Problem 8

Loading the data

```
import pandas as pd
from sklearn.cluster import SpectralClustering
from sklearn.metrics import adjusted_rand_score,
adjusted_mutual_info_score, silhouette_score
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler

data = pd.read_csv('/content/drive/MyDrive/sem 7/ID5055/Assignment
3/Problem 8/gene_expression.csv')
data.head()
```

	Gene One	Gene Two	Cancer Present
0	4.3	3.9	1
1	2.5	6.3	0
2	5.7	3.9	1
3	6.1	6.2	0
4	7.4	3.4	1

```
df = data
```

Implement spectral clustering using Python and scikit-learn to identify clusters of co-expressed genes within the dataset.

```
n_clusters = 2
X = df.iloc[:, :2]
X_scaled = StandardScaler().fit_transform(X)
spectral = SpectralClustering(n_clusters=n_clusters,
affinity='nearest_neighbors', random_state=0)
df['Cluster'] = spectral.fit_predict(X_scaled)

df.head()
```

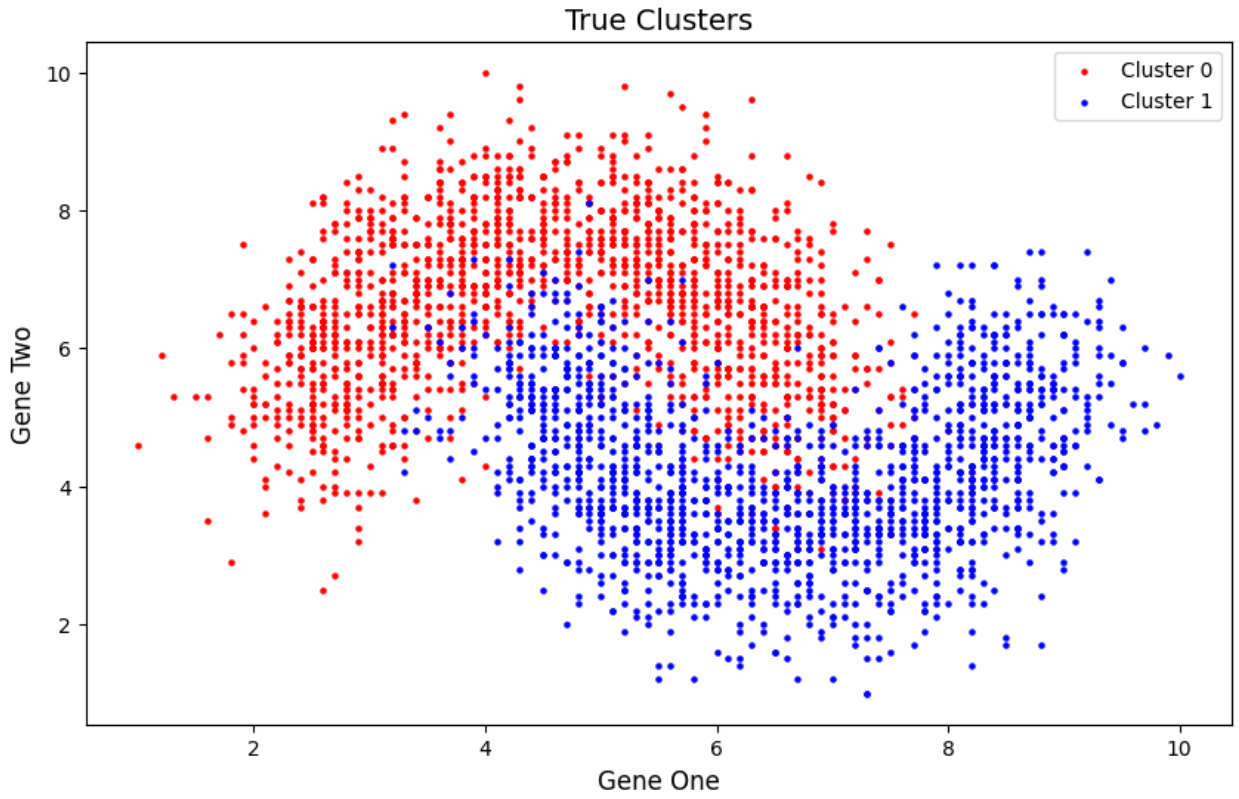
	Gene One	Gene Two	Cancer Present	Cluster
0	4.3	3.9	1	0
1	2.5	6.3	0	1
2	5.7	3.9	1	0
3	6.1	6.2	0	1
4	7.4	3.4	1	0

Create visualizations for the true clusters based on the information in the 3rd column of the dataset.

```
plt.figure(figsize=(10, 6))
colors = ['red', 'blue']

for i in range(n_clusters):
    cluster_data = df[df['Cancer Present'] == i]
    plt.scatter(cluster_data['Gene One'], cluster_data['Gene Two'],
color=colors[i], label=f'Cluster {i}', s = 5)

plt.title('True Clusters', fontsize = 14)
plt.xlabel('Gene One', fontsize = 12)
plt.ylabel('Gene Two', fontsize = 12)
plt.legend()
plt.show()
```



Evaluate and provide insights on the outcomes, including a comprehensive report on performance metrics such as Adjusted Rand Index, Adjusted Mutual Information, and Silhouette Score.

```
ari = adjusted_rand_score(df['Cancer Present'], df['Cluster'])
ami = adjusted_mutual_info_score(df['Cancer Present'], df['Cluster'])
silhouette = silhouette_score(X_scaled, df['Cluster'])
```

```
print(f'Adjusted Rand Index: {ari}')
print(f'Adjusted Mutual Information: {ami}')
print(f'Silhouette Score: {silhouette}')
```

```
Adjusted Rand Index: 0.5153624060863754
Adjusted Mutual Information: 0.413040052343147
Silhouette Score: 0.4540574993371185
```

1. Adjusted Rand Index (ARI): 0.515 -

- The ARI measures the similarity between the true clusters (Cancer Present) and the clusters generated by the spectral clustering algorithm.

- An ARI score of 0.515 indicates a moderate degree of similarity between the true clusters and the clusters identified by the algorithm.
- This suggests that while spectral clustering is capturing some underlying structure in the data, there may still be room for improvement.

1. Adjusted Mutual Information (AMI): 0.413

- The AMI is another measure of the agreement between the true clusters and the clusters produced by the algorithm.
- An AMI score of 0.413 suggests a moderate level of mutual information between the true clusters and the algorithm's clusters.
- Similar to the ARI, this indicates that the spectral clustering algorithm is providing some meaningful clustering, but it may not be capturing all of the underlying patterns in the data.

1. Silhouette Score: 0.454

- The Silhouette Score measures the quality of the clusters themselves. It assesses how well-separated the clusters are and how similar the data points within each cluster are to each other.
- A Silhouette Score of 0.454 is relatively high, indicating that the clusters are reasonably well-separated and that data points within each cluster are similar to each other.
- This suggests that the algorithm is successful in creating meaningful clusters, and the clusters are relatively distinct from each other.