ML Lab – 7

Praneesh R V
CB.SC.U4CYS23036

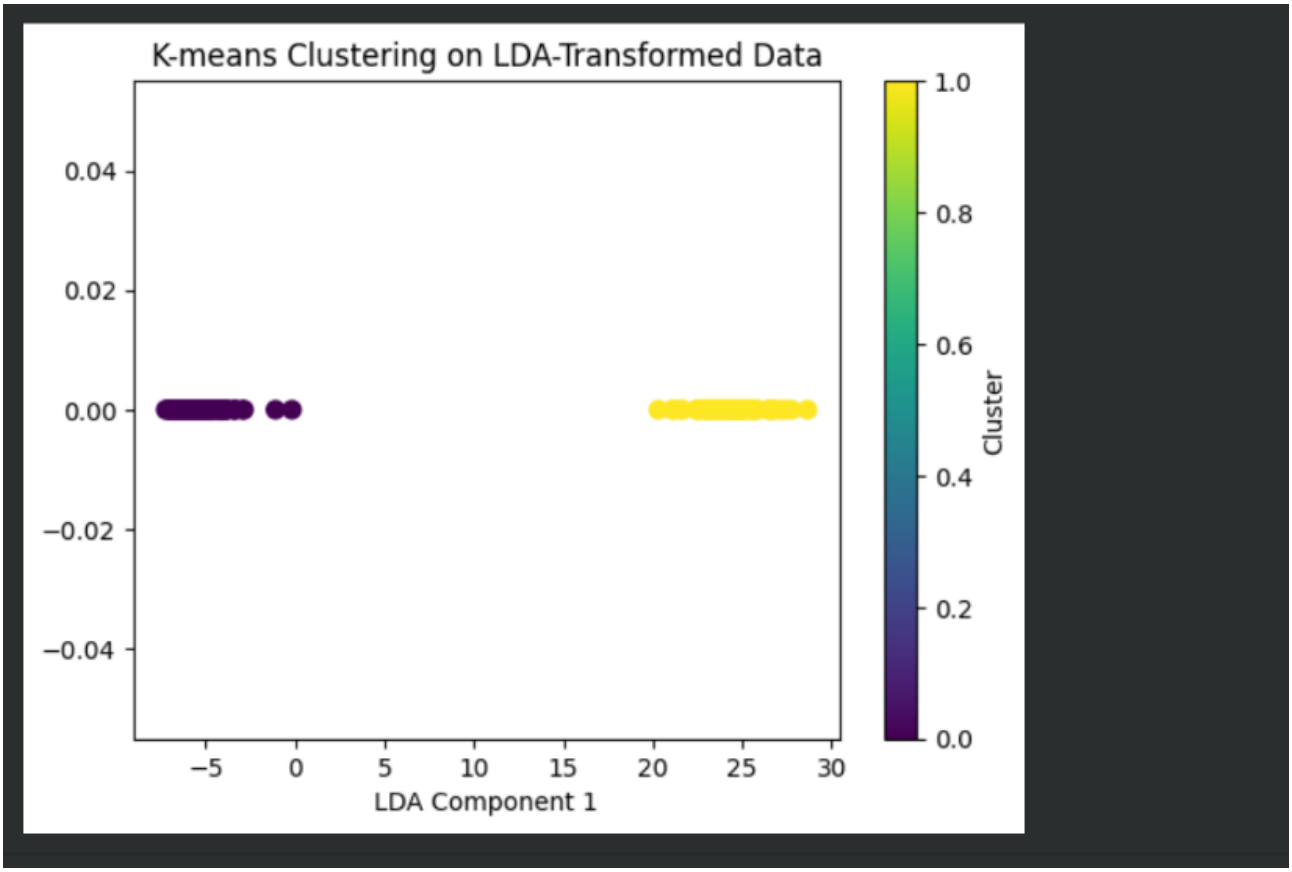
Code:

```python
import numpy as np
import pandas as pd
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.cluster import KMeans
from sklearn.metrics import confusion_matrix, classification_report
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
# Load example dataset (replace with your malware dataset)
malware = pd.read_csv("uci_malware_detection.csv")
print(malware.head())
X = malware.drop(columns=['Label'])  # Features
y = malware['Label']  # Target labels -  2 classes - malicious and non-malacious

# Convert categorical labels to numerical values
label_encoder = LabelEncoder()
y = label_encoder.fit_transform(y)

# Step 1: Apply LDA for dimensionality reduction
lda = LinearDiscriminantAnalysis(n_components=1)  # Reduce to 1 dimension
X_lda = lda.fit_transform(X, y)
# Step 2: Apply K-means clustering
kmeans = KMeans(n_clusters=2, random_state=42)  # Assume 2 malware classes - only 2
kmeans.fit(X_lda)
y_kmeans = kmeans.predict(X_lda)
# Step 3: Evaluate clustering performance
print("Confusion Matrix:")
print(confusion_matrix(y, y_kmeans))
print("\nClassification Report:")
print(classification_report(y, y_kmeans))
# Step 4: Visualize the clusters
plt.scatter(X_lda[:, 0], np.zeros_like(X_lda[:, 0]), c=y_kmeans, cmap='viridis', s=50)
plt.title("K-means Clustering on LDA-Transformed Data")
plt.xlabel("LDA Component 1")
plt.ylabel("")
plt.colorbar(label="Cluster")
plt.show()
```

Output:




```
           Label  F_1  F_2  F_3  F_4  F_5  F_6  F_7  F_8  F_9  ...  F_522  \
0  non-malicious    1    0    1    0    1    0    1    0    1  ...      0
1  non-malicious    1    0    1    0    1    0    1    0    1  ...      0
2  non-malicious    1    0    1    0    1    0    1    0    1  ...      0
3  non-malicious    1    0    1    0    1    0    1    0    1  ...      0
4  non-malicious    1    0    1    0    1    0    1    0    1  ...      0

   F_523  F_524  F_525  F_526  F_527  F_528  F_529  F_530  F_531
0      0      0      0      0      0      0      0      0      0
1      0      0      0      0      0      0      0      0      0
2      0      0      0      0      0      0      0      0      0
3      0      0      0      0      0      0      0      0      0
4      0      0      0      0      0      0      0      0      0

[5 rows x 532 columns]
Confusion Matrix:
[[301   0]
 [  0  72]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       301
           1       1.00      1.00      1.00        72

    accuracy                           1.00       373
   macro avg       1.00      1.00      1.00       373
weighted avg       1.00      1.00      1.00       373
```