Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Факультет інформатики та обчислювальної техніки

Кафедра інформаційних систем та технологій

Лабораторна робота №2

з дисципліни «Програмування інтелектуальних інформаційних систем»

Виконав:

студент групи ІП-11

Лисенко Андрій

Київ, 2023

1. Dataset1

Завдання

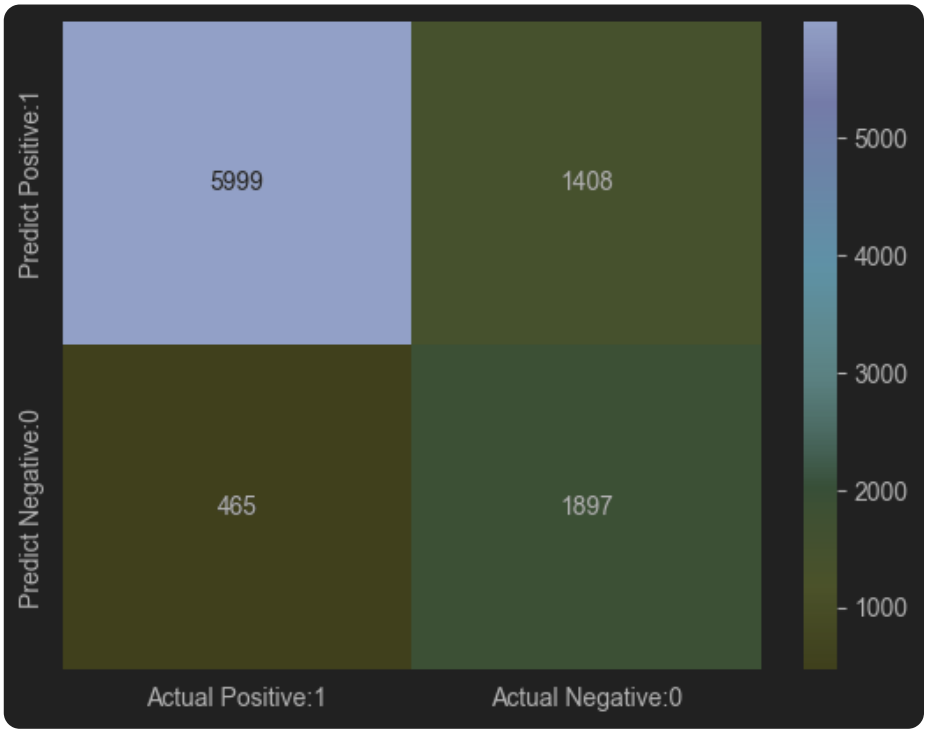
Bayesian Classification + Support Vector Machine

Зробити предікшн двома вищезгаданими алгоритмами. Порівняти наступні метрики:

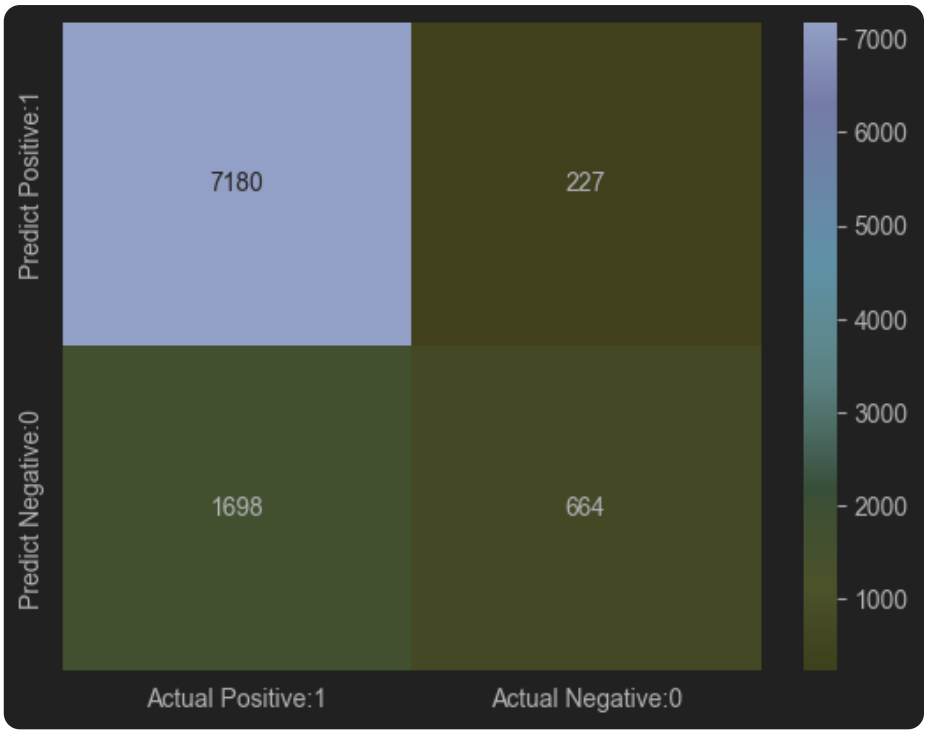
Recall, f1-score, Confusion matrix, accuracy score. Порівняти з нуль-гіпотезою і перевірити на оверфітинг. Пояснити результати.

Метрики

Metric	Bayesian Classifier	Support Vector Machine
Model accuracy	0.8083	0.8029
Training accuracy	0.8067	0.8067
Null accuracy	0.7582	0.7582
Recall	0.81	0.80
F1-score	0.82	0.77



Confusion Matrix - Bayesian Classifier



Confusion Matrix - Support Vector Machine

Висновок

1. **Model Accuracy**:
   За метрикою точності, обидва класифікатори показують схожу результативність, дуже близьку до 80%. Це означає, що обидва моделі правильно класифікують приблизно 80% випадків.

2. **Training Accuracy**:
   Обидва класифікатори мають високу точність на тестових даних, що близька до точності на навчальних даних, що свідчить про їхню здатність генералізувати на нові дані.

3. **Null Accuracy**:
   Обидва класифікатори демонструють трохи кращий результат, але цей показник все ще може бути покращений.

4. **Recall**:
   Обидва класифікатори мають схожий рівень Recall, що означає, що вони гарно реагують на виявлення позитивних випадків.

5. **F1-Score**:
   Bayesian Classifier має трохи кращий F1-Score, що вказує на більшу збалансованість між точністю і Recall, порівняно з Support Vector Machine.

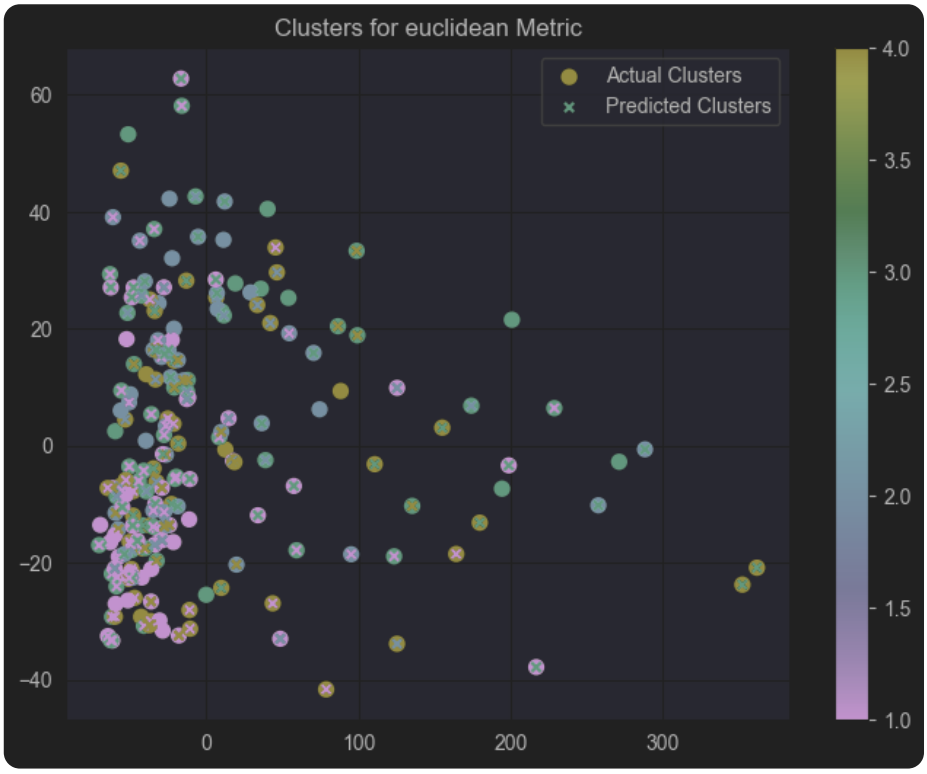
2. Dataset2

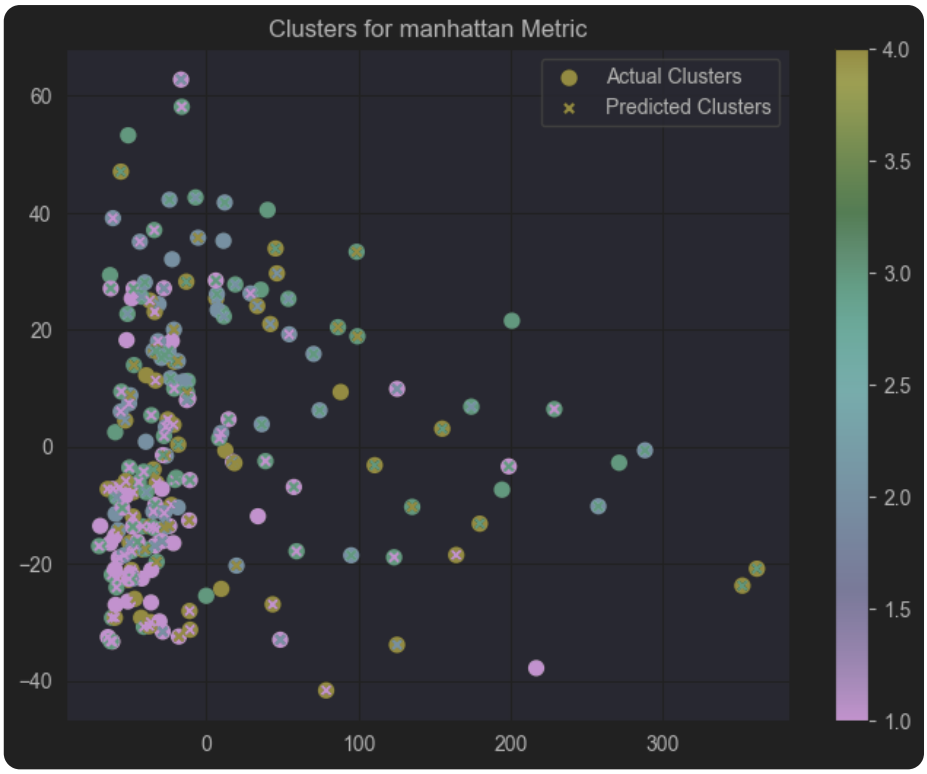
Завдання

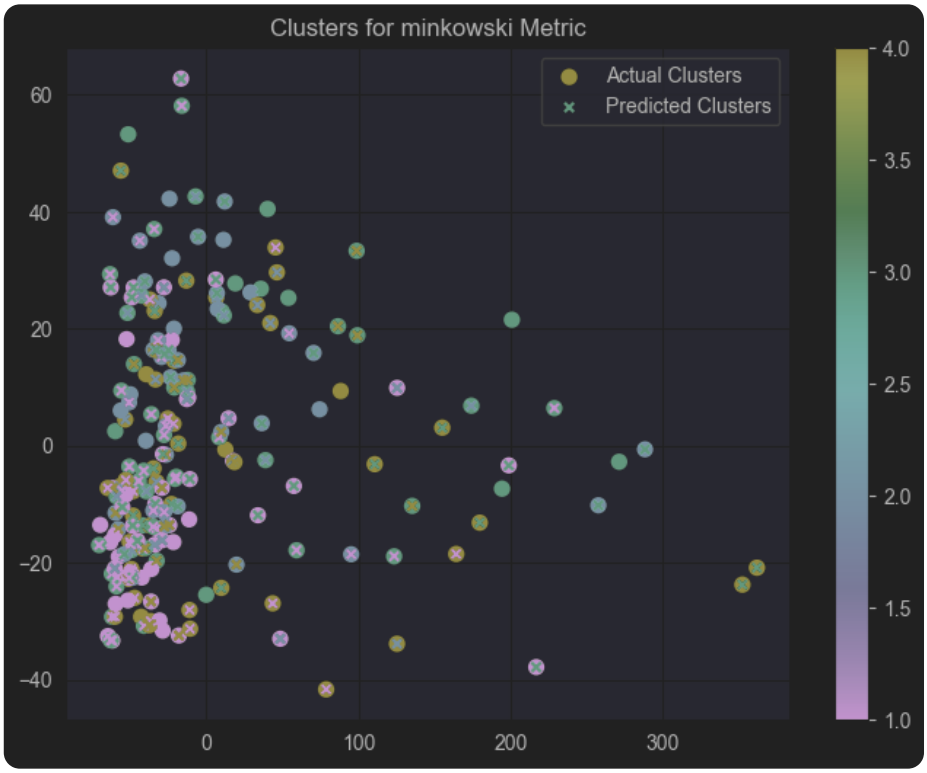
K nearest neighbours.

Те саме що і в 1 завданні, але порівнюємо між собою метрики. Euclidean, Manhattan, Minkowski. Кластери потрібно візуалізувати. Метрики аналогічно п.1

Кластери



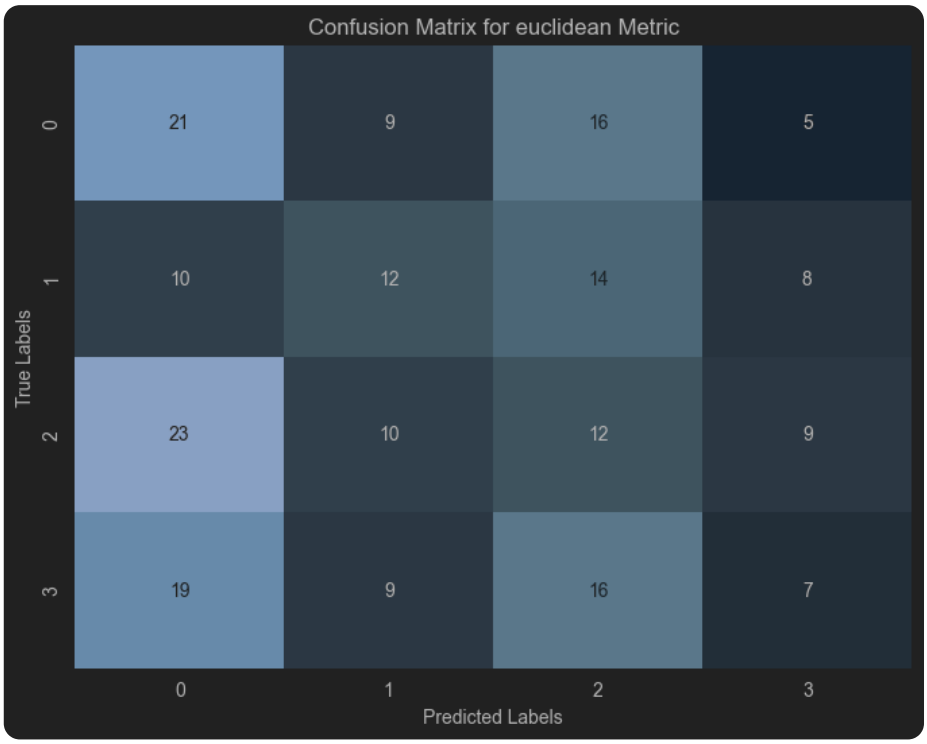


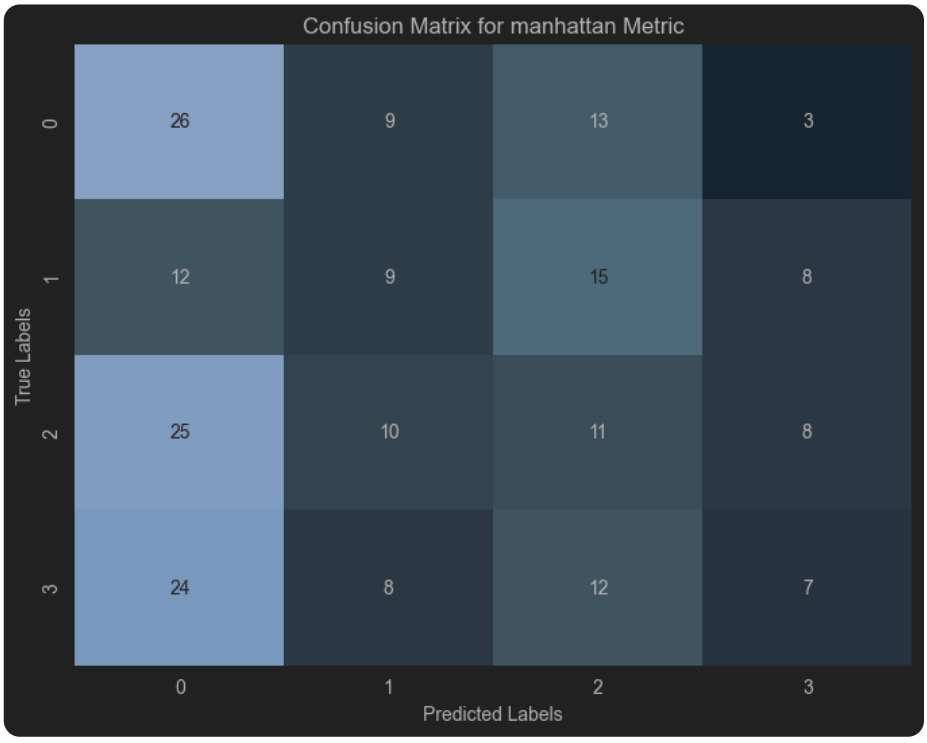


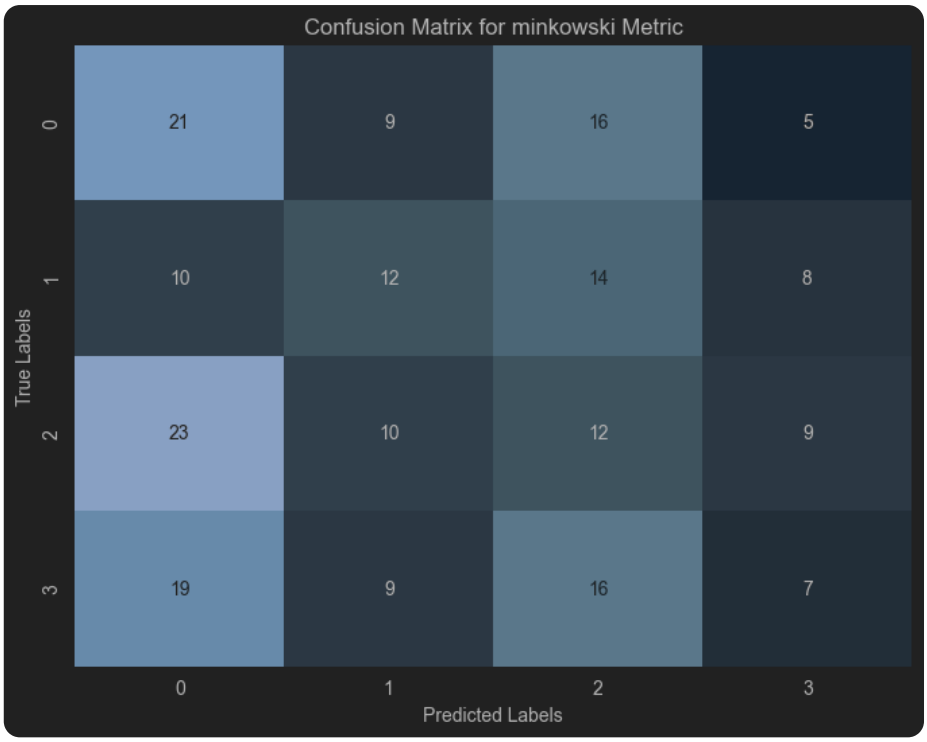
Метрики

Metric	Euclidian	Manhattan	minkowski
Model accuracy	0.26	0.265	0.26
Training accuracy	0.5575	0.576	0.5575

Metric	Euclidian	Manhattan	minkowski
Silhouette	-0.08	-0.046	-0.084
ARI	0.0001	0.004	0.0001
NMI	0.014	0.018	0.014







Висновок

1. **Model Accuracy**:
   За метрикою точності, усі три метрики показують схожі результати.

2. **Training Accuracy**:
   Помітно, що точність на навчальних даних вища, ніж на тестових даних, для всіх трьох метрик. Це свідчить про перенавчання (overfitting), оскільки модель демонструє кращу результативність на даних, які вона використовувала для навчання.

3. **Silhouette Score**:
   Усі три метрики мають від'ємні значення Silhouette Score, що свідчить про те, що об'єкти можуть бути погано кластеризовані, або що вони можуть бути віднесені до неправильних кластерів.

4. **Adjusted Rand Index (ARI)**:
   Manhattan має найкращий результат серед трьох метрик, і це може свідчити про кращу якість кластеризації в порівнянні з іншими.

5. **Normalized Mutual Information (NMI)**:
   NMI також оцінює схожість між фактичними та передбаченими кластерами. Знову ж таки, Manhattan показує найкращий результат серед трьох метрик.

Загалом, на основі цих метрик, Manhattan метрика схожа на найкращу серед розглянутих. Однак, від'ємні значення Silhouette Score вказують на проблему з якістю кластеризації.

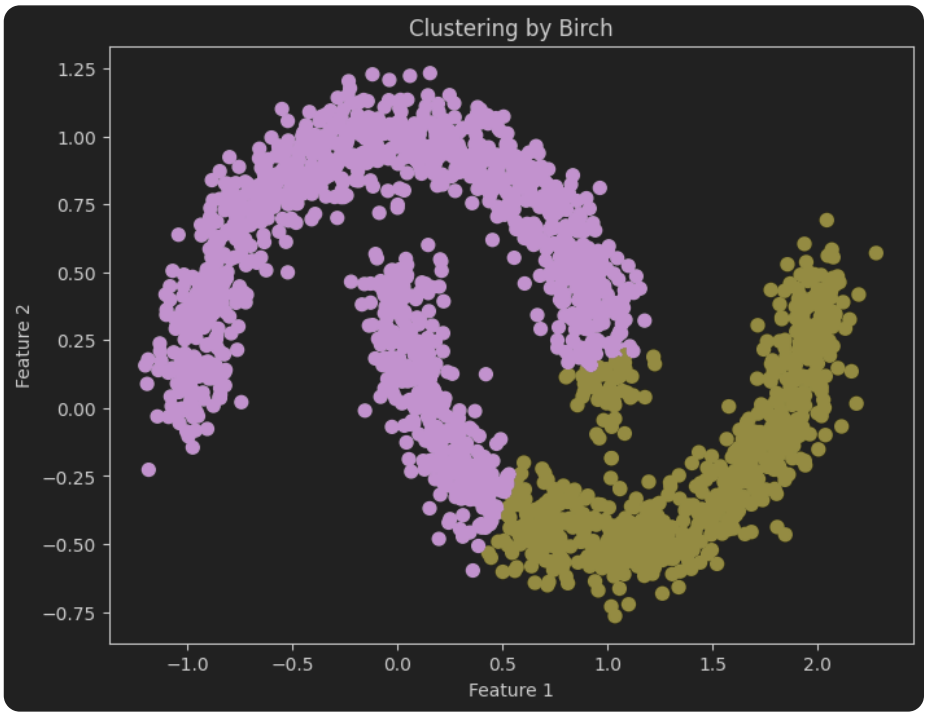
3. Dataset3

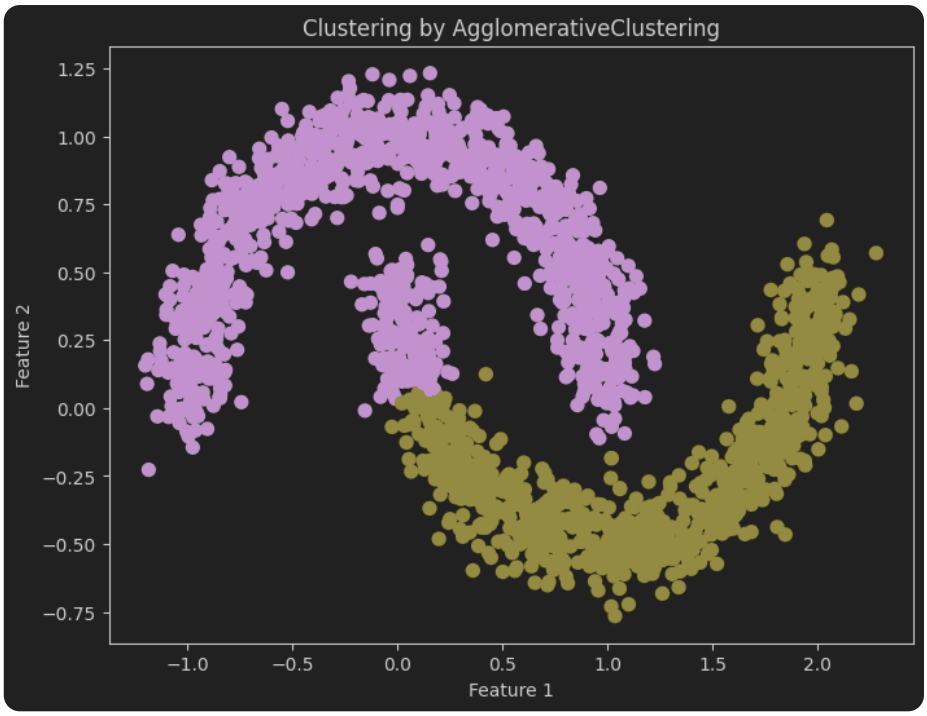
Завдання

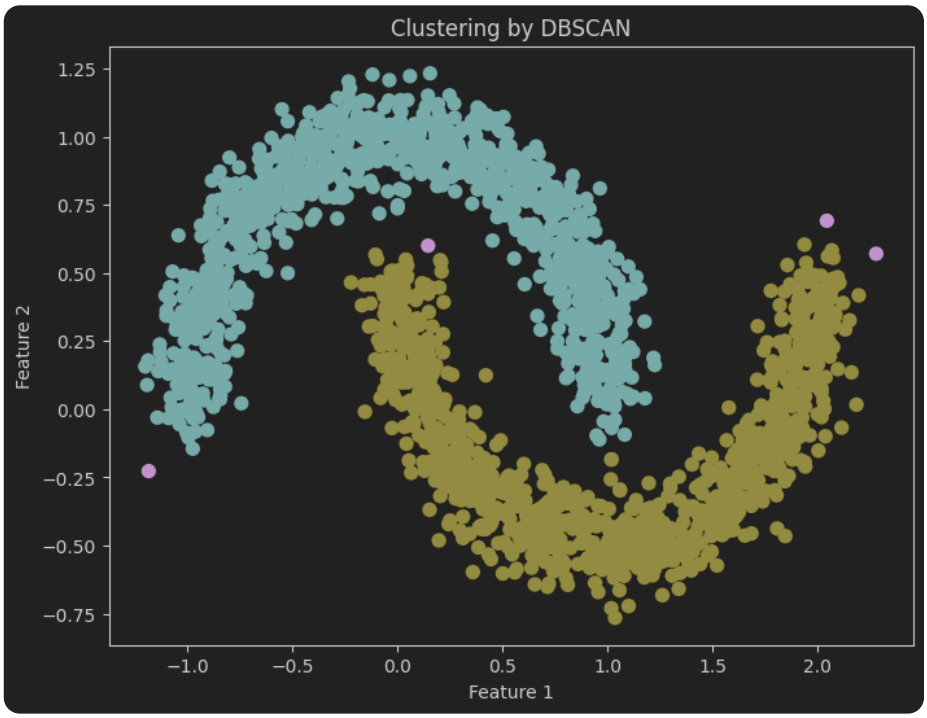
Agnes,Birch,DBSCAN

Інші методи можна ігнорувати. Зняти метрики (Silhouette Coefficient, ARI, NMI. Можна з п.1-2), пояснити.

Кластери







Метрики

Metric	Birch	AgglomerativeClustering	DBSCAN
Silhouette	0.458	0.406	0.301
ARI	0.377	0.716	0.992
NMI	0.341	0.671	0.978

Висновок

1. **Silhouette Score**:
   За метрикою Silhouette Score, Birch має найкращий результат, що вказує на гарну якість кластеризації. DBSCAN має найнижчий Silhouette Score, що може вказувати на менш якісну кластеризацію.

2. **Adjusted Rand Index (ARI)**:
   За метрикою ARI, DBSCAN має найкращий результат, що вказує на високу схожість між фактичними та передбаченими кластерами. AgglomerativeClustering також має досить високий ARI, в той час як Birch показує менший результат.

3. **Normalized Mutual Information (NMI)**:
   За метрикою NMI, DBSCAN має найкращий результат, вказуючи на високу схожість між фактичними та передбаченими кластерами. AgglomerativeClustering також показує досить високий NMI, в той час як Birch має менший результат.

Загалом, різні метрики демонструють різні аспекти якості кластеризації. DBSCAN має дуже високий ARI та NMI, що свідчить про високу якість кластеризації, особливо щодо схожості між фактичними та передбаченими кластерами. Birch має досить низький результат за всіма трьома метриками, вказуючи на можливу менш гарну якість кластеризації. AgglomerativeClustering займає проміжне положення за всіма метриками.

Зважаючи на ці результати, DBSCAN може бути найкращим вибором для завдання кластеризації, особливо якщо важлива висока схожість між фактичними та передбаченими кластерами.

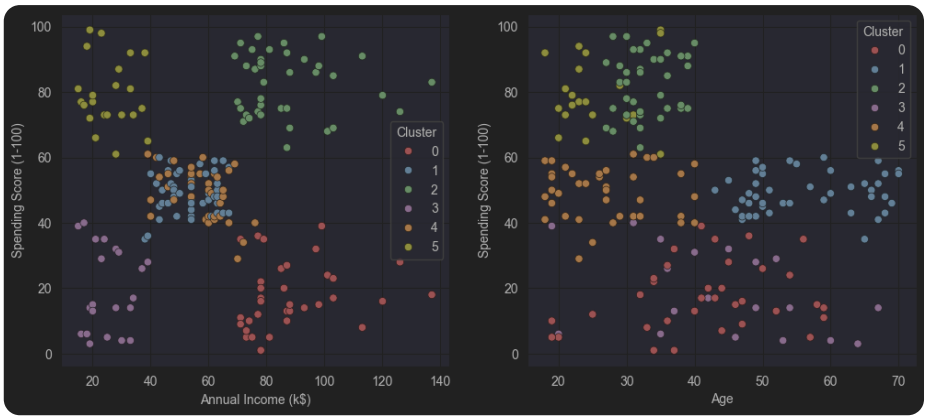
4. Dataset4

Завдання

Affinity propagation.

Порівняти з k-means. Метрики - Silhouette Coefficient, ARI, NMI

Кластери



Метрики

Metric	KMeans	Affinity Propagation
Silhouette	0.452	0.451
ARI	0.365	0.368
NMI	0.579	0.586

Висновок

1. **Silhouette Score**:
   Обидва алгоритми мають подібні значення Silhouette Score, приблизно 0.452 для K-Means та 0.451 для Affinity Propagation.

2. **Adjusted Rand Index (ARI)**:
   Обидва алгоритми мають подібні значення ARI, близько 0.365 для K-Means та 0.368 для Affinity Propagation.

3. **Normalized Mutual Information (NMI)**:
   Обидва алгоритми мають схожі значення NMI, приблизно 0.579 для K-Means та 0.586 для Affinity Propagation.

За різними метриками, результати для обох алгоритмів виявляються досить подібними. Отже, з точки зору цих метрик, обидва алгоритми демонструють приблизно однакову якість кластеризації. Вибір між ними може залежати від конкретних особливостей даних, обсягу

обчислювальних ресурсів та інших факторів, які важливі для конкретного завдання кластеризації.