Лабораторна робота № 7. Дослідження методів кластеризації: K-Means, DBSCAN, Mean-Shift, Agglomerative Clustering, Fuzzy C-Means

Мета: сформулювати практичні навички у застосуванні та аналізі різних методів кластеризації, таких як K-Means, DBSCAN, Mean-Shift, Agglomerative Clustering, Fuzzy C-Means, з метою глибшого розуміння їх особливостей, переваг та обмежень. Розвинути здатність студентів до самостійного вибору найбільш коректного методу кластеризації для конкретного аналітичного завдання, а також навички роботи з реальними даними, включаючи попередню обробку, нормалізацію даних і візуалізацію результатів.

Теоретичні відомості

Методи кластеризації – це фундаментальні інструменти в машинному навчанні та аналізі даних, які дозволяють групувати схожі об'єкти в кластери. Вони відіграють ключову роль у зниженні складності даних, виявленні шаблонів та аномалій, а також у попередній обробці даних для подальшого аналізу.

K-**Means** – один з найпопулярніших методів кластеризації, який розділяє датасет на K кластери, мінімізуючи суму квадратів відстаней між точками та центрами кластерів. Цей метод чутливий до вибору початкових центрів і кількості кластерів.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) – метод кластеризації, заснований на густині, ідеально підходить для виявлення кластерів довільної форми та викидів. Він використовує поняття досяжності густини, щоб визначити кластери.

Mean-Shift – метод кластеризації, який використовує рух до вищих густин даних, визначаючи кластери без заздалегідь заданої кількості. Це робить його ефективним для виявлення кластерів нестандартних форм.

Agglomerative Clustering – метод ієрархічної кластеризації, який починає з того, що кожен об'єкт представляє собою окремий кластер, а потім послідовно об’єднує кластери на основі міри схожості, поки не буде досягнута задана кількість кластерів або інший критерій зупинки.

Fuzzy C-Means – це розширення K-Means, яке дозволяє об'єктам належати до кількох кластерів одночасно з різним ступенем приналежності, що забезпечує більшу гнучкість у виявленні кластерів, що перекриваються.

Приклад виконання роботи

Завдання

1. Оберіть та завантажте датасет відповідно до свого варіанту. Завантажте ваш набір даних як тимчасовий файл у середовище Google Colab та ініціалізуйте його.
2. Проаналізуйте датасет та визначте ознаки, які будуть використовуватися для кластеризації. Як цільову зміну кластеризації за замовчуванням визначено останній стовпець у наборах даних.
3. Виконайте очищення даних: видаліть або заповніть пропущені значення, опрацюйте викиди якщо це потребується.
4. Виконайте тренування моделей: 1) K-Means; 2) DBSCAN; 3) Mean-Shift; 4) Agglomerative Clustering; 5) Gaussian Mixture. Виконайте оцінку та візуалізуйте результати за наступними метриками: Silhouette для кожної моделі.
5. Проаналізуйте отримані результати, обґрунтуйте вибір найкращої моделі для даного набору даних.
6. За бажанням проекспериментуйте з гіперпараметрами для моделей задля покращення результатів.

Контрольне запитання №1

Вкажіть результуюче значення Silhouette для K-Means.
Відповідь: _____

Контрольне запитання №2

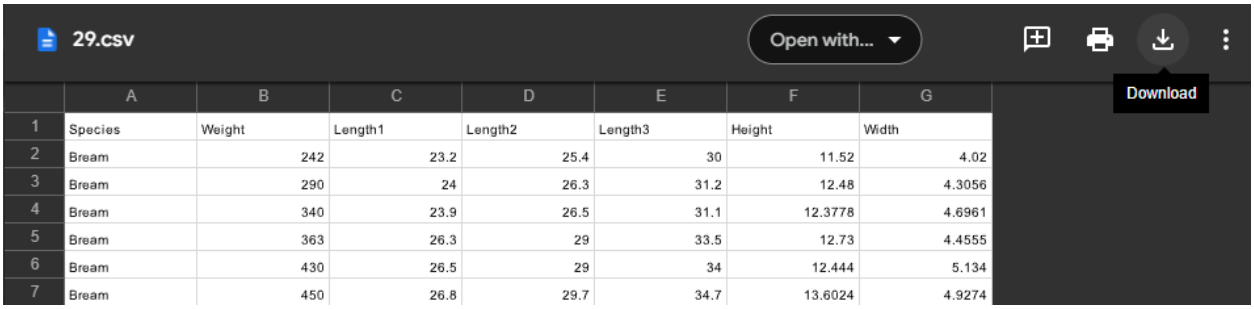
Вкажіть результуюче значення Silhouette для Agglomerative Clustering.
Відповідь: _____

Контрольне запитання №3

Вкажіть результуюче значення Silhouette для Gaussian Mixture.
Відповідь: _____

Хід роботи

1. Збережемо та імпортуємо набір даних за варіантом.



	A	B	C	D	E	F	G
1	Species	Weight	Length1	Length2	Length3	Height	Width
2	Bream	242	23.2	25.4	30	11.52	4.02
3	Bream	290	24	26.3	31.2	12.48	4.3056
4	Bream	340	23.9	26.5	31.1	12.3778	4.6961
5	Bream	363	26.3	29	33.5	12.73	4.4555
6	Bream	430	26.5	29	34	12.444	5.134
7	Bream	450	26.8	29.7	34.7	13.6024	4.9274

Рисунок 6.1. Датасет на Google Drive

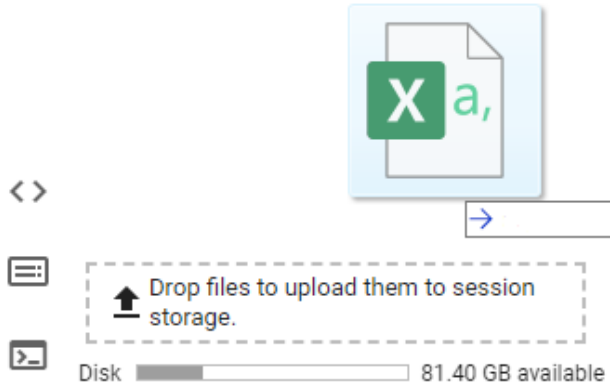


Рисунок 6.2. Завантаження датасету до Google Colab

2. Встановлення необхідних бібліотек та модулів

```
import warnings
warnings.filterwarnings('ignore')
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.cluster import KMeans, DBSCAN, MeanShift,
AgglomerativeClustering
from sklearn.mixture import GaussianMixture
from sklearn.metrics import silhouette_score
import ipywidgets as widgets
from IPython.display import display
```

3. Ініціалізація набору даних в Google Colab.

```
# Завдання №1. # Завантаження датасету
data = pd.read_csv("/content/5.csv")
features = data.columns[:-1].tolist()  # Отримання списку ознак
data.head()
```

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	1	337	118	4	4.5	4.5	9.65	1	0.92
1	2	324	107	4	4.0	4.5	8.87	1	0.76
2	3	316	104	3	3.0	3.5	8.00	1	0.72
3	4	322	110	3	3.5	2.5	8.67	1	0.80
4	5	314	103	2	2.0	3.0	8.21	0	0.65

Рисунок 6.3. Перші 5 рядків набору даних

4. Визначимо цільову змінну, як Chance of Admit, ознаками будуть: GRE Score та TOEFL Score. Обрані ознаки простираються на широкий діапазон значень та можуть бути досить інформативними під час кластеризації.

5. Виконаємо очищення даних

```
# Перевірка на відсутні значення
print(data.isnull().sum())
```

```
Serial No.          0
GRE Score           0
TOEFL Score         0
University Rating   0
SOP                 0
LOR                 0
CGPA                0
Research            0
Chance of Admit     0
dtype: int64
```

Рисунок 6.4. Пошук нульових значень у датасеті

Порожні значення відсутні, тож видалення або заповнення виконувати не треба.

```
# Видалення або заповнення відсутніх значень
# data.dropna(inplace=True) # Видалення

# data.fillna(data.mean(), inplace=True) # Заповнення середніми значеннями
```

Виконаємо візуалізацію однієї з ознак набору даних, а саме «Length2»

```
# Візуалізація викидів
plt.figure(figsize=(10, 6))
sns.boxplot(x=data['GRE Score'])
plt.title('Boxplot before removing outliers')
plt.show()
```

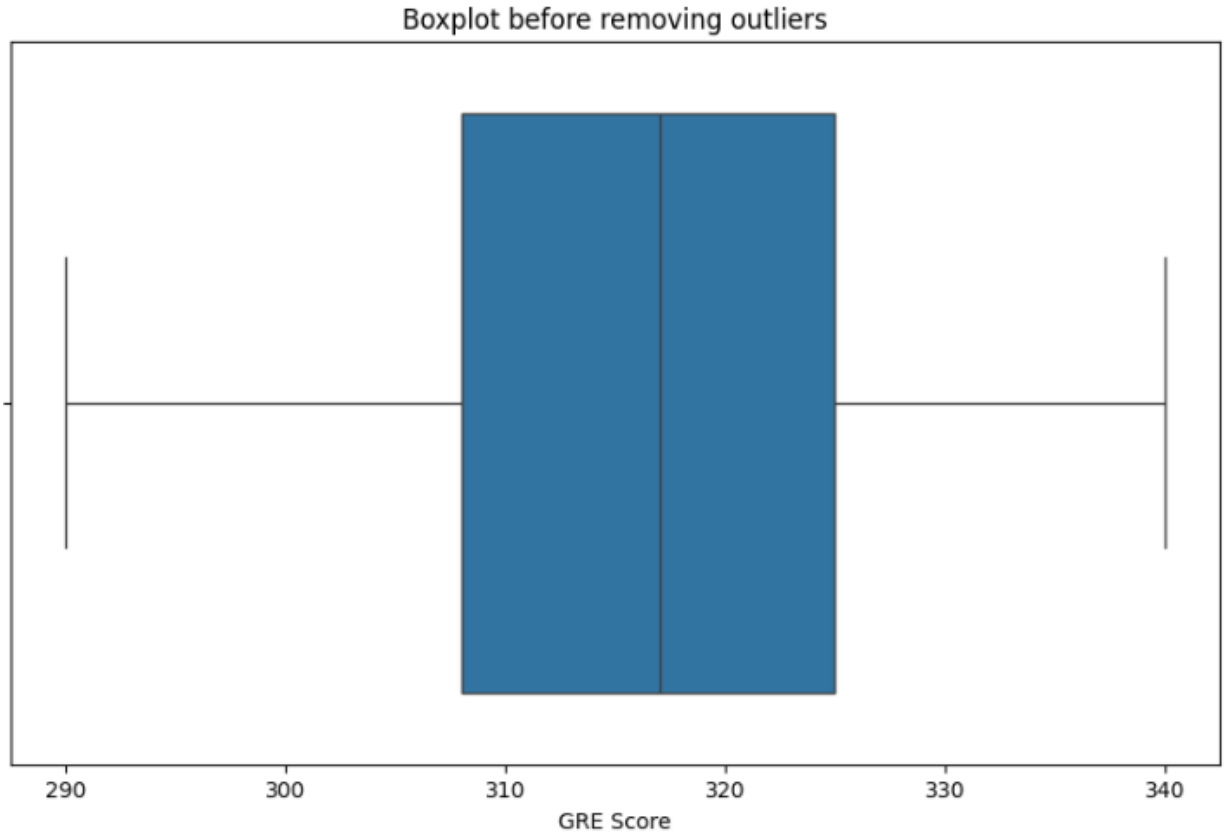


Рисунок 6.5. Діаграма викидів для GRE Score перед очищенням

Викидів не зафіксовано, тож додаткове очищення виконувати немає сенсу.

6. Побудуємо логіку взаємодії з віджетом задля обробки методів кластеризації

```
# Віджети для вибору ознак
selected_features = []

def add_feature(b):
    if feature_dropdown.value and feature_dropdown.value not in
selected_features:
        selected_features.append(feature_dropdown.value)
        update_feature_list()
```

```
def remove_feature(b):
    if feature_list_widget.value:
        selected_features.remove(feature_list_widget.value)
        update_feature_list()

def update_feature_list():
    feature_list_widget.options = selected_features

feature_dropdown = widgets.Dropdown(options=features,
description='Features:')
add_button = widgets.Button(description="Add Feature")
remove_button = widgets.Button(description="Remove Feature")
feature_list_widget = widgets.Select(options=selected_features,
description="Selected Features:")

add_button.on_click(add_feature)
remove_button.on_click(remove_feature)

# Віджет для вибору методу кластеризації
clustering_widget = widgets.Dropdown(
    options=['K-Means', 'DBSCAN', 'Mean-Shift', 'Agglomerative
Clustering', 'Gaussian Mixture'],
    description='Method:',
)

# Віджет для вибору гіперпараметру EPS для DBSCAN
eps_widget = widgets.FloatSlider(description='EPS:', min=0.1, max=10.0,
value=0.5, step=0.1)

# Віджет для вибору гіперпараметру Min Samples для DBSCAN
min_samples_widget = widgets.IntSlider(description='Min Samples:', min=1,
max=20, value=5)

n_clusters_widget = widgets.IntSlider(
    value=3,
    min=2,
    max=10,
    step=1,
    description='N Clusters:',
    disabled=False,
    continuous_update=False,
    orientation='horizontal',
    readout=True,
    readout_format='d'
)

build_button = widgets.Button(description="Build Clusters")

output = widgets.Output()
```

```
method_widget = widgets.Dropdown(
    options=['K-Means', 'DBSCAN', 'Mean-Shift', 'Agglomerative
Clustering', 'Gaussian Mixture'],
    description='Method:',
    disabled=False,
)
```

7. Виконаємо кластеризацію та оцінку моделей кластеризації.

```
def build_clusters(b):
    with output:
        output.clear_output()
        if not selected_features:
            print("Please add features to cluster.")
            return
        # Попередня обробка та нормалізація даних
        scaler = StandardScaler()
        scaled_data = scaler.fit_transform(data[selected_features])

        n_clusters = n_clusters_widget.value
        method = method_widget.value

        if method == 'K-Means':
            model = KMeans(n_clusters=n_clusters)
        elif method == 'DBSCAN':
            model = DBSCAN(eps=eps_widget.value,
min_samples=min_samples_widget.value)
        elif method == 'Mean-Shift':
            model = MeanShift()
        elif method == 'Agglomerative Clustering':
            model = AgglomerativeClustering(n_clusters=n_clusters)
        elif method == 'Gaussian Mixture':
            model = GaussianMixture(n_components=n_clusters)
        model.fit(scaled_data)

        labels = model.labels_ if hasattr(model, 'labels_') else
model.predict(scaled_data)

        # Візуалізація
        plt.figure(figsize=(10, 6))
        sns.scatterplot(x=scaled_data[:, 0], y=scaled_data[:, 1],
hue=labels, palette='viridis')
        plt.title(f'Clustering Result with {method}')
        plt.show()

        silhouette = silhouette_score(scaled_data, labels)
        print(f'Silhouette Score: {silhouette:.2f}')
```

```
build_button.on_click(build_clusters)

# Відображення віджетів
feature_selector_widgets = widgets.VBox([feature_dropdown, add_button,
remove_button, feature_list_widget])
settings_widgets = widgets.VBox([method_widget, eps_widget,
min_samples_widget, n_clusters_widget, build_button])
display(widgets.HBox([feature_selector_widgets, settings_widgets]),
output)
```

Рисунок 6.6. Результати моделі K-Means

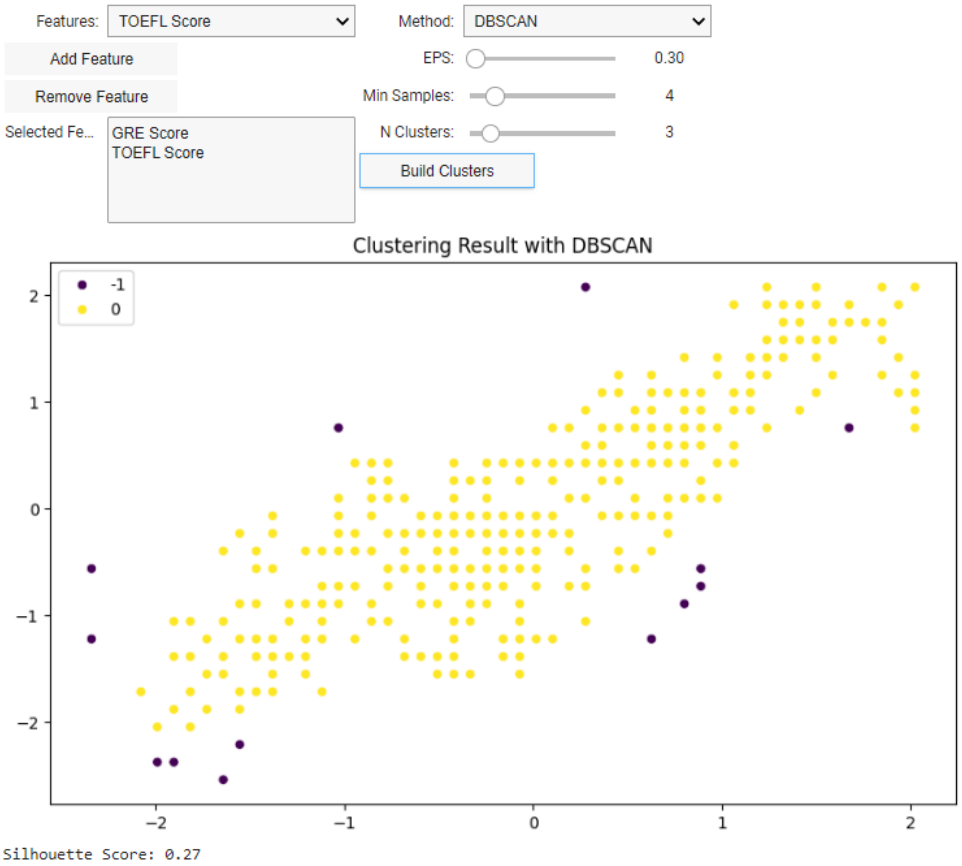


Рисунок 6.7. Результати моделі DBSCAN

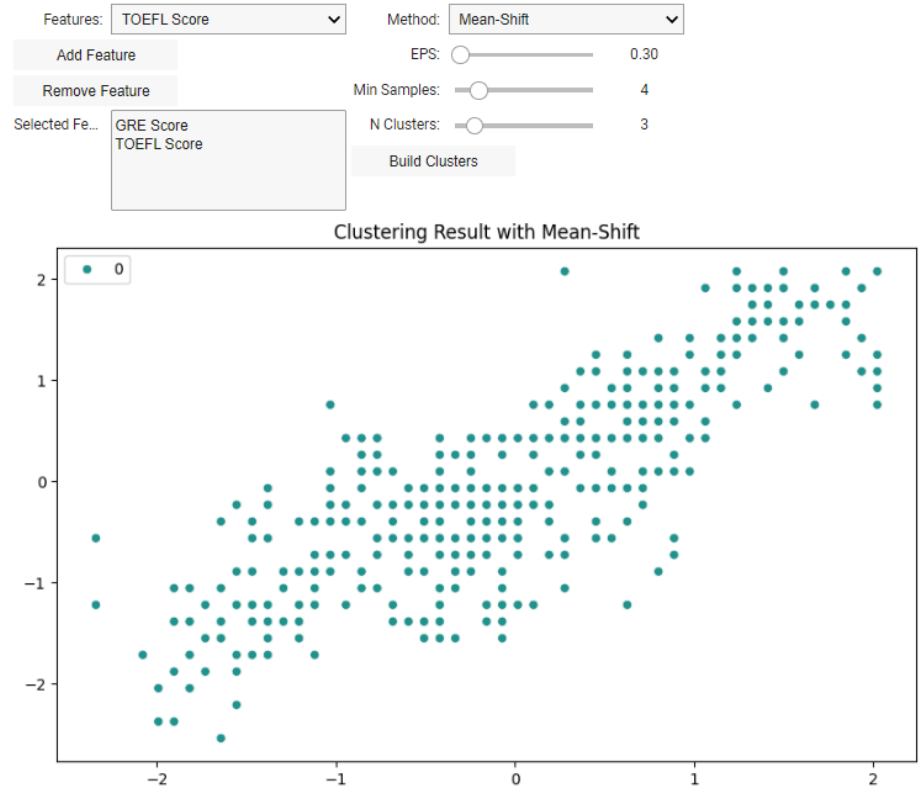


Рисунок 6.8. Результати моделі Mean-Shift

Рисунок 6.9. Результати моделі Agglomerative Clustering

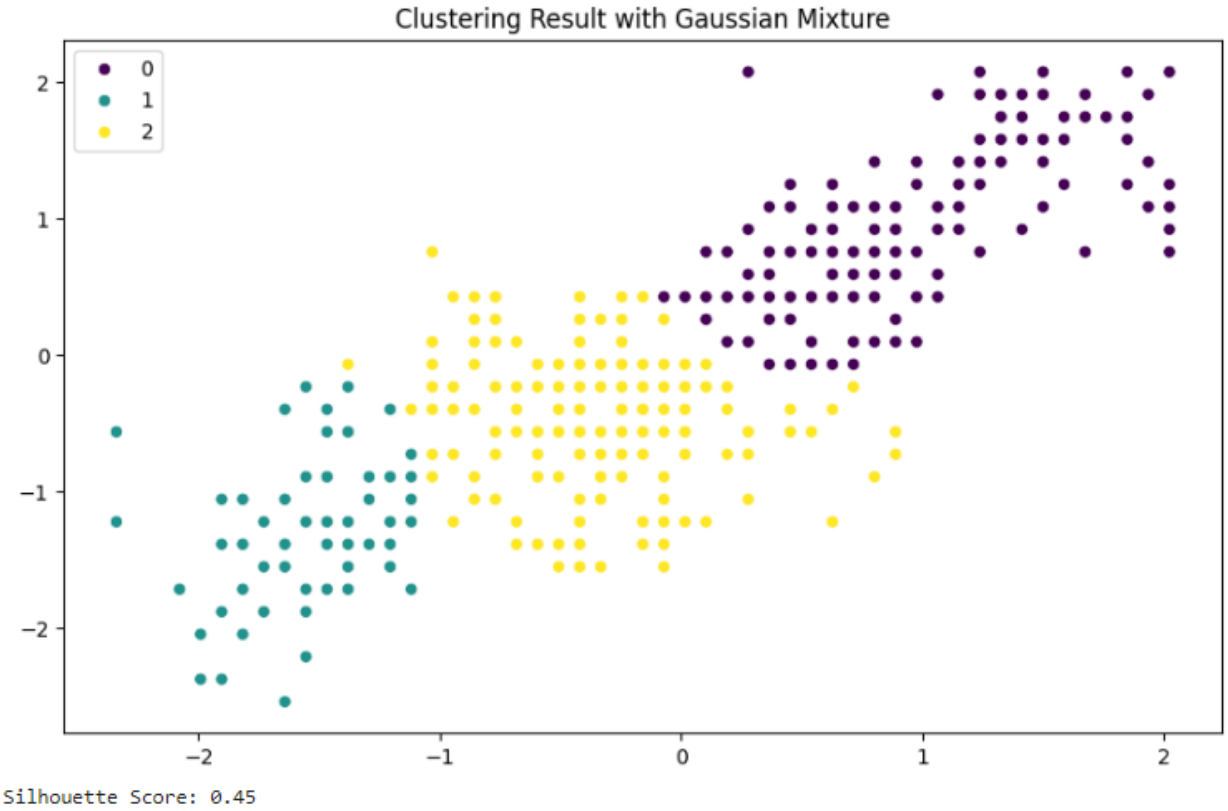


Рисунок 6.10. Результати моделі Gaussian Mixture

8. Виконаємо тестування гіперпараметрів

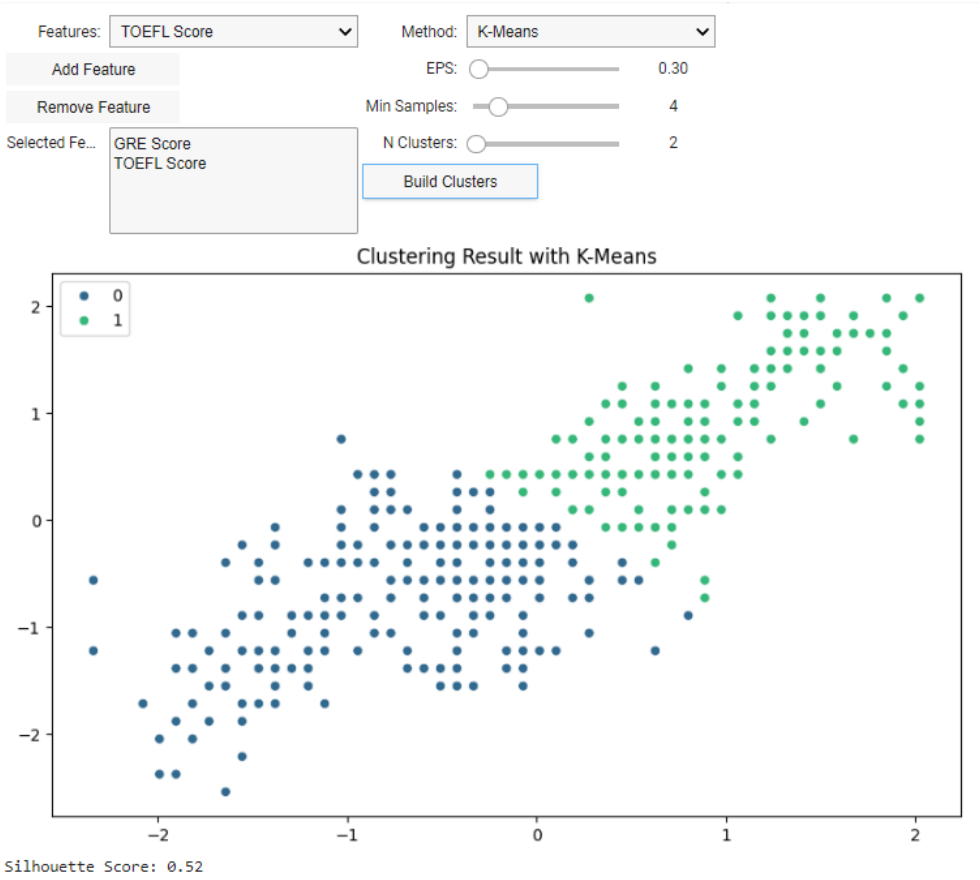


Рисунок 6.11. Зменшення кількості кластерів до «2» у моделі K-Means

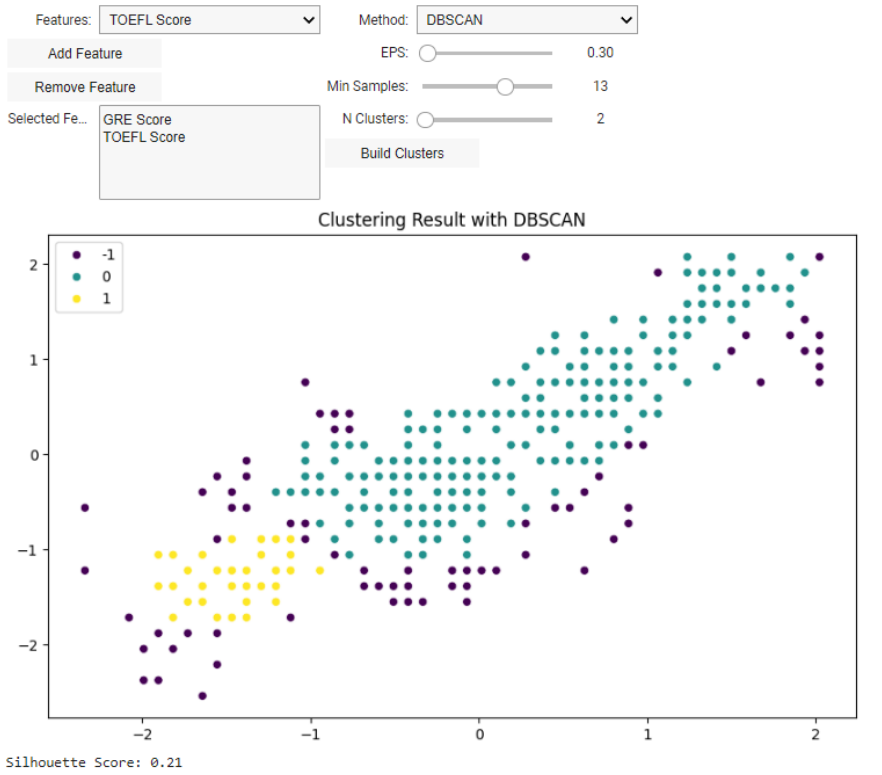


Рисунок 6.12. Збільшення гіперпараметру Min Samples до «13» у моделі DBSCAN

Таким чином, аналізуючи отримані дані, можна зазначити, що використання моделі Gaussian Mixture продемонструвало найкращі результати за показником Silhouette (0.45) за умови поділу на 3 кластери. Моделі Mean-Shift не вистачило даних або ознак, тому вона не надала коректних результатів. Модель K-Means (0.44) та Agglomerative Clustering (0.39) також продемонстрували непогані результати Silhouette. Що стосується моделі DBSCAN (0.07) – її показник Silhouette є незадовільними та не дозволяє якісно виконувати кластеризацію на заданому наборі даних, потребується доопрацювання. За рахунок маніпуляції з гіперпараметрами EPS та Min Samples вдалося підвищити показник Silhouette до 0.21.

Контрольне запитання №1

Вкажіть результуюче значення Silhouette для K-Means.
Відповідь: **0.44**

Контрольне запитання №2

Вкажіть результуюче значення Silhouette для Agglomerative Clustering.
Відповідь: **0.39**

Контрольне запитання №3

Вкажіть результуюче значення Silhouette для Gaussian Mixture.
Відповідь: **0.45**

Висновки: …

Завдання для виконання лабораторної роботи

1. Оберіть та завантажте датасет відповідно до свого варіанту. Завантажте ваш набір даних як тимчасовий файл у середовище Google Colab та ініціалізуйте його.
2. Проаналізуйте датасет та визначте ознаки, які будуть використовуватися для кластеризації. Як цільову зміну кластеризації за замовчуванням визначено останній стовпець у наборах даних.
3. Виконайте очищення даних: видаліть або заповніть пропущені значення, опрацюйте викиди якщо це потребується.
4. Виконайте тренування моделей: 1) K-Means; 2) DBSCAN; 3) Mean-Shift; 4) Agglomerative Clustering; 5) Gaussian Mixture. Виконайте оцінку та візуалізуйте результати за наступними метриками: Silhouette для кожної моделі.
5. Проаналізуйте отримані результати, обґрунтуйте вибір найкращої моделі для даного набору даних.
6. За бажанням проекспериментуйте з гіперпараметрами для моделей задля покращення результатів.

	Посилання на датасети				
№	**1**	**5**	**9**	**13**	**17**
Clusterization	Customer_Segment	Chance of Admit	Shape	category	profit
№	**2**	**6**	**10**	**14**	**18**
Clusterization	Profit	Sales	diabetes	R	Population
№	**3**	**7**	**11**	**15**	**19**
Clusterization	epsilon	Species	CO2	fiber_ID	class
№	**4**	**8**	**12**	**16**	**20**
Clusterization	median_house_value	Profit	quality	Median_Family_Income	Total

Контрольне запитання №1

Вкажіть результуюче значення Silhouette для K-Means.
Відповідь: _____

Контрольне запитання №2

Вкажіть результуюче значення Silhouette для Agglomerative Clustering.
Відповідь: _____

Контрольне запитання №3

Вкажіть результуюче значення Silhouette для Gaussian Mixture.

Відповідь: _____

Welcome to Google Colab

https://colab.research.google.com/drive/131_zrC7dhzn9LpNjBzhRZ_e51pgKpcdi?usp=sharing