Subject: 19CSE305

Lab Session: 05

Notes:

1. Please read the assignment notes carefully and comply to the guidelines provided.
2. Code should be checked into the GitHub. These details shall be provided in the Lab.
3. If you have not completed the prerequisite assignments, please complete them before the next lab session.

Coding Instructions:

1. The code should be modularized; The asked functionality should be available as a function. Please create multiple functions if needed. However, all functions should be present within a single code block, if you are using Jupyter or Colab notebooks.
2. There should be no print statement within the function. All print statements should be in the main program.
3. Please use proper naming of variables.
4. For lists, strings and matrices, you may use your input values as appropriate.
5. Please make inline documentation / comments as needed within the code blocks.

Main Section (Mandatory):

Please use the data associated with your own project. This assignment deals with classification models.

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabasz_score.html

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html

Please use help manuals of sklearn package to gain understanding of the model behaviors as well as ways to use various package functionalities.

A1. If your project deals with a regression problem, please use one attribute of your dataset (*X_train*) along with the target values (*y_train*) for training a linear regression model. Sample code suggested below.

```
from sklearn.linear_model import LinearRegression
reg = LinearRegression().fit(X_train, y_train)
y_train_pred = reg.predict(X_train)
```

(P.S.: If your project deals with a classification or clustering problem, please consider one of the attributes with numerical values as target values.)

A2. Calculate MSE, RMSE, MAPE and R2 scores for prediction made by the trained model in A1. Perform prediction on the test data and compare the metric values between train and test set.

A3. Repeat the exercises A1 and A2 with more than one attribute or all attributes.

A4. Perform k-means clustering on your data. Please remove / ignore the target variable for performing clustering. Sample code suggested below.

```
from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=2, random_state=0,
n_init="auto").fit(X_train)

kmeans.labels_

kmeans.cluster_centers_
```

(P.S.: Above code uses k = 2.)

A5. For the clustering done in A4, calculate the: (i) Silhouette Score, (ii) CH Score and (iii) DB Index.

```
from sklearn.metrics import silhouette_score

from sklearn.metrics import calinski_harabasz_score

from sklearn.metrics import davies_bouldin_score


kmeans = KMeans(n_clusters=2, random_state=42).fit(X_train)

silhouette_score(X_train, kmeans.labels_)

calinski_harabasz_score(X_train, kmeans.labels_)

davies_bouldin_score(X_train, kmeans.labels_)
```

A6. Perform k-means clustering for different values of k. Evaluate the above scores for each k value. Make a plot of the values against the k value to determine the optimal cluster count.

A7. Using elbow plot, determine the optimal k value for k-means clustering. Use below code.

```
for k in range(2, 20):
    kmeans = KMeans(n_clusters=k).fit(X_train)
    distorsions.append(kmeans.inertia_)
plot(distortions)
```

Report Assignment:

1. Update your understanding of your project in the introduction section of the report.
2. Study the downloaded papers & update the literature survey section of your report.
3. Expand the methodology and results sections with outcomes of this experiments & results obtained. Please discuss your observations, inferences in results & discussion section. Please conclude the report appropriately with these experiments. Consider following points for observation analysis & inferences.