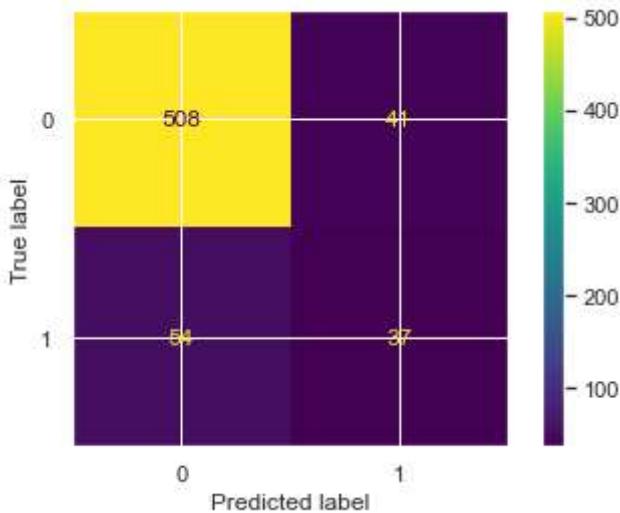
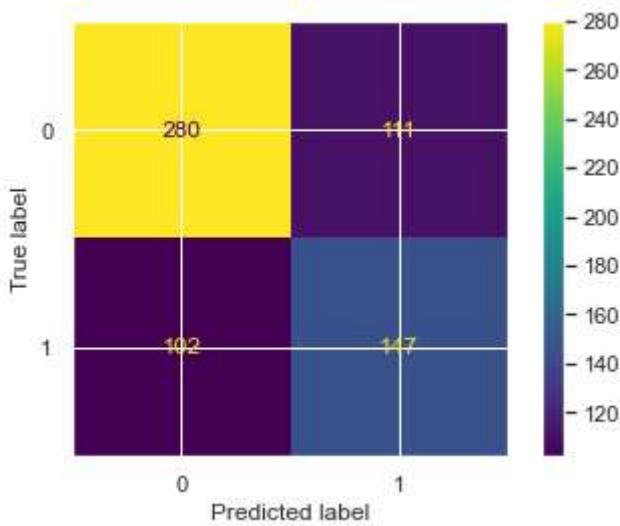
decision_predict = decisiontree.predict(test_x)
#fpr, tpr, _ = roc_curve(test_y, decision_predict)
#aucscore= round(roc_auc_score(test_y, decision_predict), 2)

#plt.plot(fpr, tpr, label = "Decision Tree, AUC Score:" + str(aucscore))

cmatrix = multilabel_confusion_matrix(decision_predict, test_y)
num = 0

for x in range(6):
    cmat = ConfusionMatrixDisplay(cmatrix[num]).plot()
    num = num + 1
```





## K-Means Clustering

In [148]:

```
#normalize
norm_redwine = redwine_df.apply(preprocessing.scale, axis = 0)

#kmeans clustering
```

```
inertia = []
for k in range(1, 7):
    kclustering = KMeans(n_clusters = k, random_state = 0).fit(norm_redwine)
    kmeans_predict = kclustering.fit_predict(redwine_df)

cluster1 = redwine_df[kmeans_predict == 0]
cluster2 = redwine_df[kmeans_predict == 8]

#print(cluster1[:, 12])

#plt.scatter(true_data, redwine_df)

ax.legend().set_visible(False)
plt.show()
```