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. 6307 _____ Ходос А.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами кластеризации модуля Sklearn.

Ход работы

Был загружен датасет. Откинув наблюдения с пропущенными значениями и убрав столбец с метками были получены данные для работы, содержащие 8636 записей и 17 признаков. Некоторые записи и признаки представлены на рисунке 1.

	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES
0	40.900749	0.818182	95.40	0.00
1	3202.467416	0.909091	0.00	0.00
2	2495.148862	1.000000	773.17	773.17
4	817.714335	1.000000	16.00	16.00
5	1809.828751	1.000000	1333.28	0.00
...	...	...	...	...
8943	5.871712	0.500000	20.90	20.90
8945	28.493517	1.000000	291.12	0.00
8947	23.398673	0.833333	144.40	0.00
8948	13.457564	0.833333	0.00	0.00
8949	372.708075	0.666667	1093.25	1093.25

Рисунок 1 — загруженный датасет

После стандартизации данных, провели кластеризацию методом DBSCAN при параметрах по умолчанию. Были получены 36 кластеров, при этом 75% данных не удалось кластеризовать.

Для исследования были проведены кластеризации сначала варьируя максимальную дистанцию между наблюдениями, а затем минимальное значение количества точек, образующих кластер. Результаты представлены на рисунках 2-5.

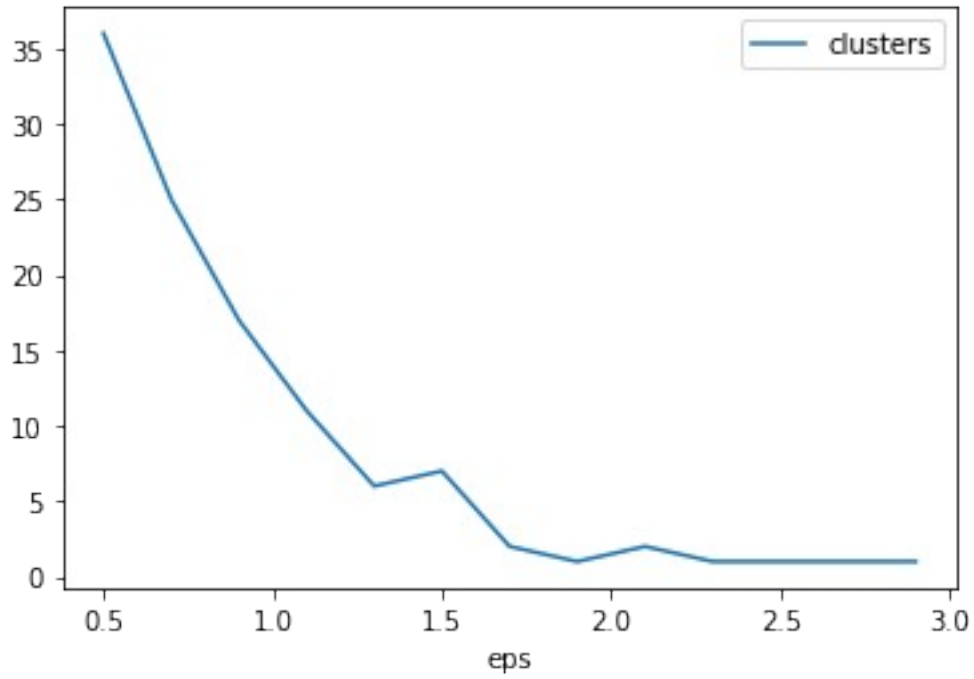

Рисунок 2 — График зависимости количества кластеров от максимальной дистанции между наблюдениями

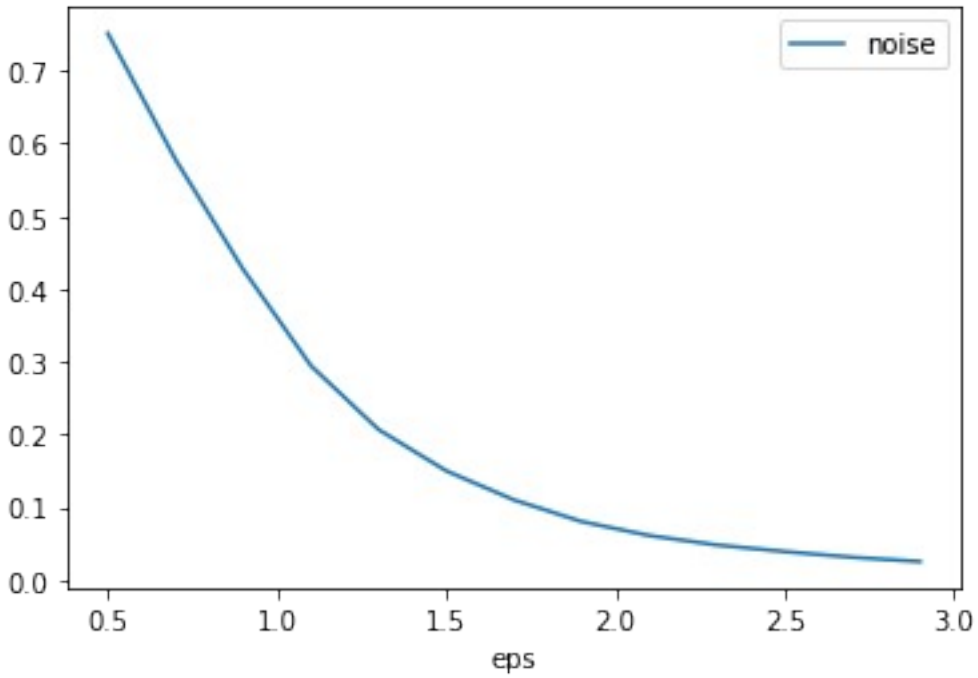

Рисунок 3 — График зависимости процента некластеризованных данных от максимальной дистанции между наблюдениями

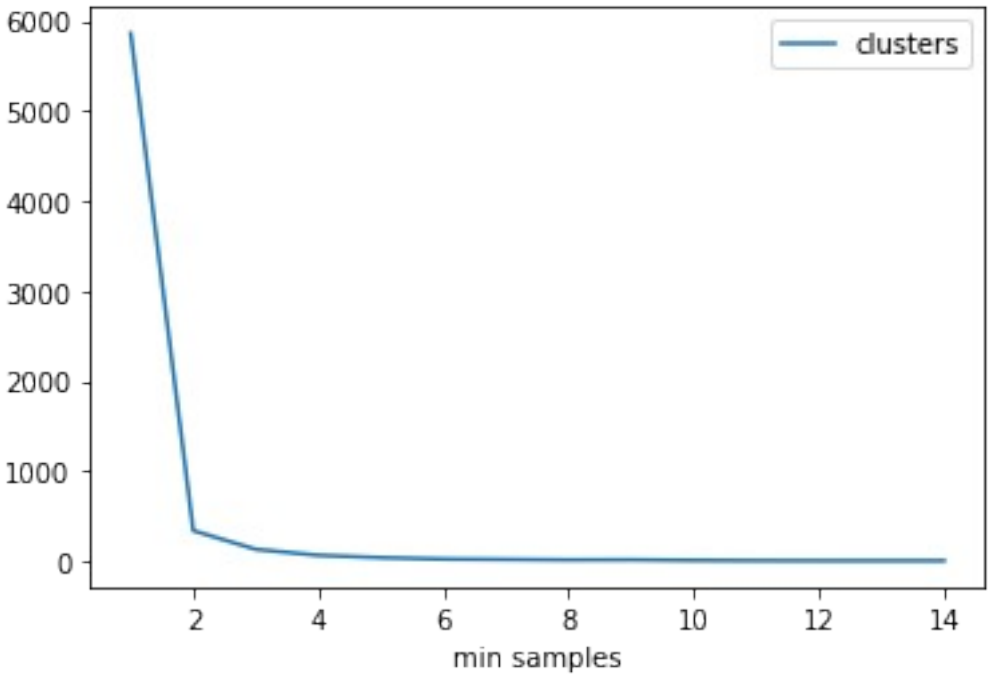

Рисунок 4 — График зависимости количества кластеров от минимального значение количества точек

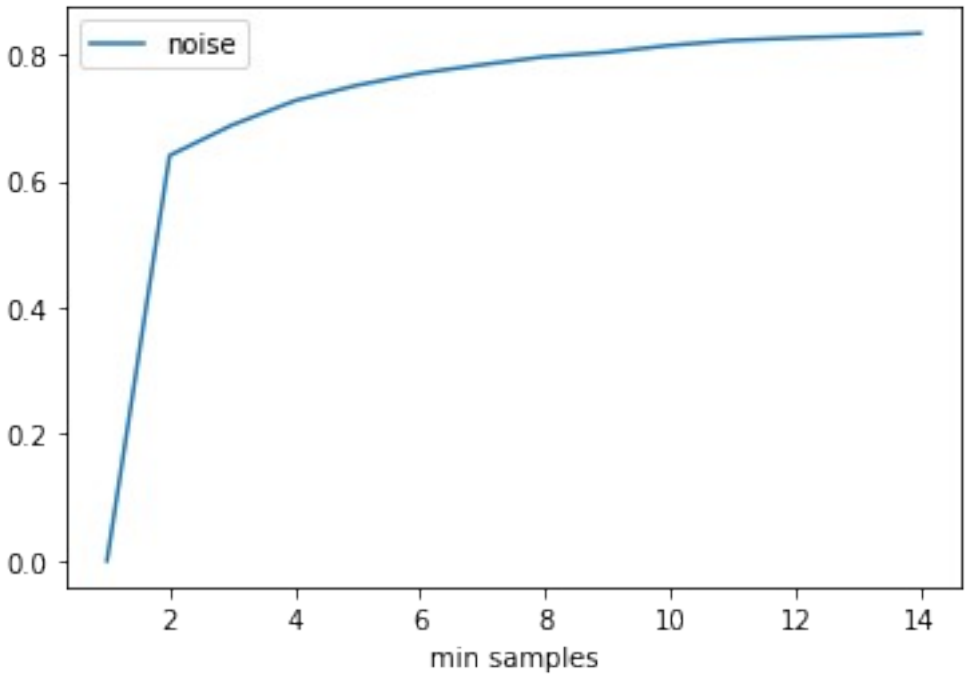

Рисунок 5 — График зависимости процента некластеризованных данных от минимального значение количества точек

Была проведена кластеризация, при которой оличество кластеров получается 5, и процент не кластеризованных наблюдений не превышает 12.

Получили минимальное количество точек равное 3, и максимальное расстояние равное 2.9.

Понизив размерность данных до 2 с использованием метода главных компонент, была проведена визуализация результатов кластеризации (метки получены до уменьшения размерности). Результаты представлены на рисунке 6.

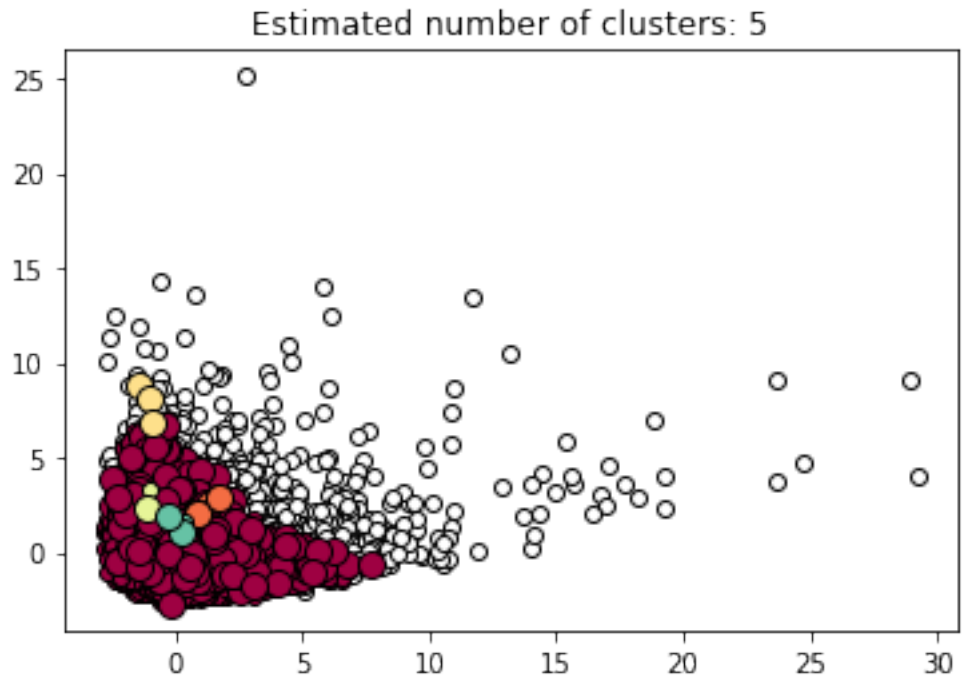

Рисунок 6 — визуализация результатов кластеризации DBSCAN

Был исследован метод кластеризации OPTICS. При тех же параметрах, как и DBSCAN, и параметре cluster_method „dbscan“ кластеризация дала результаты, близкие к кластеризации DBSCAN. Результаты представлены на рисунке 7.

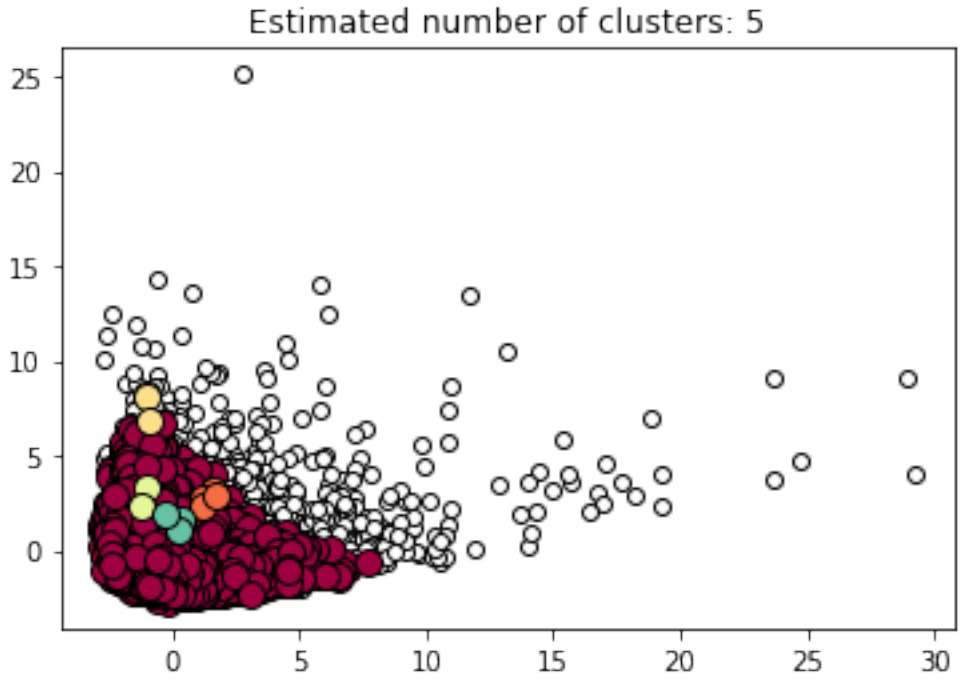

Рисунок 7 — визуализация результатов кластеризации OPTICS

Была исследована работа алгоритма при разных метриках, при этом для анализа были построены графики достижимости. Результаты представлены на рисунках 8-12.

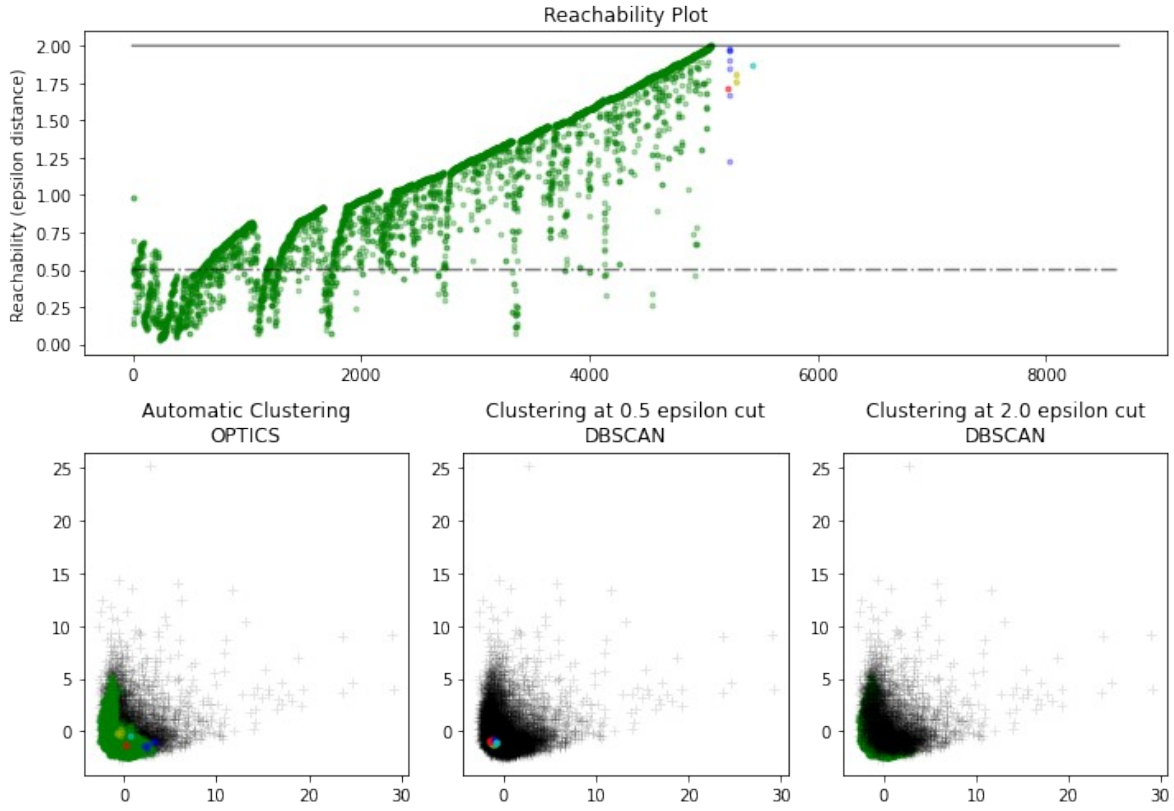

Рисунок 8 — кластеризация OPTICS с метрикой cityblock

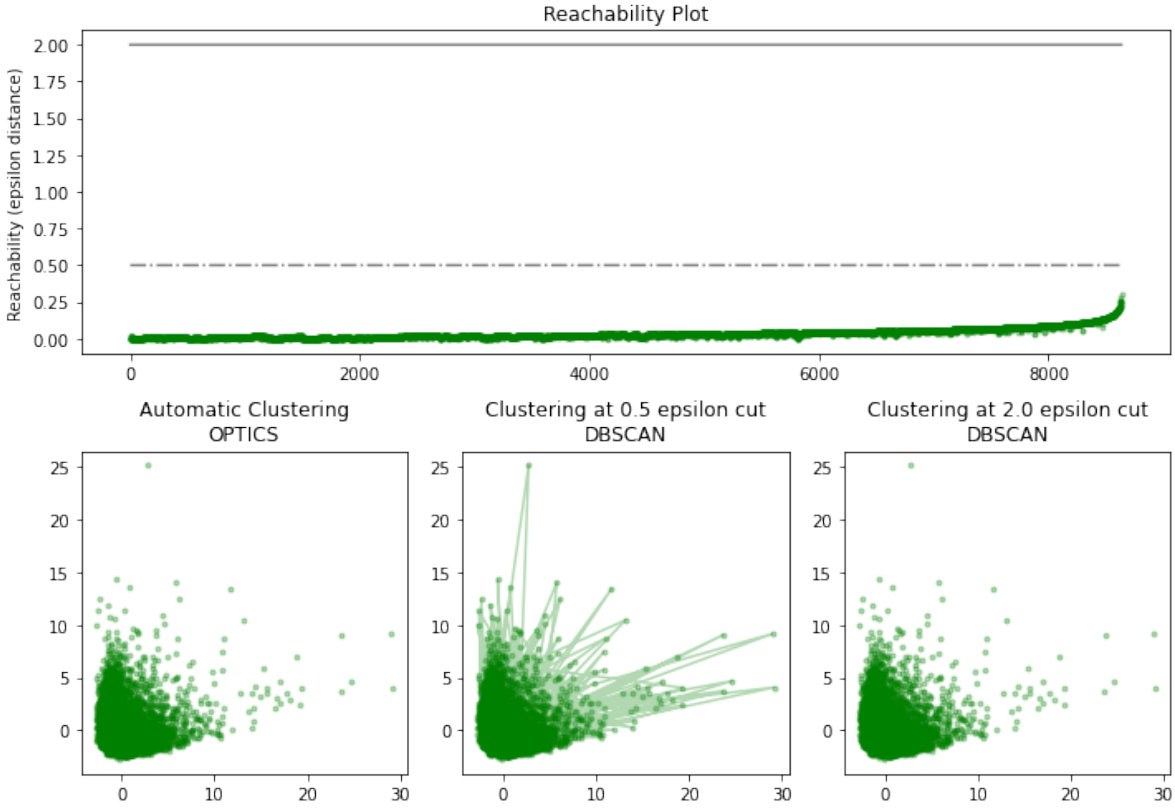

Рисунок 9 — кластеризация OPTICS с метрикой cosine

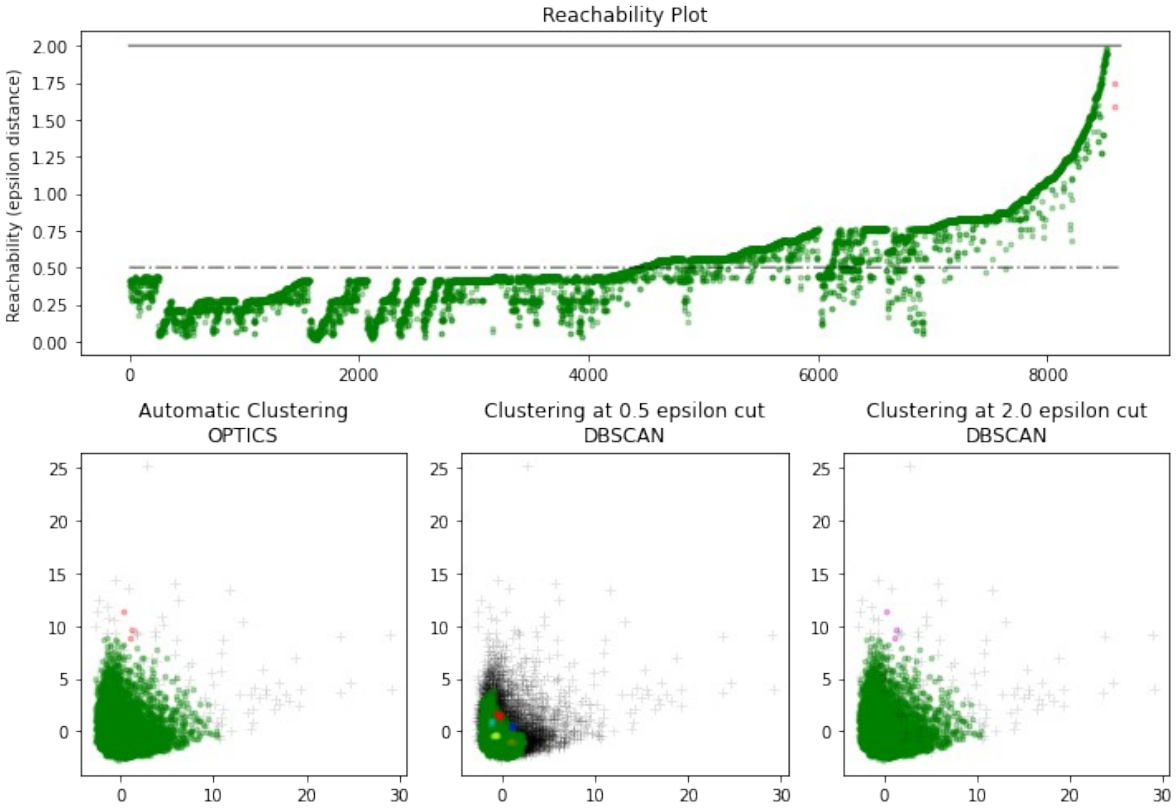

Рисунок 10 — кластеризация OPTICS с метрикой chebyshev

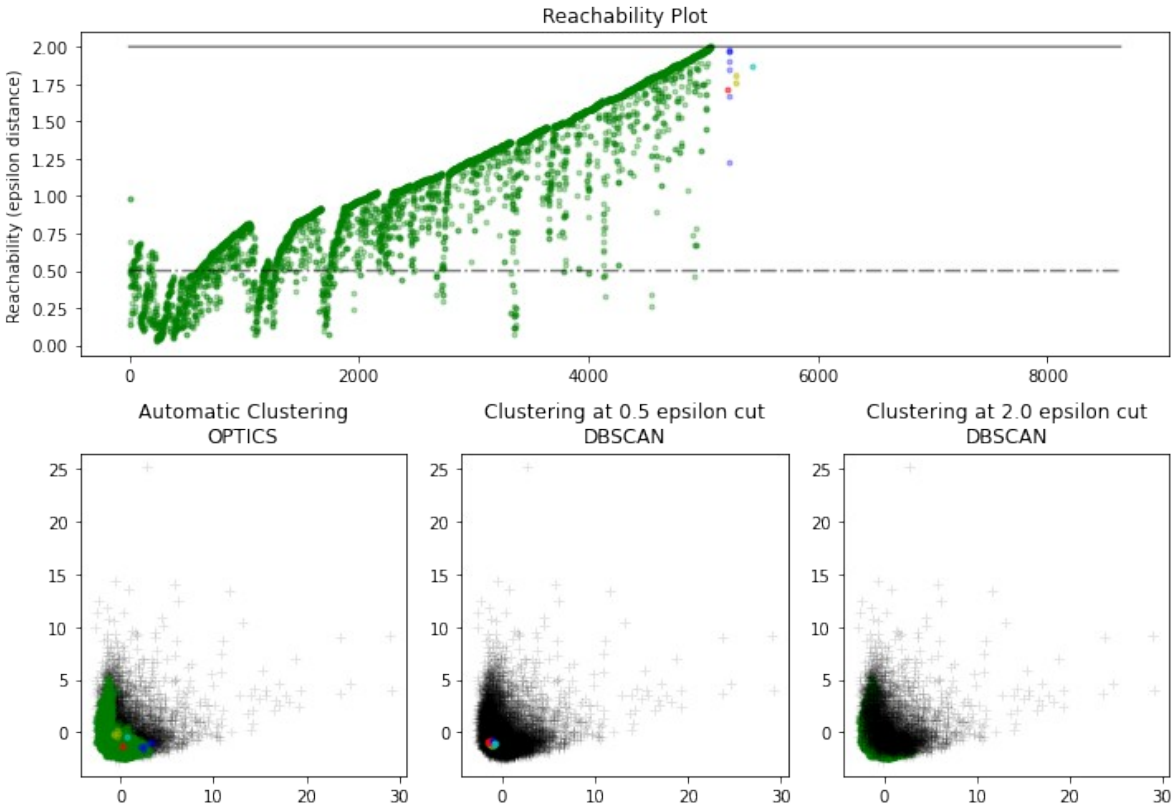

Рисунок 11 — кластеризация OPTICS с метрикой l1

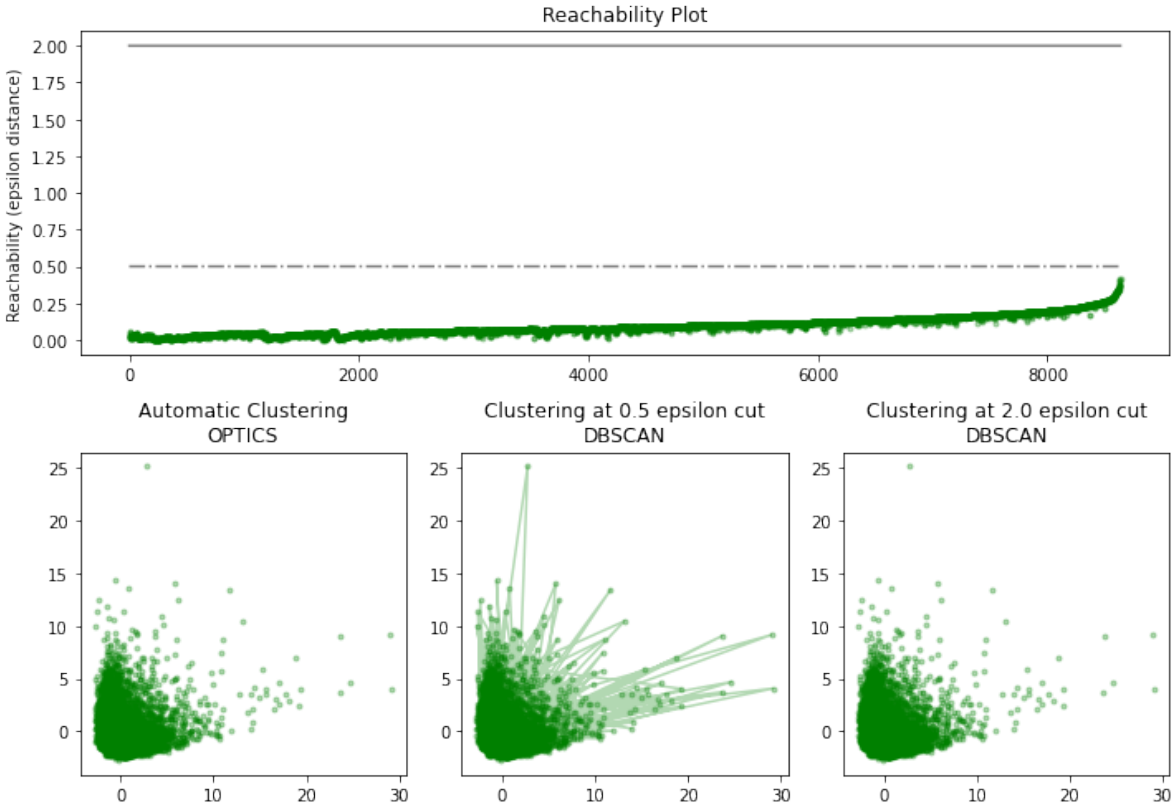

Рисунок 12 — кластеризация OPTICS с метрикой braycurtis

Вывод

Были получены навыки работы с методами кластеризации DBSCAN и OPTICS.