Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Рубежный контроль №1
по дисциплине
«Методы машинного обучения»
на тему
«Кластеризация и метрики качества»

Выполнил:
студент группы ИУ5-24М
Зубаиров В. А.

_____

Москва — 2020 г.

1. Задание

1.1. 1. решить задачу кластеризации с использованием методов:

1) MeanShift
2) спектральная кластеризация
3) иерархическая кластеризация.

1.2. 2. Оценить качество модели на основе подходящих метрик качества (не менее двух метрик, если это возможно).

1.3. 3. Сделать выводы о качестве построенных моделей?

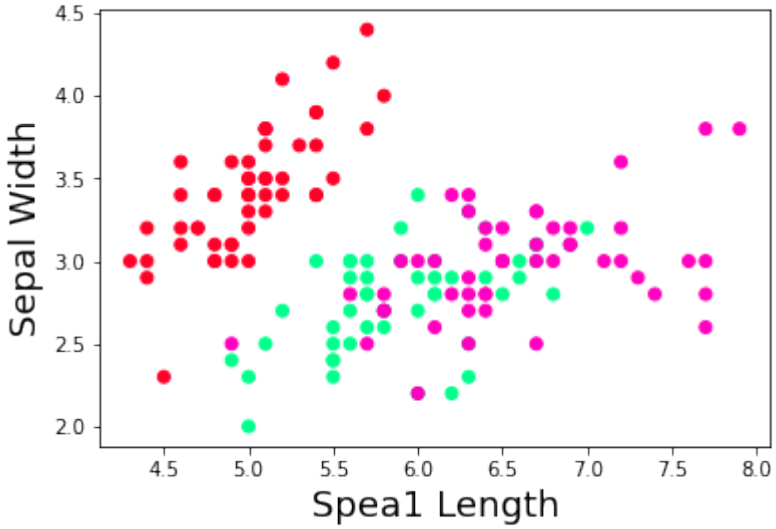
[73]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
from sklearn import preprocessing
from sklearn.datasets import load_iris
```

[74]:
```
iris = load_iris()
```

[75]:
```
X = iris.data[:, :2]
Y = iris.target
```

[76]:
```
plt.scatter(X[:,0], X[:,1], c=Y, cmap='gist_rainbow')
plt.xlabel('Spea1 Length', fontsize=18)
plt.ylabel('Sepal Width', fontsize=18)
```

[76]: Text(0, 0.5, 'Sepal Width')



[77]:
```
from sklearn.cluster import MeanShift
```

```
ms = MeanShift()
ms.fit(X)
```

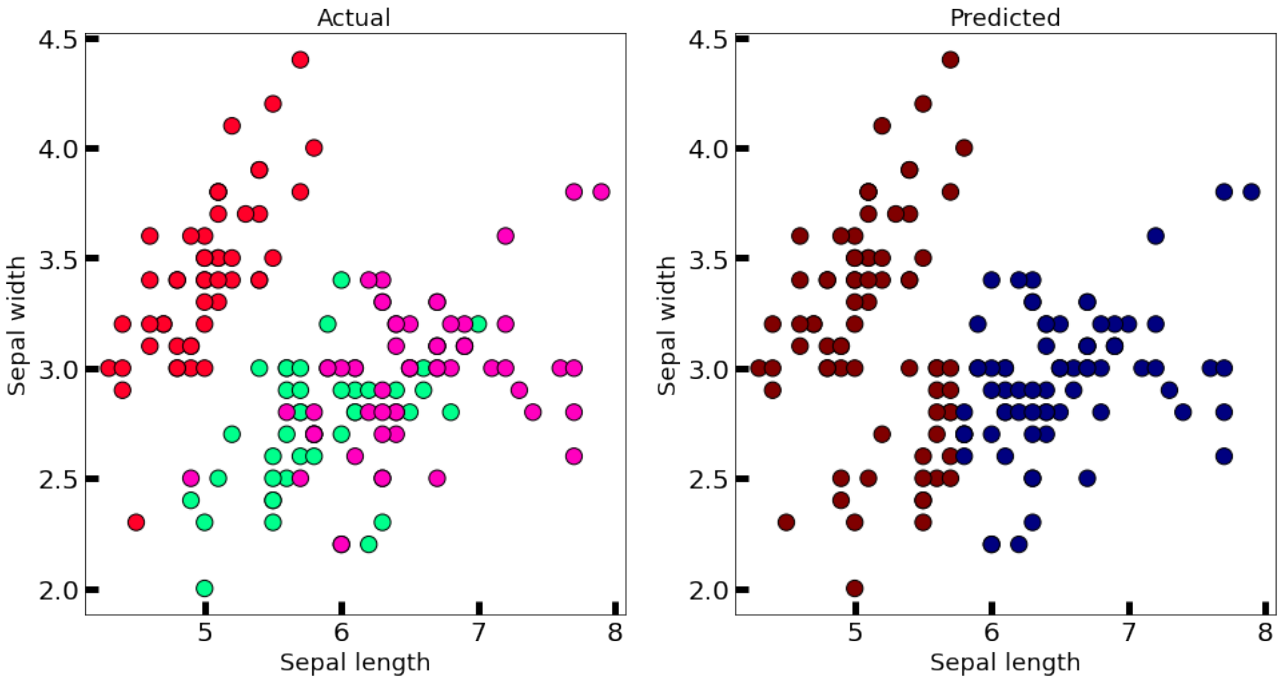
[77]:
```
MeanShift(bandwidth=None, bin_seeding=False, cluster_all=True, max_iter=300,
     min_bin_freq=1, n_jobs=None, seeds=None)
```

[78]:
```
centers = ms.cluster_centers_
centers
```

[78]:
```
array([[6.22      , 2.892     ],
       [5.41142857, 3.03285714]])
```

[79]:
```
new_labels = ms.labels_
# Plot the identified clusters and compare with the answers
fig, axes = plt.subplots(1, 2, figsize=(16,8))
axes[0].scatter(X[:, 0], X[:, 1], c=Y, cmap='gist_rainbow',
edgecolor='k', s=150)
axes[1].scatter(X[:, 0], X[:, 1], c=new_labels, cmap='jet',
edgecolor='k', s=150)
axes[0].set_xlabel('Sepal length', fontsize=18)
axes[0].set_ylabel('Sepal width', fontsize=18)
axes[1].set_xlabel('Sepal length', fontsize=18)
axes[1].set_ylabel('Sepal width', fontsize=18)
axes[0].tick_params(direction='in', length=10, width=5, colors='k', labelsize=20)
axes[1].tick_params(direction='in', length=10, width=5, colors='k', labelsize=20)
axes[0].set_title('Actual', fontsize=18)
axes[1].set_title('Predicted', fontsize=18)
```

[79]:
```
Text(0.5, 1.0, 'Predicted')
```



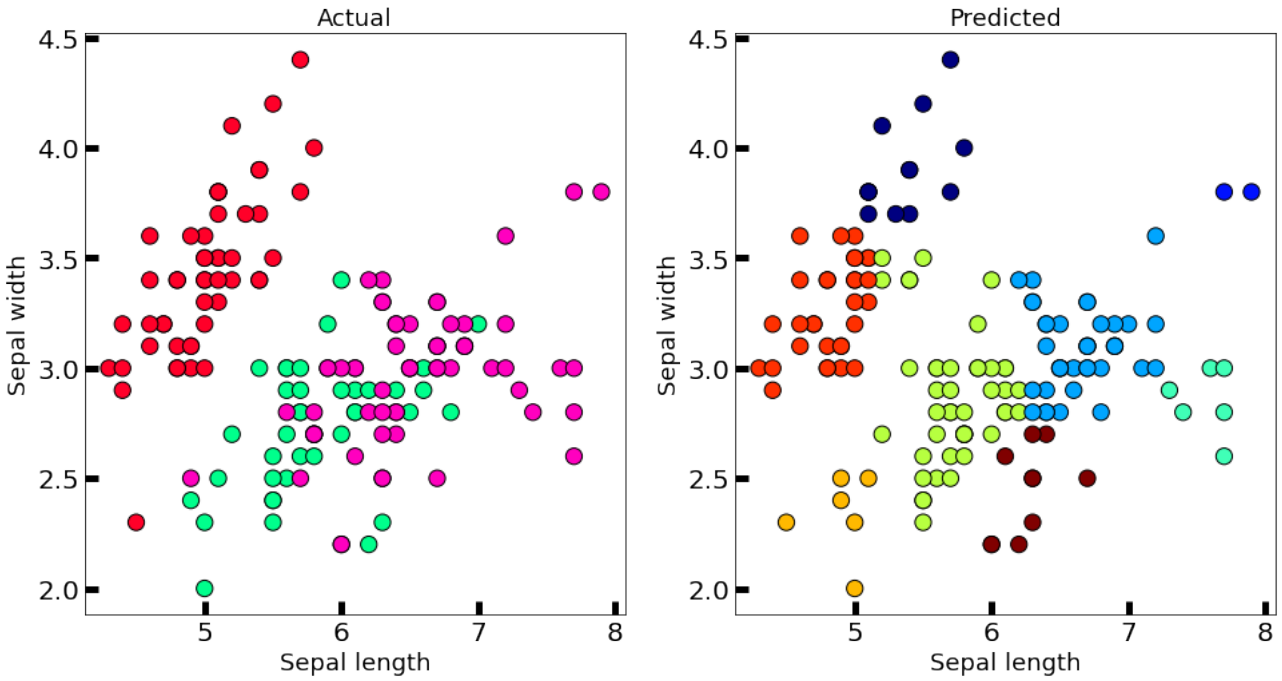
[83]:
```
from sklearn.cluster import SpectralClustering

sc = SpectralClustering()
sc.fit(X)
```

[83]:
```
SpectralClustering(affinity='rbf', assign_labels='kmeans', coef0=1, degree=3,
                   eigen_solver=None, eigen_tol=0.0, gamma=1.0,
                   kernel_params=None, n_clusters=8, n_components=None,
                   n_init=10, n_jobs=None, n_neighbors=10, random_state=None)
```

[84]:
```
new_labels = sc.labels_
# Plot the identified clusters and compare with the answers
fig, axes = plt.subplots(1, 2, figsize=(16,8))
axes[0].scatter(X[:, 0], X[:, 1], c=Y, cmap='gist_rainbow',
edgecolor='k', s=150)
axes[1].scatter(X[:, 0], X[:, 1], c=new_labels, cmap='jet',
edgecolor='k', s=150)
axes[0].set_xlabel('Sepal length', fontsize=18)
axes[0].set_ylabel('Sepal width', fontsize=18)
axes[1].set_xlabel('Sepal length', fontsize=18)
axes[1].set_ylabel('Sepal width', fontsize=18)
axes[0].tick_params(direction='in', length=10, width=5, colors='k', labelsize=20)
axes[1].tick_params(direction='in', length=10, width=5, colors='k', labelsize=20)
axes[0].set_title('Actual', fontsize=18)
axes[1].set_title('Predicted', fontsize=18)
```

[84]:
```
Text(0.5, 1.0, 'Predicted')
```

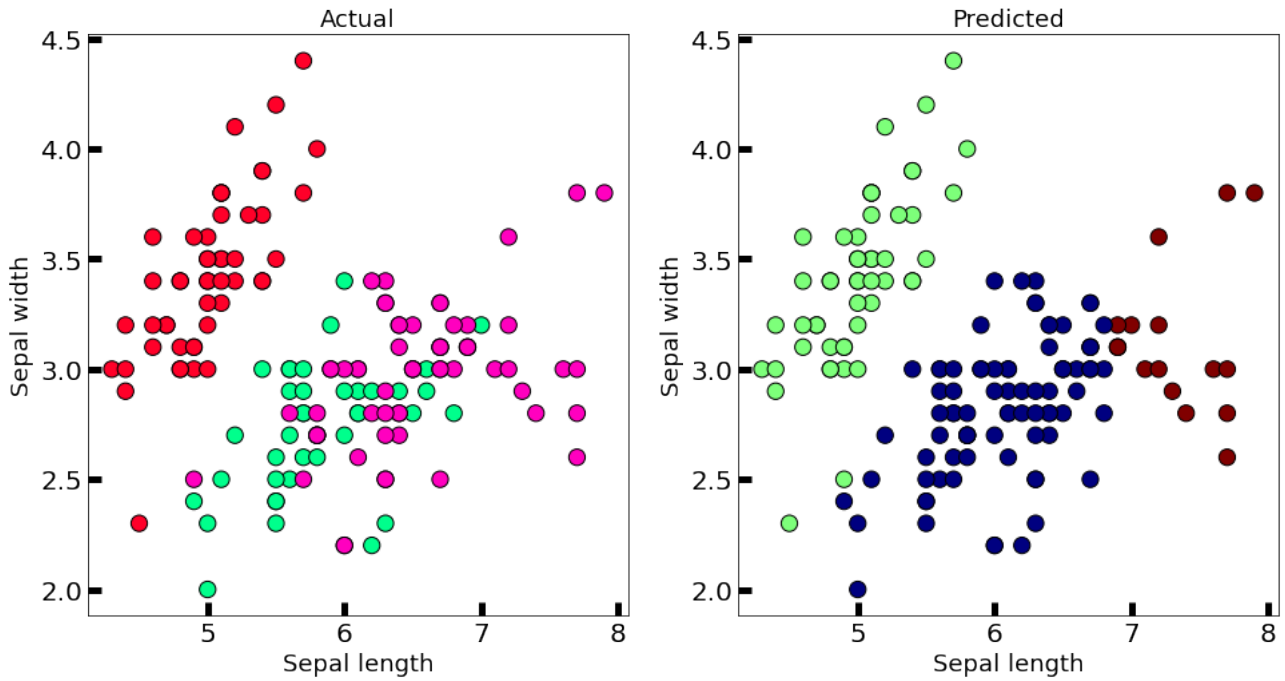


[85]:
```
sc2 = SpectralClustering(n_clusters=3)
sc2.fit(X)
```

[85]:
```
SpectralClustering(affinity='rbf', assign_labels='kmeans', coef0=1, degree=3,
                   eigen_solver=None, eigen_tol=0.0, gamma=1.0,
                   kernel_params=None, n_clusters=3, n_components=None,
                   n_init=10, n_jobs=None, n_neighbors=10, random_state=None)
```

[86]:
```
new_labels = sc2.labels_
# Plot the identified clusters and compare with the answers
fig, axes = plt.subplots(1, 2, figsize=(16,8))
axes[0].scatter(X[:, 0], X[:, 1], c=Y, cmap='gist_rainbow',
edgecolor='k', s=150)
axes[1].scatter(X[:, 0], X[:, 1], c=new_labels, cmap='jet',
edgecolor='k', s=150)
axes[0].set_xlabel('Sepal length', fontsize=18)
axes[0].set_ylabel('Sepal width', fontsize=18)
axes[1].set_xlabel('Sepal length', fontsize=18)
axes[1].set_ylabel('Sepal width', fontsize=18)
axes[0].tick_params(direction='in', length=10, width=5, colors='k', labelsize=20)
axes[1].tick_params(direction='in', length=10, width=5, colors='k', labelsize=20)
axes[0].set_title('Actual', fontsize=18)
axes[1].set_title('Predicted', fontsize=18)
```

[86]:
```
Text(0.5, 1.0, 'Predicted')
```



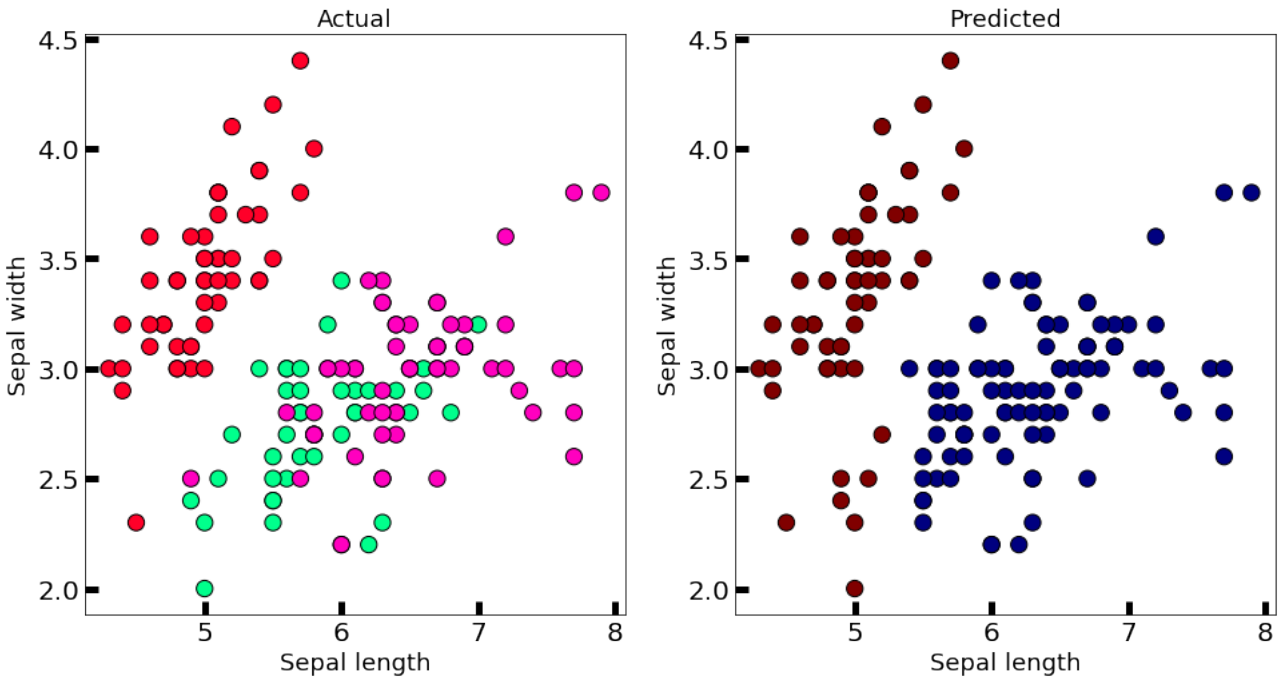
[87]:
```
from sklearn.cluster import AgglomerativeClustering
```

[88]:
```
ag = AgglomerativeClustering()
ag.fit(X)
```

[88]:
```
AgglomerativeClustering(affinity='euclidean', compute_full_tree='auto',
                        connectivity=None, distance_threshold=None,
                        linkage='ward', memory=None, n_clusters=2)
```

[89]:
```
new_labels = ag.labels_
# Plot the identified clusters and compare with the answers
fig, axes = plt.subplots(1, 2, figsize=(16,8))
axes[0].scatter(X[:, 0], X[:, 1], c=Y, cmap='gist_rainbow',
edgecolor='k', s=150)
axes[1].scatter(X[:, 0], X[:, 1], c=new_labels, cmap='jet',
edgecolor='k', s=150)
axes[0].set_xlabel('Sepal length', fontsize=18)
axes[0].set_ylabel('Sepal width', fontsize=18)
axes[1].set_xlabel('Sepal length', fontsize=18)
axes[1].set_ylabel('Sepal width', fontsize=18)
axes[0].tick_params(direction='in', length=10, width=5, colors='k', labelsize=20)
axes[1].tick_params(direction='in', length=10, width=5, colors='k', labelsize=20)
axes[0].set_title('Actual', fontsize=18)
axes[1].set_title('Predicted', fontsize=18)
```

[89]:
```
Text(0.5, 1.0, 'Predicted')
```

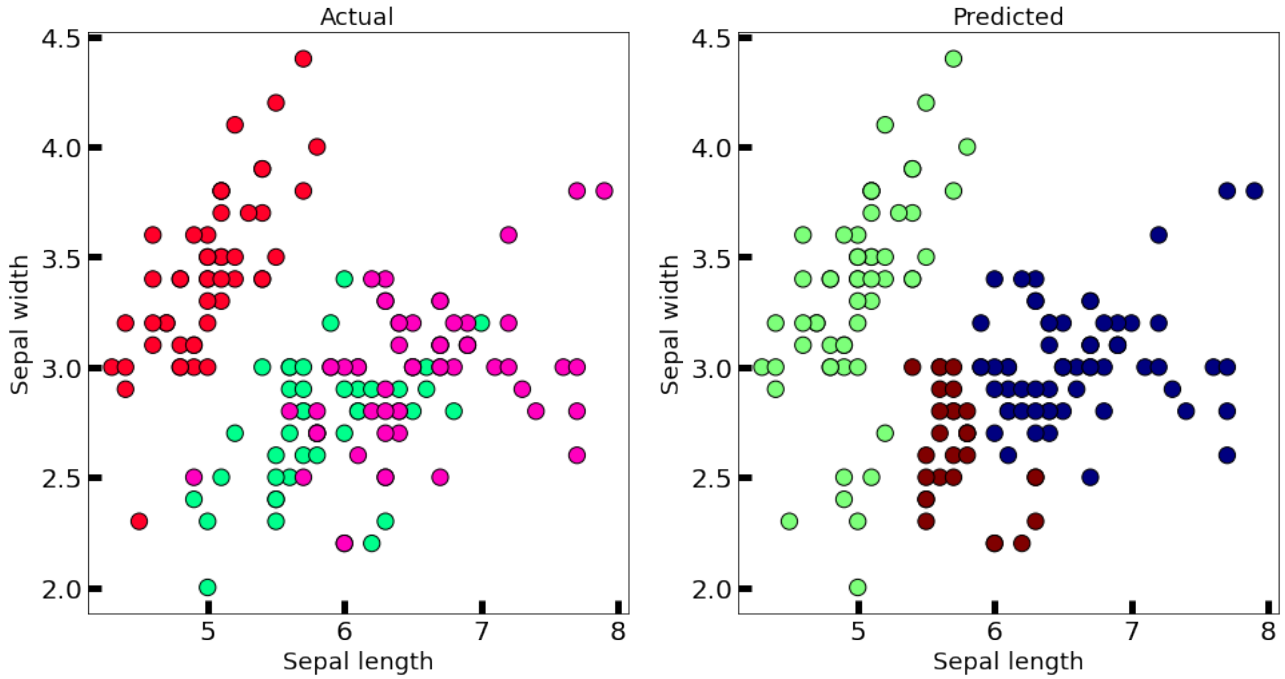


[90]:
```
ag2 = AgglomerativeClustering(n_clusters=3)
ag2.fit(X)

new_labels = ag2.labels_
```

```
# Plot the identified clusters and compare with the answers
fig, axes = plt.subplots(1, 2, figsize=(16,8))
axes[0].scatter(X[:, 0], X[:, 1], c=Y, cmap='gist_rainbow',
edgecolor='k', s=150)
axes[1].scatter(X[:, 0], X[:, 1], c=new_labels, cmap='jet',
edgecolor='k', s=150)
axes[0].set_xlabel('Sepal length', fontsize=18)
axes[0].set_ylabel('Sepal width', fontsize=18)
axes[1].set_xlabel('Sepal length', fontsize=18)
axes[1].set_ylabel('Sepal width', fontsize=18)
axes[0].tick_params(direction='in', length=10, width=5, colors='k', labelsize=20)
axes[1].tick_params(direction='in', length=10, width=5, colors='k', labelsize=20)
axes[0].set_title('Actual', fontsize=18)
axes[1].set_title('Predicted', fontsize=18)
```

[90]: Text(0.5, 1.0, 'Predicted')



1.4. Метрики

[91]:
```
from sklearn.metrics import adjusted_rand_score
from sklearn.metrics import adjusted_mutual_info_score
from sklearn.metrics import homogeneity_completeness_v_measure
from sklearn.metrics import silhouette_score
```

[99]:
```
def count_metrics(name, method):
    tmp = method.fit_predict(X)
    print("Dataset: " + name)
```

```
    print("ARI: "+ str(adjusted_rand_score(Y, tmp)))
    print("AMI: "+  str(adjusted_mutual_info_score(Y, tmp)))
    h, c, v = homogeneity_completeness_v_measure(Y, tmp)
    print("HCVm: Homogeneity - " + str(h) + "\nCompleteness - " + str(c) + "\nV-measure -
    ↪"+str(v))
    print("SL: " + str(silhouette_score(X, tmp)))
    print("==================================")
```

[100]:
```
count_metrics("MeanShift", MeanShift())
count_metrics("Spectral default", SpectralClustering())
count_metrics("Spectral 3", SpectralClustering(n_clusters=3))
count_metrics("Agglomerative def", AgglomerativeClustering())
count_metrics("Agglomerative 3", AgglomerativeClustering(n_clusters=3))
```

```
Dataset: MeanShift
ARI: 0.3944401908806803
AMI: 0.4317743582900882
HCVm: Homogeneity - 0.355574438925241
Completeness - 0.5636444355672562
V-measure - 0.43606057162569084
SL: 0.4644681851183547
==================================
Dataset: Spectral default
ARI: 0.3103895058381067
AMI: 0.4078924556814485
HCVm: Homogeneity - 0.5607839300056536
Completeness - 0.34798815614173934
V-measure - 0.4294721828965487
SL: 0.36454192316615136
==================================
Dataset: Spectral 3
ARI: 0.5529473055759424
AMI: 0.6353736832348081
HCVm: Homogeneity - 0.595146173209358
Completeness - 0.6928341566599039
V-measure - 0.6402855500855817
SL: 0.4131437626307253
==================================
Dataset: Agglomerative def
ARI: 0.5114270772970757
AMI: 0.5875852748543551
HCVm: Homogeneity - 0.4730196835308308
Completeness - 0.7865303025387341
V-measure - 0.590757522780414
SL: 0.47767996898758924
==================================
Dataset: Agglomerative 3
ARI: 0.5112126489117526
AMI: 0.5240179186847511
HCVm: Homogeneity - 0.5190720845536648
```

Completeness - 0.5414839345877656
V-measure - 0.5300412040588491
SL: 0.3653346819163389
==========

Были использованы метрики кластеризации ARI (так как известны истинные метки), AMI, HSV, коэфф. силуэта).

По результатам можно сказать, что лучше всего справились спектральная классификация и иерархическая, так как ARI у них ближе к 1. Но значение коэффицентов все равно небольшие, поэтому сказать, что модель получилась хорошего качества, нельзя.