МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №6
по дисциплине «Машинное обучение»
Тема: Кластеризация (DBSCAN, OPTICS)

Студент гр. 6304 _____ Виноградов К.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург
2020

Загрузка данных

Датасет загружен в датафрейм.

DBSCAN

Проведена кластеризация методом DBSCAN при параметрах по умолчанию. Выведены метки кластеров, количество кластеров, а также процент наблюдений, которые кластеризовать не удалось, что показано на рис. 1. В табл. 1 представлены все параметры, которые принимает DBSCAN.

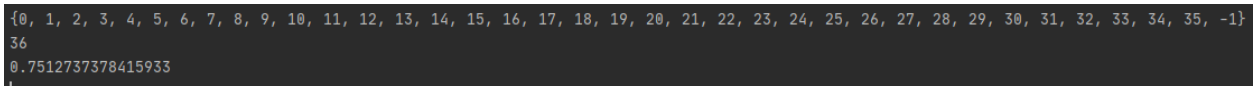

Рисунок 1 – Кластеризация DBSCAN при параметрах по умолчанию

Таблица 1 – Параметры DBSCAN

Параметр	Описание
eps: float, default=0.5	Максимальное расстояние между двумя наблюдениями, чтобы один считался соседним с другим (радиус окрестности наблюдения).
min_samples: int, default=5	Минимальное количество наблюдений в окрестности точки, чтобы считать ее базовой (включая саму точку).
metric: string or callable, default=’euclidean’	Метрика для вычисления расстояния между экземплярами в массиве признаков.
algorithm: {‘auto’, ‘ball_tree’, ‘kd_tree’, ‘brute’}, default = ‘auto’	Алгоритм, который будет использоваться для вычисления точечных расстояний и поиска ближайших соседей.

Построены график количества кластеров и процента не кластеризованных наблюдений в зависимости от максимальной

рассматриваемой дистанции (минимальное значение точек, образующих кластер, оставлено по умолчанию) и график количества кластеров и процента не кластеризованных наблюдений в зависимости от минимального значения количества точек, образующих кластер (максимальная рассматриваемая дистанция между наблюдениями оставлена по умолчанию). Графики представлены на рис. 2 и 3.

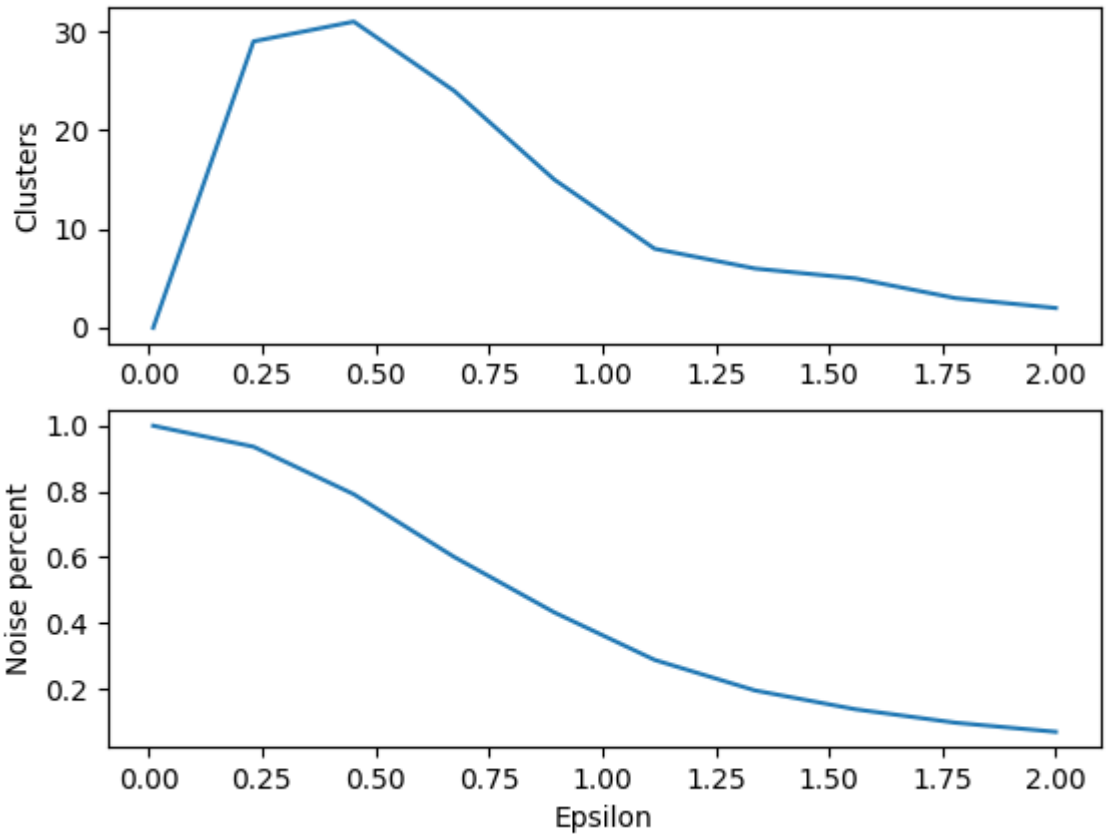

Рисунок 2 – Зависимости количества кластеров и процента шума от eps

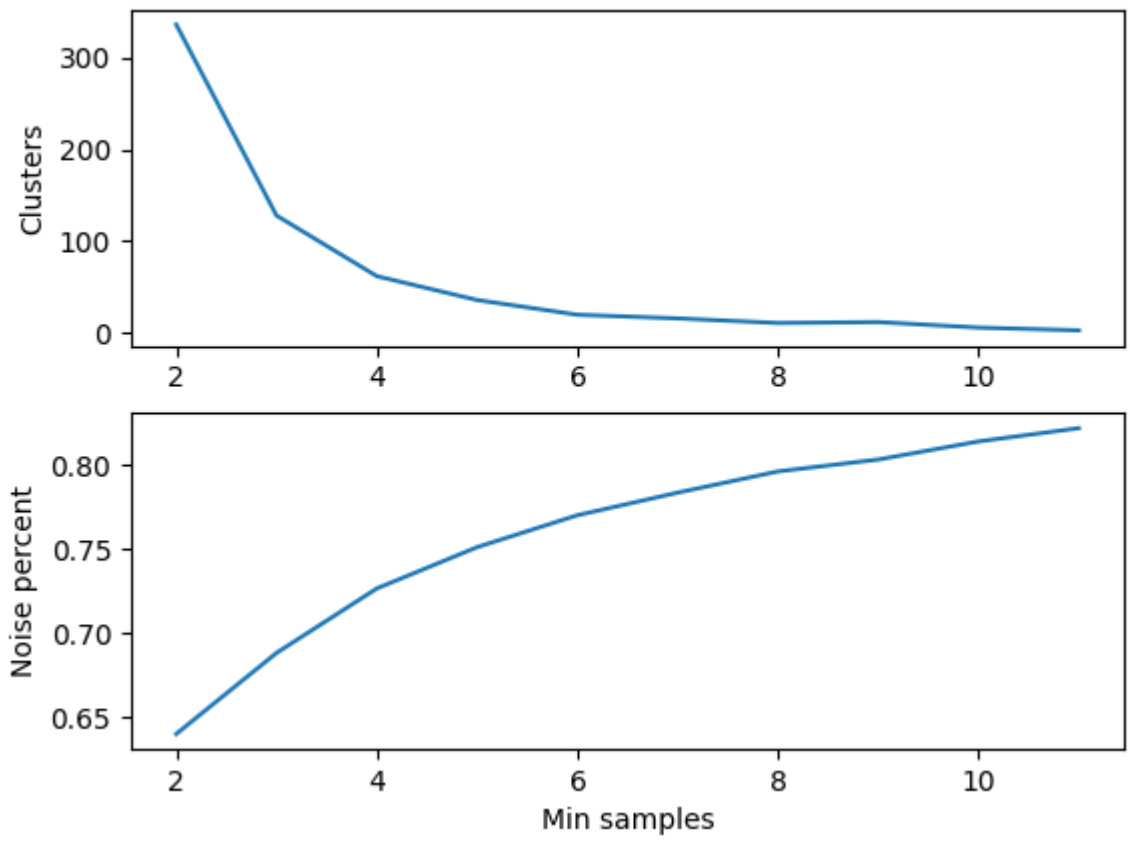

Рисунок 3 – Зависимости количества кластеров и процента шума от min_samples

Размерность данных понижена до 2 с помощью метода главных компонент. Результаты кластеризации данных пониженной размерности представлены на рис. 4.

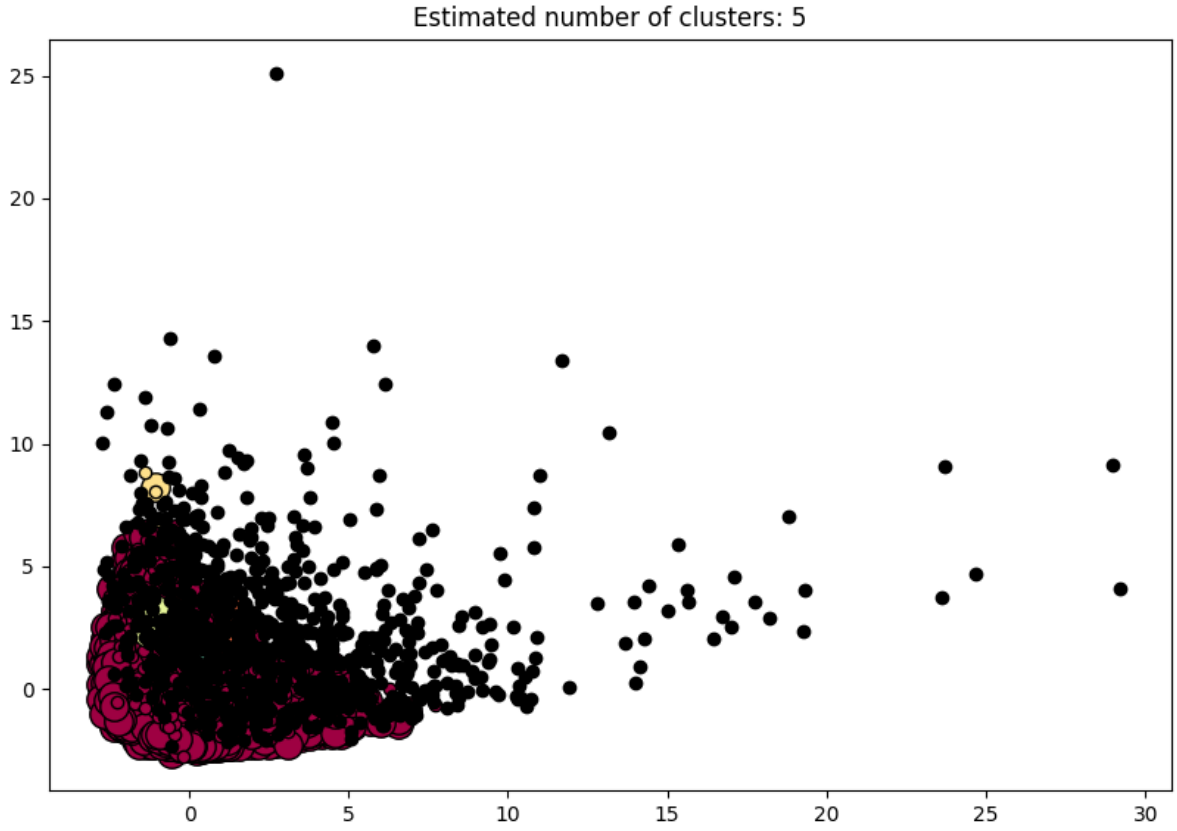

Рисунок 4 – Кластеризация данных пониженной размерности

OPTICS

В табл. 2 представлены все параметры, которые принимает OPTICS.

Таблица 2 – Параметры OPTICS

Параметр	Описание
max_eps: float, default=∞	Максимальное расстояние между двумя наблюдениями, чтобы один считался соседним с другим (радиус окрестности наблюдения).
min_samples: int>1 or float in [0, 1], default=5	Количество наблюдений в окрестности точки, чтобы считать ее базовой.

metric: string or callable, default=’minkowski’	Метрика для вычисления расстояния.
p: int, default = 2	Параметр для метрики Минковского.
cluster_method: string, default = ‘xi’	Метод извлечения кластеров. Также можно поставить 'dbscan'.
eps	Максимальное расстояние между двумя наблюдениями, чтобы один считался соседним с другим (радиус окрестности наблюдения). Нужен только при cluster_method=dbscan.
xi: float in [0,1], default=0.05	Определяет минимальную крутизну на графике достижимости, который составляет границу кластера.
predecessor_correction: bool, default=True	Коррекция кластеров в соответствии с предшественниками, рассчитанными OPTICS. Этот параметр оказывает минимальное влияние на большинство наборов данных. Используется только когда cluster_method = 'xi'.
min_cluster_size: int>1 or float in [0, 1], default=None	Минимальное количество выборок в кластере OPTICS, выраженное в виде абсолютного числа или доли от количества выборок (округленное до не менее 2). Если None, вместо этого используется значение min_samples. Используется только когда cluster_method = 'xi'.
algorithm: {‘auto’, ‘ball_tree’, ‘kd_tree’, ‘brute’}, default = ‘auto’	Алгоритм, который будет использоваться для вычисления точечных расстояний и поиска ближайших соседей.

Результаты кластеризации методом OPTICS близки к результатам DBSCAN при cluster_method=dbscan, max_eps=0.5, min_samples=5. Результат представлен на рис. 5.

```
{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, -1}
36
0.7550949513663733
```

Рисунок 5 – Результат кластеризации OPTICS

Процесс определения базовых точек в OPTICS идентичен DBSCAN, однако в OPTICS для точек вычисляются и сохраняются расстояния достижимости, на основе которых наблюдения выстраиваются в кластере, сохраняя при этом иерархическую структуру.

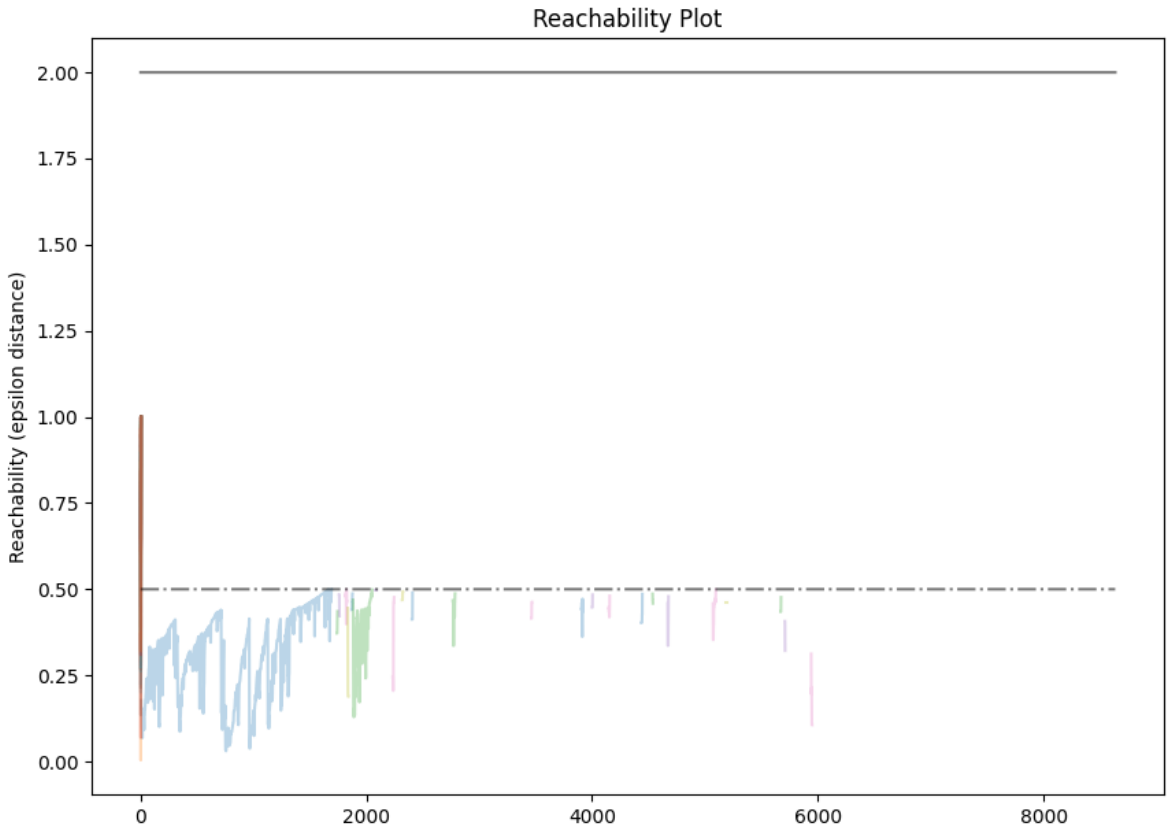

Рисунок 6 – График достижимости

Работа классификатора исследована при различных параметрах metric. Результаты представлены на рис. 7 – 11 и в табл. 3.

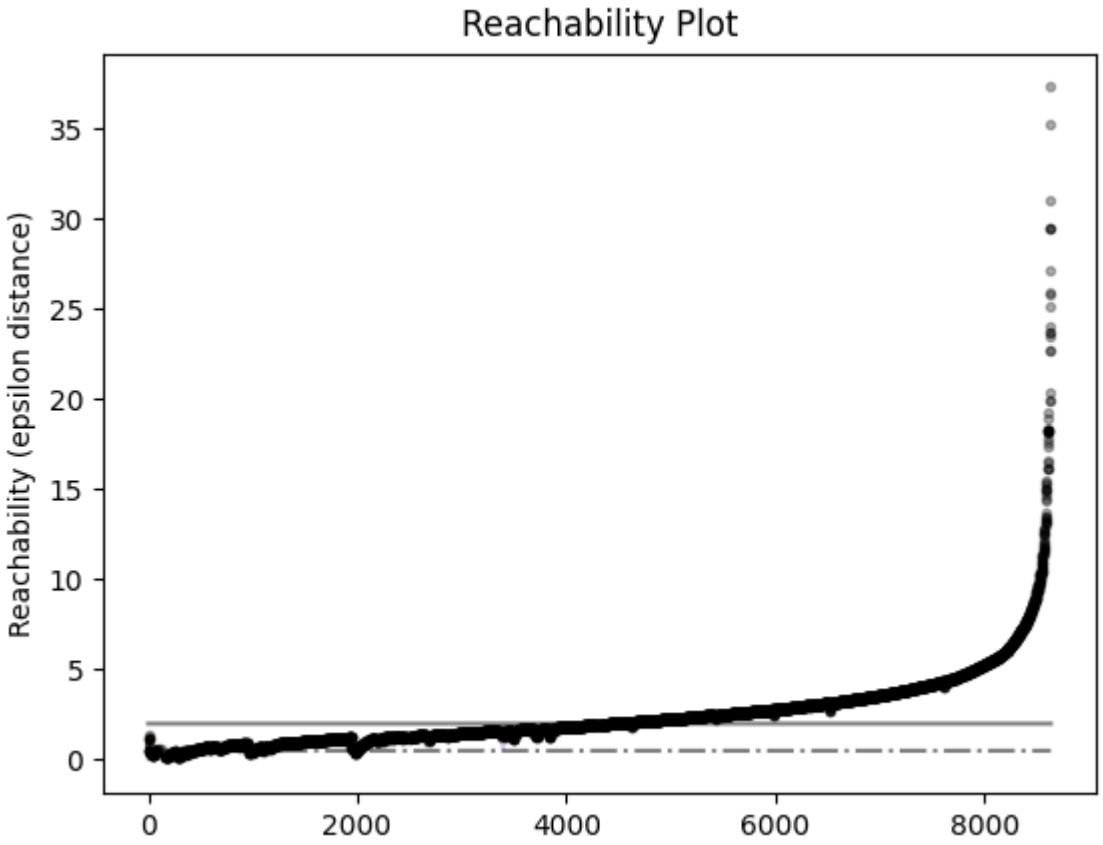

Рисунок 7 – График достижимости при метрике cityblock

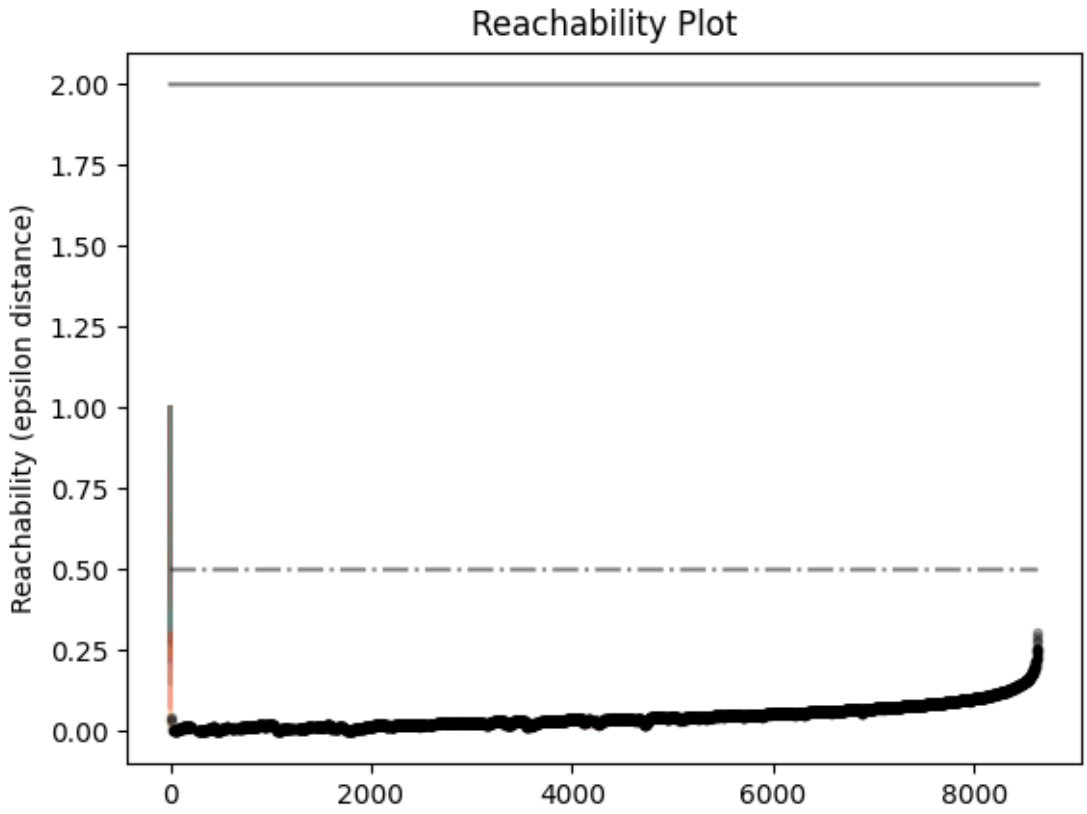

Рисунок 7 – График достижимости при метрике cosine

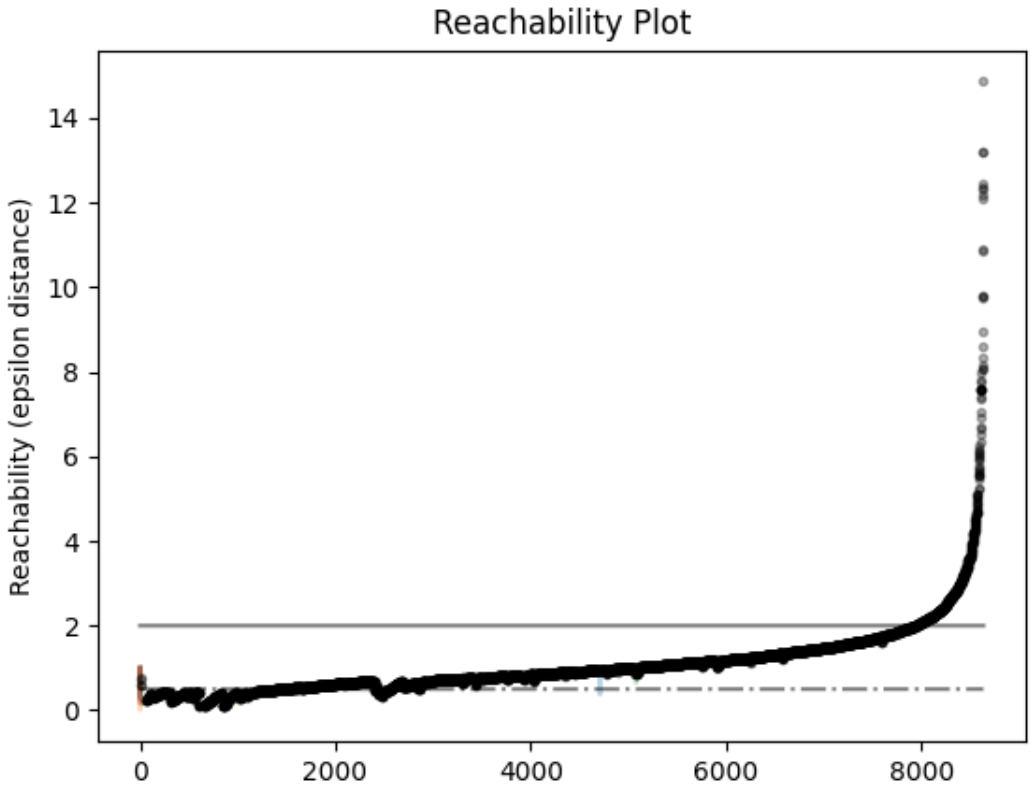

Рисунок 7 – График достижимости при метрике Euclidean

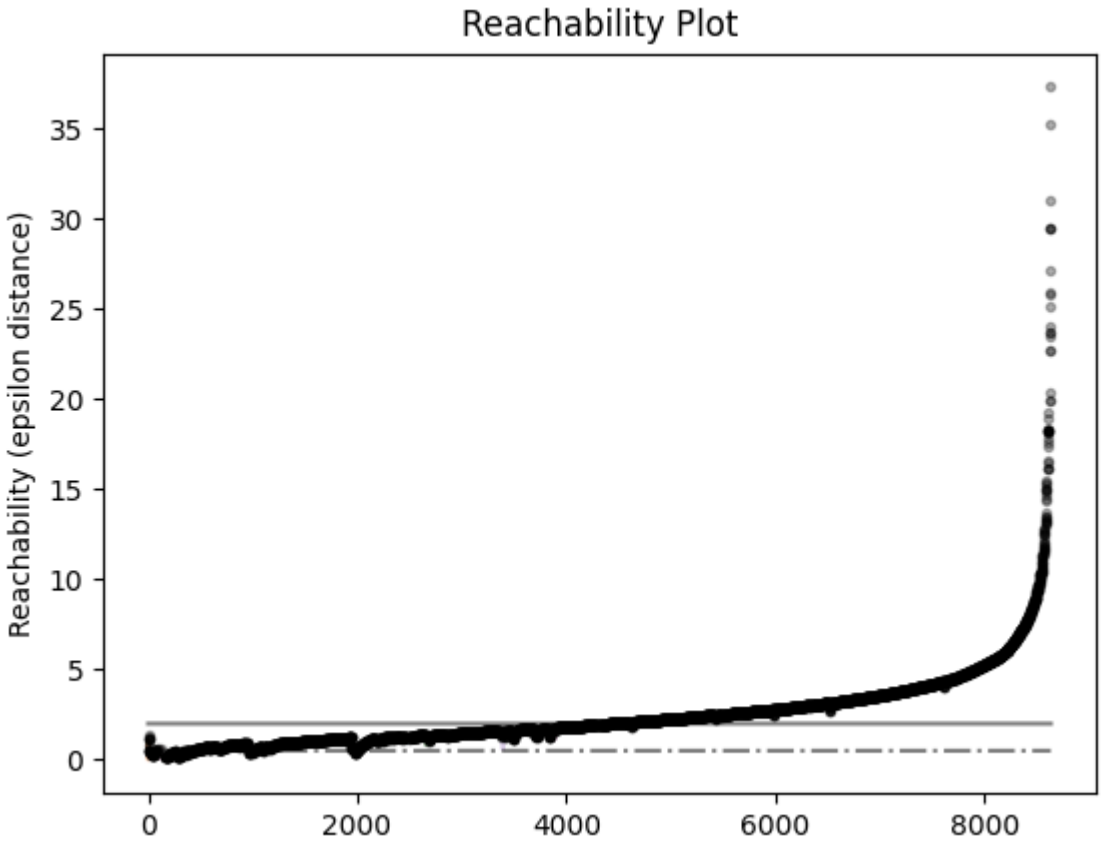

Рисунок 7 – График достижимости при метрике l1

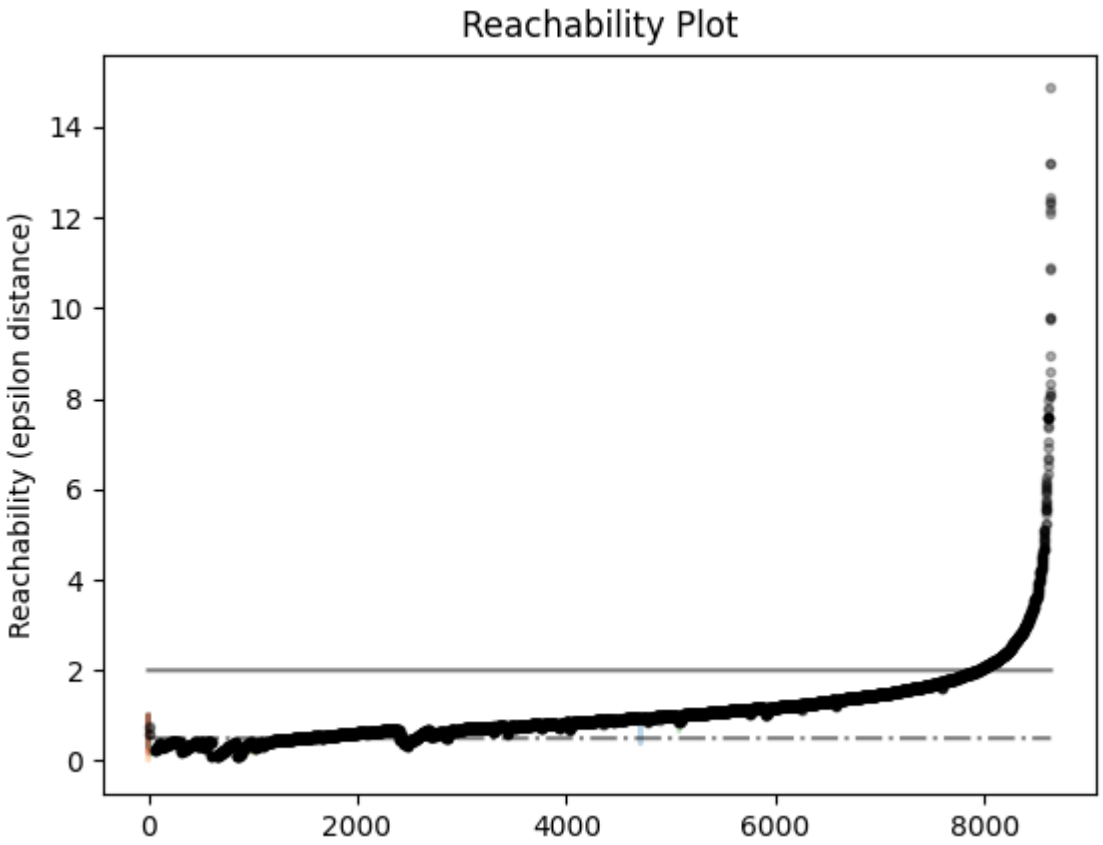

Рисунок 7 – График достижимости при метрике l2

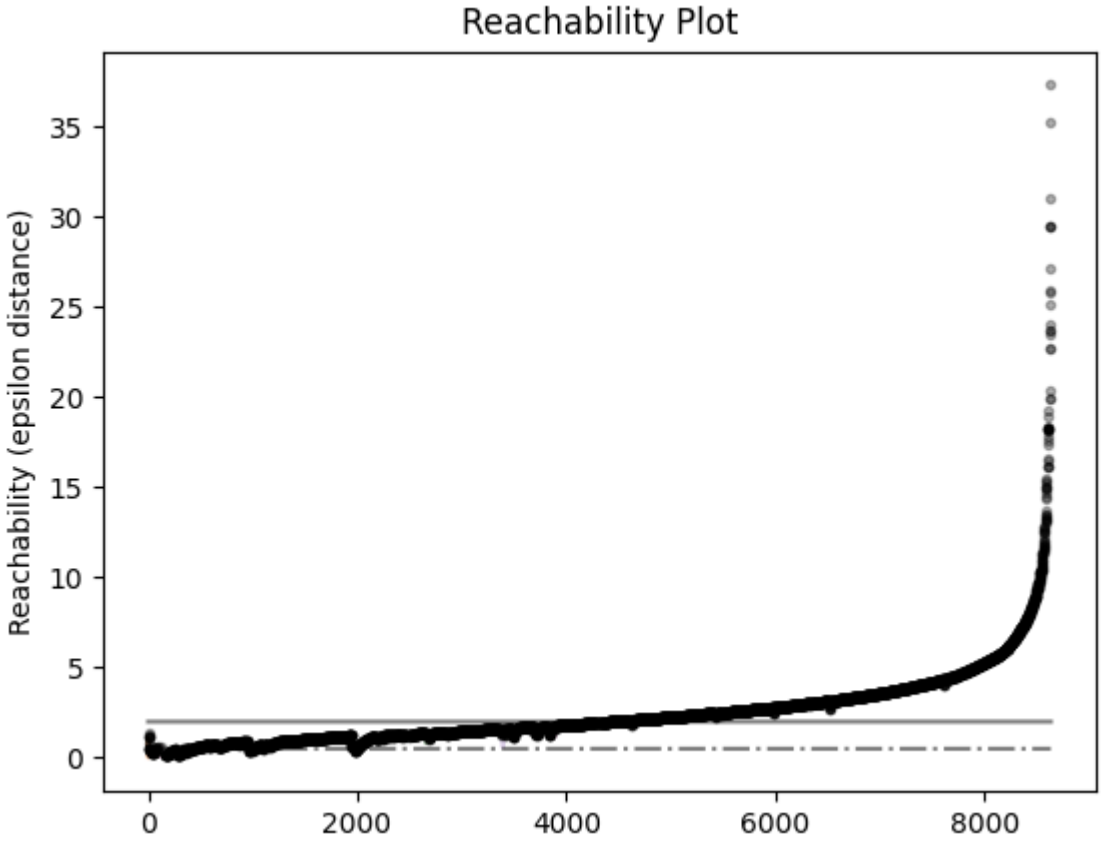

Рисунок 7 – График достижимости при метрике manhattan

Таблица 3 – Результаты OPTICS для различных метрик

Метрика	Количество кластеров	Процент выпавших наблюдений
cityblock	15	0.031
cosine	34	0.077
euclidean	17	0.03
l1	15	0.03
l2	17	0.037
manhattan	15	0.03

Выводы

В ходе лабораторной работы изучены такие методы кластеризации модуля Sklearn, как DBSCAN и OPTICS. При cluster_method=’xi’ OPTICS разделяет данные на большое число кластеров, малое количество кластеров достигается только при большом количестве выпавших наблюдений.