EDS Assignment 6

By:-
Suryansh Ambekar (202)
Atharva Bramhankar (211)
Shrushty Dhamange (215)

Code:- Linear Regression

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.linear_model import LinearRegression

df1=pd.read_csv("/content/winequalityN.csv")
data = df1.dropna()
print(data)



# Extract the columns for linear regression
X = data['fixed acidity'].values.reshape(-1, 1)  # Input feature
y = data['quality'].values  # Target variable

# Create and fit the linear regression model
model = LinearRegression()
model.fit(X, y)

# Predict the target variable
y_pred = model.predict(X)

# Plot the data points and the regression line
plt.scatter(X, y, color='blue', label='Actual')
plt.plot(X, y_pred, color='red', label='Regression Line')
plt.xlabel('pH')
plt.ylabel('alcohol')
plt.legend()
plt.show()
```

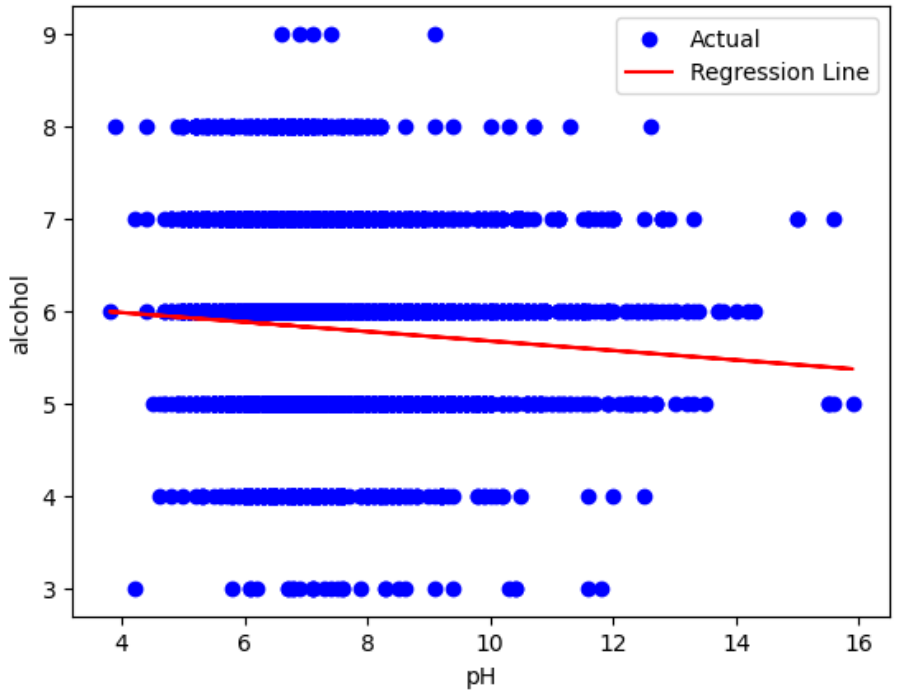
Output:-

	type	fixed acidity	volatile acidity	citric acid	residual sugar	\
0	white	7.0	0.270	0.36	20.7	
1	white	6.3	0.300	0.34	1.6	
2	white	8.1	0.280	0.40	6.9	
3	white	7.2	0.230	0.32	8.5	
4	white	7.2	0.230	0.32	8.5	
...	...	...	...	...	...	
6491	red	6.8	0.620	0.08	1.9	
6492	red	6.2	0.600	0.08	2.0	
6494	red	6.3	0.510	0.13	2.3	
6495	red	5.9	0.645	0.12	2.0	
6496	red	6.0	0.310	0.47	3.6	

	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	\
0	0.045	45.0	170.0	1.00100	3.00	
1	0.049	14.0	132.0	0.99400	3.30	
2	0.050	30.0	97.0	0.99510	3.26	
3	0.058	47.0	186.0	0.99560	3.19	
4	0.058	47.0	186.0	0.99560	3.19	
...	...	...	...	...	...	
6491	0.068	28.0	38.0	0.99651	3.42	
6492	0.090	32.0	44.0	0.99490	3.45	
6494	0.076	29.0	40.0	0.99574	3.42	
6495	0.075	32.0	44.0	0.99547	3.57	
6496	0.067	18.0	42.0	0.99549	3.39	

	sulphates	alcohol	quality	reviews
0	0.45	8.8	6	BAD
1	0.49	9.5	6	VERY BAD
2	0.44	10.1	6	GOOD
3	0.40	9.9	6	GOOD
4	0.40	9.9	6	AVERAGE
...	...	...	...	...
6491	0.82	9.5	6	EXCELLENT
6492	0.58	10.5	5	BAD
6494	0.75	11.0	6	GOOD
6495	0.71	10.2	5	VERY GOOD
6496	0.66	11.0	6	AVERAGE

[6463 rows x 14 columns]



Code:- KNN

```
# Drop the missing values
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.linear_model import LinearRegression

df1=pd.read_csv("/content/winequalityN.csv")
df = df1.dropna()
print(data)

X=df['pH']
df=df.dropna()
Y=df['fixed acidity']
X=np.array(df['pH']).reshape(-1,1)
Y=np.array(df['fixed acidity']).reshape(-1,1)
X_train, X_test,Y_train, Y_test = train_test_split(X,Y,test_size=0.30)
from sklearn.metrics import classification_report,\
confusion_matrix
knn = KNeighborsClassifier(n_neighbors=1)
knn.fit(X_train, y_train)
pred = knn.predict(X_test)
# Predictions and Evaluations
# Let's evaluate our KNN model !
print(confusion_matrix(y_test, pred))
```

Output:-

[[0 0 8 0 0 20 7 7 15 5 11 12 6 0 8 0] [0 0 7 0 0 19 6 6 11 6 15 9 9 0 9 0] [0 0 8 0 0 18 4 11 16 4 10 15 11 0 17 0] [0 0 8 0 0 19 11 3 11 10 11 15 6 0 10 0] [0 0 9 0 0 14 3 9 6 9 9 7 4 0 7 0] [0 0 6 0 0 21 3 4 12 7 13 8 1 0 9 0] [0 0 4 0 0 14 5 9 14 3 17 5 5 0 16 0] [0 0 5 0 0 19 7 4 20 9 17 9 7 0 14 0] [0 0 7 0 0 12 4 4 17 10 15 11 8 0 18 0] [0 0 5 0 0 19 7 8 9 9 20 12 7 0 11 0] [0 0 5 0 0 21 3 4 19 3 12 8 9 0 11 0] [0 0 9 0 0 20 6 10 4 4 13 15 11 0 10 0] [0 0 7 0 0 17 8 2 14 5 11 17 4 0 11 0] [0 0 7 0 0 15 6 2 10 5 10 14 5 0 10 0] [0 0 12 0 0 25 6 7 12 3 10 10 9 0 13 0] [0 0 5 0 0 16 4 6 11 5 12 7 8 0 8 0]] precision recall f1-score support 35 0.00 0.00 0.00 99 36 0.00 0.00 0.00 97 37 0.07 0.07 0.07 114 38 0.00 0.00 0.00 104 39 0.00 0.00 0.00 77 40 0.07 0.25 0.11 84 41 0.06 0.05 0.05 92 42 0.04 0.04 0.04 111 43 0.08 0.16 0.11 106 44 0.09 0.08 0.09 107 45 0.06 0.13 0.08 95 46 0.09 0.15 0.11 102 47 0.04 0.04 0.04 96 48 0.00 0.00 0.00 84 49 0.07 0.12 0.09 107 50 0.00 0.00 0.00 82 accuracy 0.07 1557 macro avg 0.04 0.07 0.05 1557 weighted avg 0.04 0.07 0.05 1557

Code:-KMeans Clustering

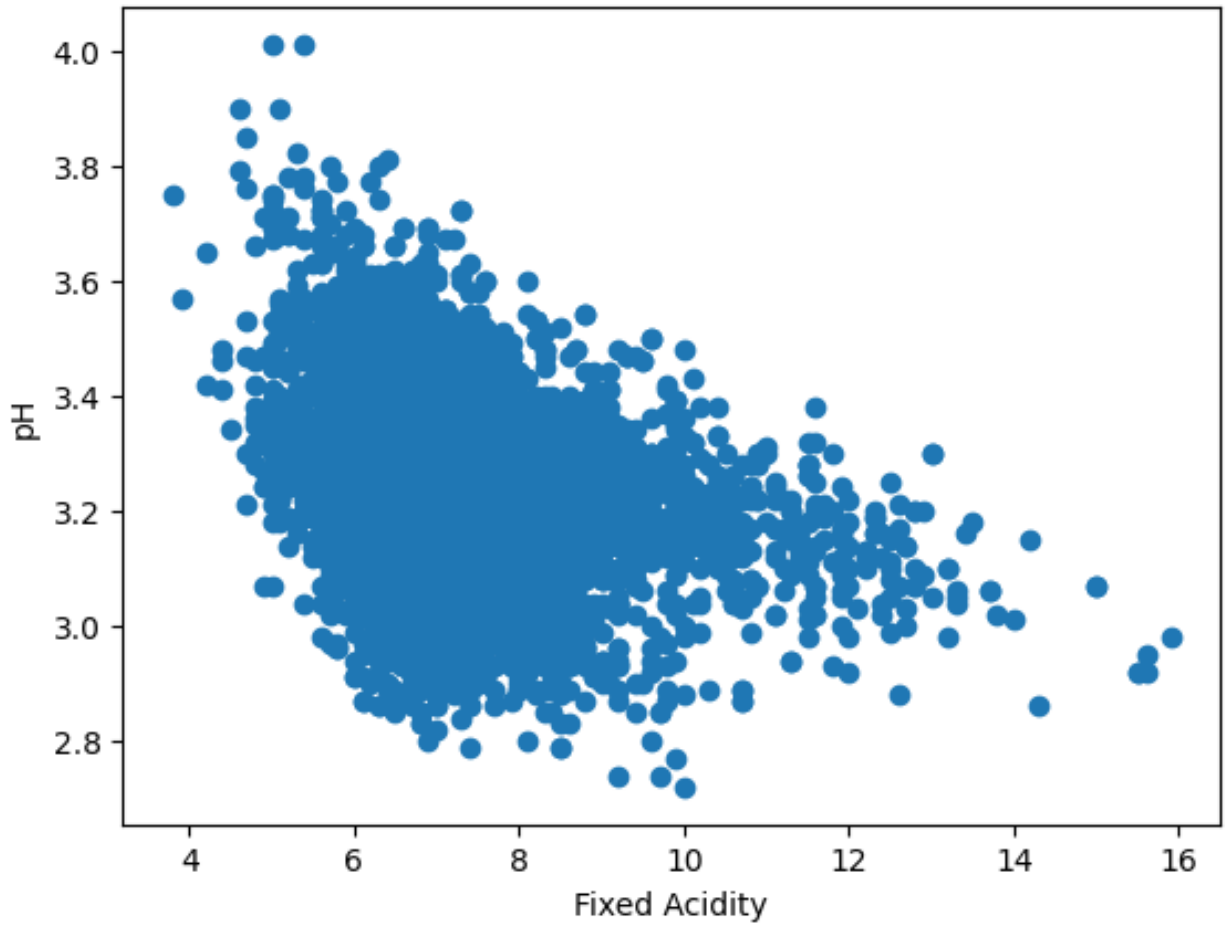
```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

df = pd.read_csv("/content/winequalityN.csv")
Data = {'x': df["fixed acidity"], 'y': df["pH"]}
df=pd.DataFrame(Data, columns=['x', 'y'])

plt.xlabel("Fixed Acidity")

plt.ylabel("pH")

plt.scatter(df['x'], df['y'])
plt.show()
```



```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

df = pd.read_csv("/content/winequalityN.csv")
Data = {'x': df["fixed acidity"], 'y': df["pH"]}
df = pd.DataFrame(Data, columns=['x', 'y'])

# Drop rows with missing values
df.dropna(inplace=True)

km = KMeans(n_clusters=5).fit(df)
centroids = km.cluster_centers_

plt.xlabel("Fixed Acidity")
plt.ylabel("pH")
plt.scatter(df['x'], df['y'], c=km.labels_.astype(float), s=60,
alpha=1)
plt.scatter(centroids[:, 0], centroids[:, 1], c='red', s=190)
plt.show()
```