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

ОТЧЁТ

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

Студент гр. 6304 _____ Корытов П.В.
Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург
2020

1. Цель

Ознакомиться с методами кластеризации модуля Sklearn

2. Выполнение

2.1. Загрузка данных

- Проведена загрузка и стандартизация данных. Для стандартизованных данные построены гистограммы (рис. 1)

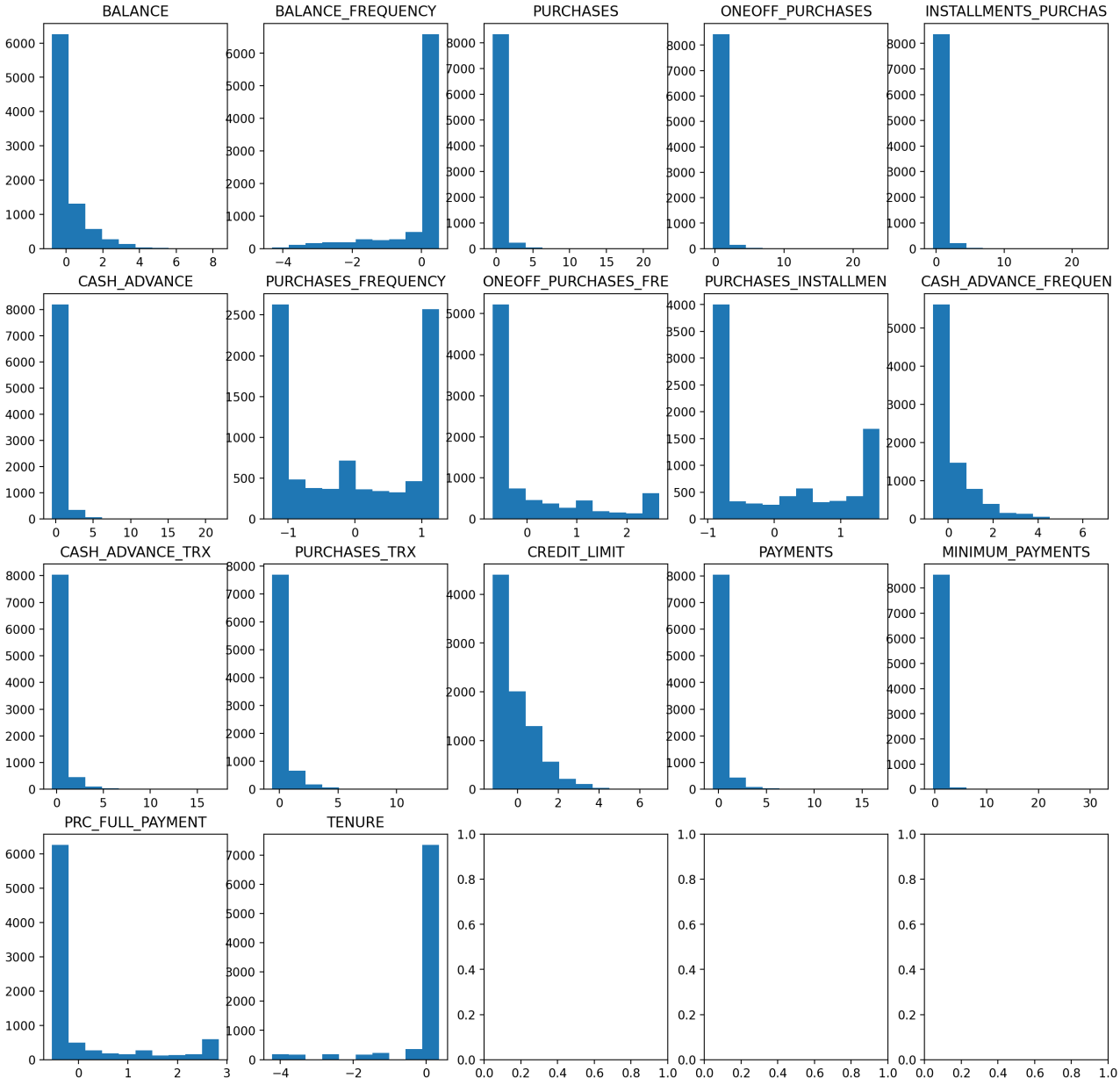

Рисунок 1 – Гистограммы

2.2. DBSCAN

1. Произведена кластеризация DBSCAN с параметрами по умолчанию. Результаты на листинге 1

Листинг 1. Результаты DBSCAN

```
Метки кластеров: {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}
Количество кластеров: 36
Процент выпавших: 0.7512737378415933
```

2. Для анализа работы DBSCAN выбрано 3 метрики — количество кластеров, процент выпавших и процент объектов в самом большом кластере. Графики изменения этих показателей в зависимости от значений параметров алгоритма представлены на рис. 2 и 3.

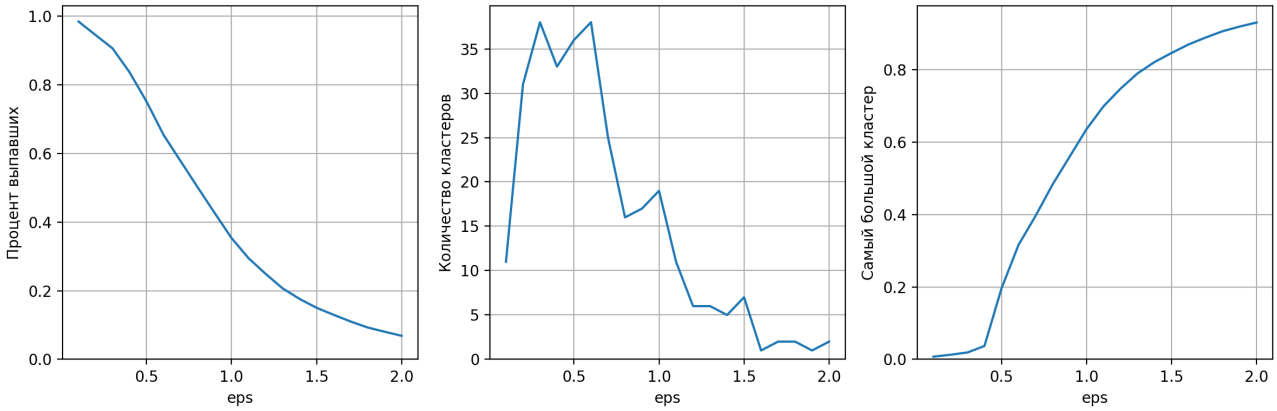

Рисунок 2 – Изменение `eps`

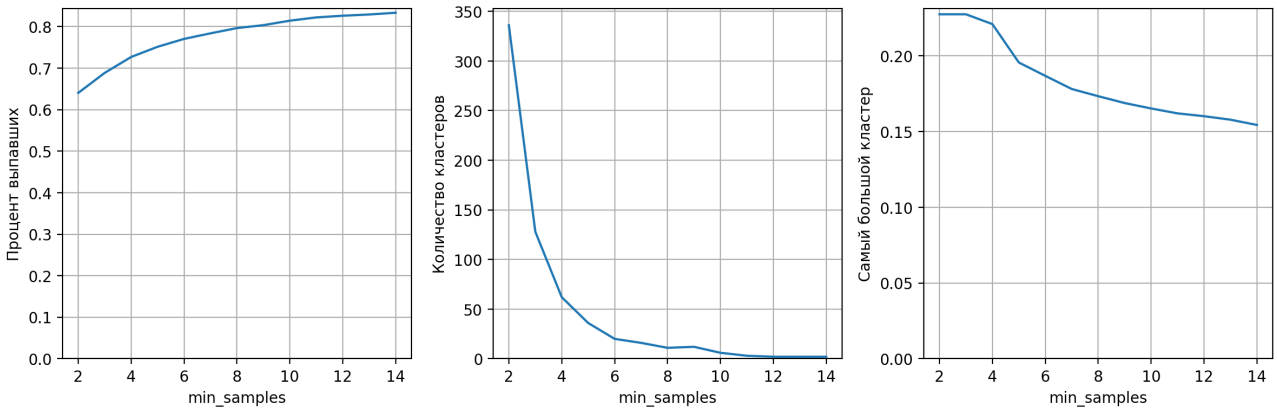

Рисунок 3 – Изменение `min_samples`

Как видно, при увеличении `eps` точки из выпавших переезжают в один

большой кластер. Увеличение `min_samples` приводит к увеличению числа выпавших точек.

Эти же данные представлены на матрице на рис. 4.

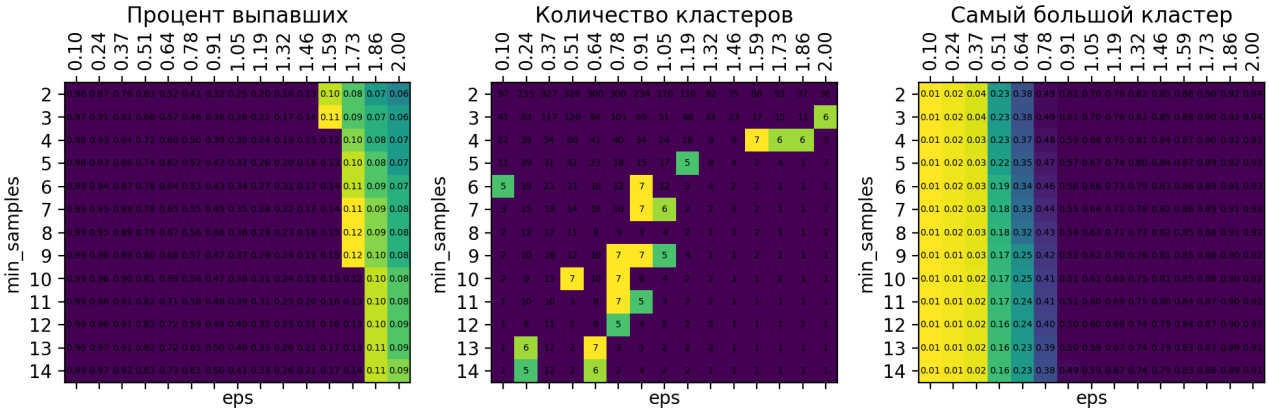

Рисунок 4 – Матрица

Таким образом, для параметров вроде `min_samples=3`, `eps=2` цель формально выполняется, но большая часть данных записывается в единственный кластер.

Ситуация сохраняется при других подходах к обработке данных (`QuantileTran` `PowerTransformer`) и других значениях параметров.

3. Для полученных параметров проведена визуализация с понижением размерности. Результаты на рис. 5.
4. Параметры `DBSCAN`
    - `eps` — радиус окрестности точки;
    - `min_samples` — минимальное число точек в окрестности, чтобы посчитать её основной;
    - `metric` — метрика;
    - `metric_params` — параметры метрики;
    - `algorithm` — алгоритм вычисления соседей;
    - `p` — степень метрики Минковского (напр. 1 — Манхэттенское расстояние, 2 — Евклидово)

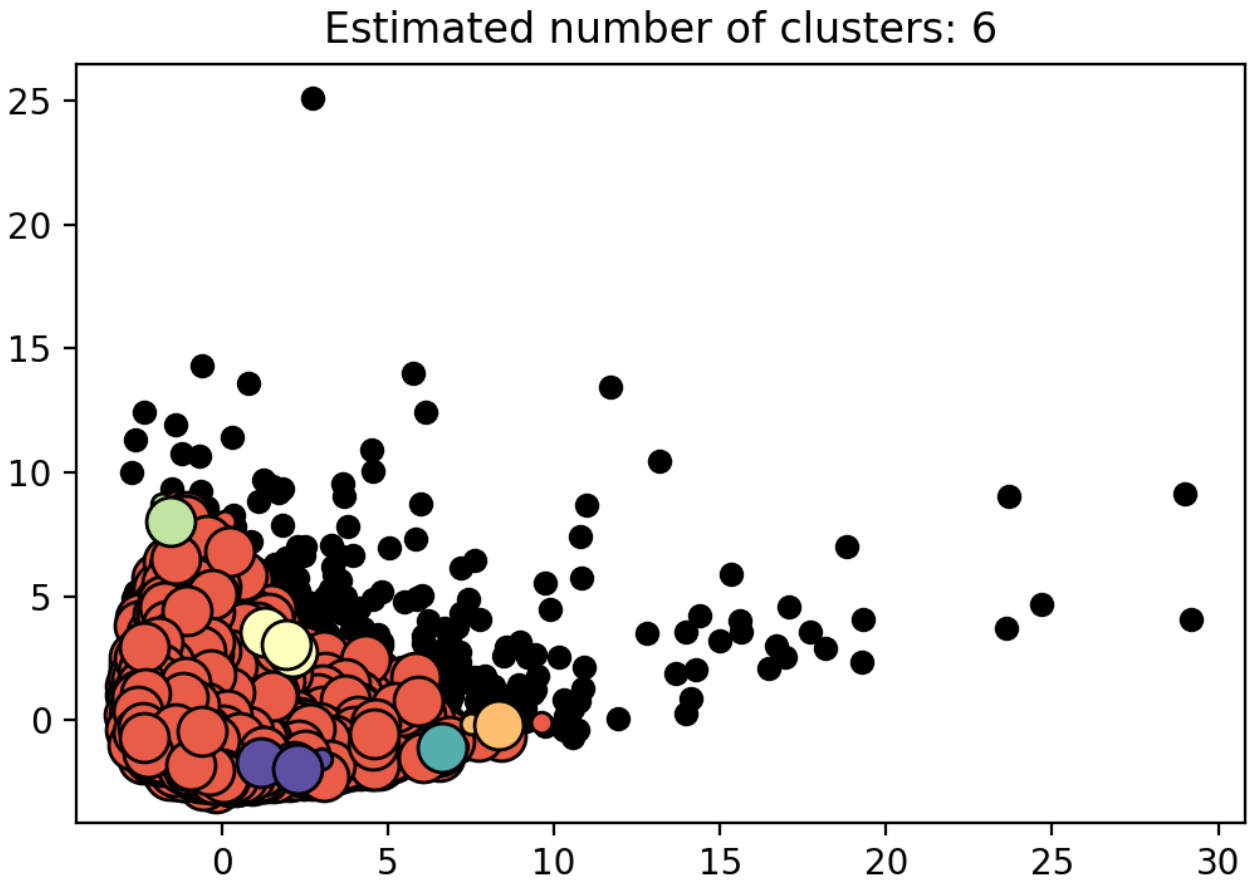

Рисунок 5 – Визуализация с понижением размерности

2.3. OPTICS

1. Произведен поиск параметров `OPTICS`, при которых результаты близки к результатам DBSCAN. Из-за длительности работы поиск проведен вручную. Некоторые результаты представлены в таблице 1.

Таблица 1. Ручной поиск параметров `OPTICS`

	min_samples	max_eps	Кластеров	Выпало	Самый большой
0	5	inf	112	0.898448	0.00266327
1	4	inf	229	0.847499	0.00243168
2	3	inf	524	0.733673	0.00185271
3	2	inf	1629	0.506369	0.000926355
4	2	1	1417	0.56994	0.000926355
5	2	0.5	877	0.73541	0.000926355
6	5	1	107	0.901575	0.00266327
7	5	0.5	76	0.926934	0.00277906
8	15	inf	11	0.965609	0.0081056

Как видно, `OPTICS` кластеризует данные на чрезвычайно большое количе-

ство кластеров с большим количество некластеризованных данных. Похожие на DBSCAN результаты получаются для `cluster_method='dbscan'`.

2. Для описанного случая построен график достижимости. Результат на рис. 6.

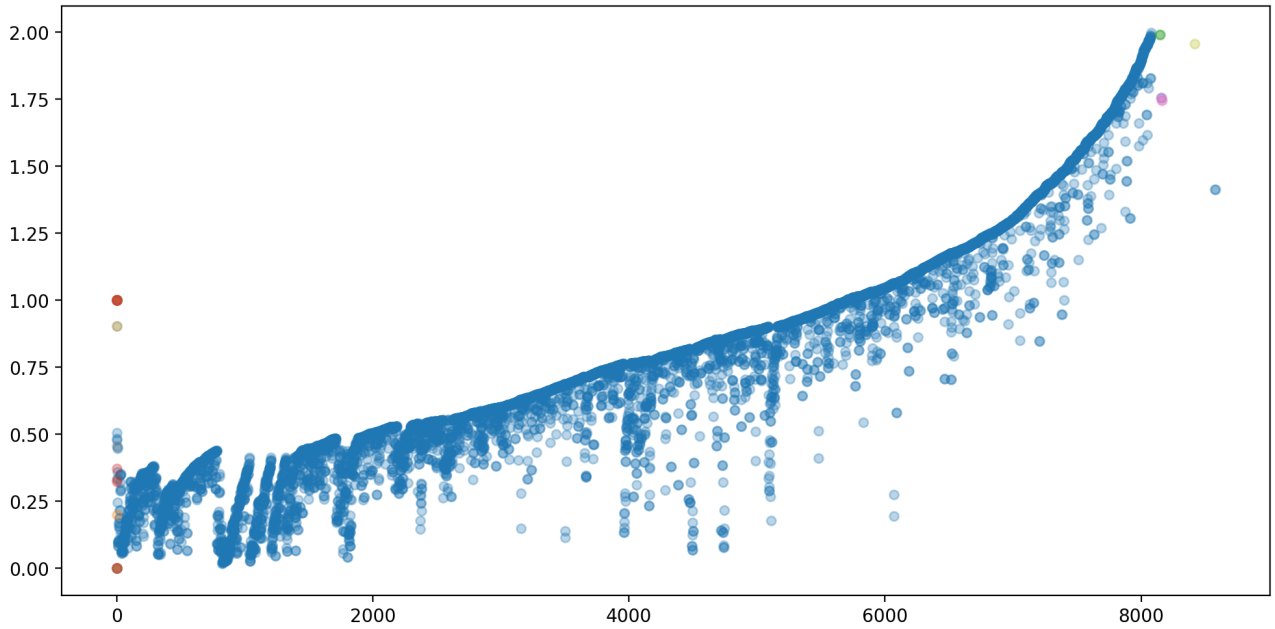

Рисунок 6 – График достижимости

3. Проведен запуск алгоритма с использованием различных метрик. Результаты представлены в таблице 2.
   Как видно, ни одна из метрик для заданных параметров не смогла разделить набор данных на более четкие кластеры.
4. В OPTICS основные точки определяеются также, как в DBSCAN, но дополнительно каждой точке присваивается дистанция достижимости:

$$\mathrm{rd}_{\varepsilon,MinPts}(o,p) = \begin{cases} \text{UNDEFINED} & \text{если } |N_\varepsilon(p)| < MinPts \\ \max(\mathrm{cd}_{\varepsilon,MinPts}(p), \mathrm{dist}(p,o)) & \text{иначе,} \end{cases}$$

где

$$\mathrm{cd}_{\varepsilon,MinPts}(p) = \begin{cases} \text{UNDEFINED} & \text{если } |N_\varepsilon(p)| < MinPts \\ MinPts\text{-й наименьшее в } N_\varepsilon(p) & \text{иначе} \end{cases}$$

Таблица 2. Метрики

	params	Кластеров	Выпавшие	Самый большой
0	{’metric’: ’l1’}	99	0.909217	0.00289486
1	{’metric’: ’l1’, ’min_samples’: 2}	99	0.909217	0.00289486
2	{’metric’: ’l2’}	1545	0.543191	0.000926355
3	{’metric’: ’l2’, ’min_samples’: 2}	112	0.901112	0.00266327
4	{’metric’: ’l2’, ’min_samples’: 10}	1629	0.506369	0.000926355
5	{’metric’: ’l2’, ’min_samples’: 40}	23	0.952987	0.00451598
6	{’metric’: ’l2’, ’max_eps’: 1}	2	0.984831	0.00949514
7	{’metric’: ’manhattan’}	108	0.903659	0.00266327
8	{’metric’: ’manhattan’, ’min_samples’: 2}	99	0.909217	0.00289486
9	{’metric’: ’manhattan’, ’max_eps’: 1}	1545	0.543191	0.000926355
10	{’metric’: ’manhattan’, ’max_eps’: 1, ’min_samples’: 10}	68	0.934808	0.00289486
11	{’metric’: ’chebyshev’}	17	0.967114	0.00312645
12	{’metric’: ’chebyshev’, ’min_samples’: 2}	141	0.870195	0.00301065
13	{’metric’: ’chebyshev’, ’min_samples’: 10}	1550	0.527559	0.00138953
14	{’metric’: ’chebyshev’, ’min_samples’: 10, ’max_eps’: 1}	28	0.914775	0.00706346
15	{’metric’: ’chebyshev’}	28	0.914775	0.00706346
16	{’metric’: ’sqeuclidean’}	141	0.870195	0.00301065

3. Выводы

Произведено знакомство с реализацией методов `DBSCAN` и `OPTICS` в модуле `Sklearn`.

Для заданного набора данных оба метода либо производят разбиение на очень большое число маленьких кластеров, либо на один большой кластер, либо не классифицируют большую часть результатов.

Предположительно, это связано со структурой набора данных — большая часть транзакций однородна, но выделяется несколько видов банковского мошенничества, которые встречаются достаточно редко.