

SCDM LAB ASSIGNMENT- 4

▼ 1. Load 'spambase.csv' file in jupyter notebook.

```
import pandas as pd
```

```
df = pd.read_csv("spambase.csv")
```

▼ 2. Print the total instances per class of the above dataset

```
df
```

	0	0.64	0.64.1	0.1	0.32	0.2	0.3	0.4	0.5	0.6	0.7	0.64.2	0.8	0.9
0	0.21	0.28	0.50	0.0	0.14	0.28	0.21	0.07	0.00	0.94	0.21	0.79	0.65	0.21
1	0.06	0.00	0.71	0.0	1.23	0.19	0.19	0.12	0.64	0.25	0.38	0.45	0.12	0.00
2	0.00	0.00	0.00	0.0	0.63	0.00	0.31	0.63	0.31	0.63	0.31	0.31	0.31	0.00
3	0.00	0.00	0.00	0.0	0.63	0.00	0.31	0.63	0.31	0.63	0.31	0.31	0.31	0.00
4	0.00	0.00	0.00	0.0	1.85	0.00	0.00	1.85	0.00	0.00	0.00	0.00	0.00	0.00
...
4595	0.31	0.00	0.62	0.0	0.00	0.31	0.00	0.00	0.00	0.00	0.00	1.88	0.00	0.00
4596	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4597	0.30	0.00	0.30	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.80	0.30	0.00
4598	0.96	0.00	0.00	0.0	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.00	0.00
4599	0.00	0.00	0.65	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.65	0.00

4600 rows × 58 columns



▼ 3. Split the data part and label part of the above dataset.

```
[row, col] = df.shape
Data = df.iloc[0 : row, 0 : (col - 1)]
Label = df.iloc[0 : row, (col - 1)]
```

Data

	0	0.64	0.64.1	0.1	0.32	0.2	0.3	0.4	0.5	0.6	0.7	0.64.2	0.8	0.9
0	0.21	0.28	0.50	0.0	0.14	0.28	0.21	0.07	0.00	0.94	0.21	0.79	0.65	0.21
1	0.06	0.00	0.71	0.0	1.23	0.19	0.19	0.12	0.64	0.25	0.38	0.45	0.12	0.00
2	0.00	0.00	0.00	0.0	0.63	0.00	0.31	0.63	0.31	0.63	0.31	0.31	0.31	0.00
3	0.00	0.00	0.00	0.0	0.63	0.00	0.31	0.63	0.31	0.63	0.31	0.31	0.31	0.00
4	0.00	0.00	0.00	0.0	1.85	0.00	0.00	1.85	0.00	0.00	0.00	0.00	0.00	0.00
...
4595	0.31	0.00	0.62	0.0	0.00	0.31	0.00	0.00	0.00	0.00	0.00	1.88	0.00	0.00
4596	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4597	0.30	0.00	0.30	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.80	0.30	0.00
4598	0.96	0.00	0.00	0.0	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.00	0.00
4599	0.00	0.00	0.65	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.65	0.00

4600 rows × 57 columns



Label

```
0      1
1      1
2      1
3      1
4      1
..
4595   0
4596   0
4597   0
4598   0
4599   0
Name: 1, Length: 4600, dtype: int64
```



```
from sklearn.model_selection import train_test_split
from sklearn.cluster import KMeans
X=Data
y=Label
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=42)
kmns=KMeans(n_clusters=2, random_state=0)
kmns.fit(X_train,y_train)
```

```
KMeans(n_clusters=2, random_state=0)
```

```
predictions = kmns.predict(X_test)
print(predictions)
```

[illegible]

```
from sklearn.metrics import accuracy_score
accuracy_score(y_test, predictions)
```

0.5956521739130435

5. Compare the accuracy of spamhase dataset with the

3. Compare the accuracy of spainbase dataset with the heart disease dataset by using k-means algorithm.

```
df = pd.read_csv("heart_failure_clinical_records_dataset.csv")

[row, col] = df.shape
Data = df.iloc[0 : row, 0 : (col - 1)]
Label = df.iloc[0 : row, (col - 1)]

from sklearn.model_selection import train_test_split
from sklearn.cluster import KMeans
X=Data
y=Label
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20,random_state=42)
kmns=KMeans(n_clusters=2, random_state=0)
kmns.fit(X_train,y_train)

KMeans(n_clusters=2, random_state=0)

predictions = kmns.predict(X_test)
print(predictions)

[0 1 1 1 0 0 0 1 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1 1 1
 0 0 0 0 0 0 1 0 0 1 0 1 0 1 0 1 1 0 0 0 0 0 0]

from sklearn.metrics import accuracy_score
accuracy_score(y_test, predictions)

0.5833333333333334
```

✓ 0s completed at 11:31 AM

● ✕