

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

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

Студент гр. 6304

Ковынев М.В.

Преподаватель

Жангиров Т.Р.

Санкт-Петербург

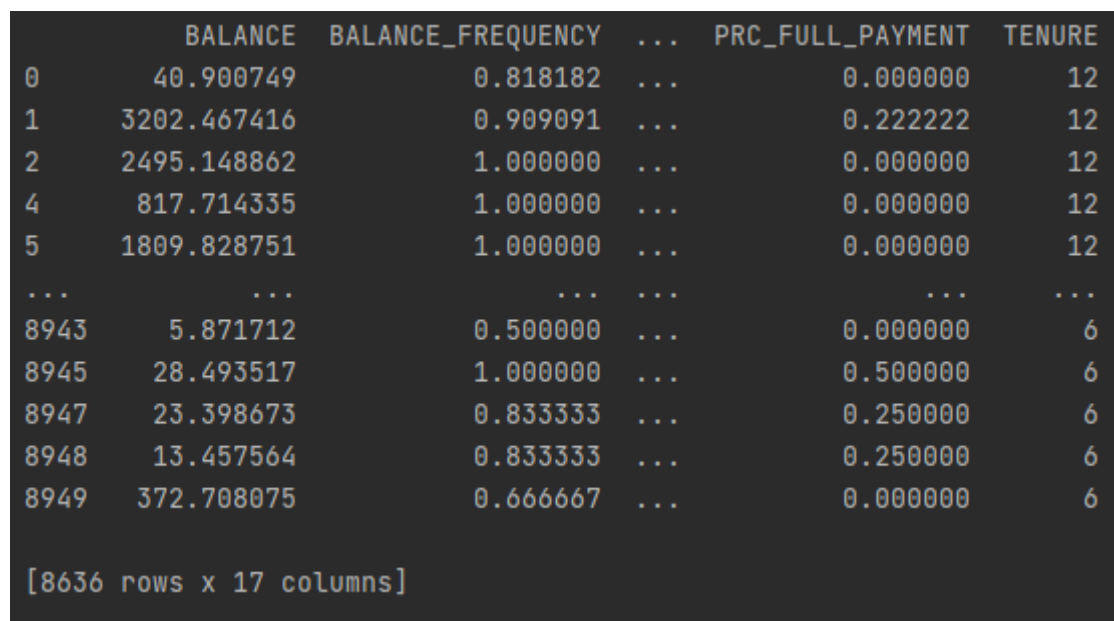
2020

## Цель

Ознакомиться с методами кластеризации модуля Sklearn

## Ход работы

1. Загружен датасет по ссылке: <https://www.kaggle.com/arjunbhasin2013/ccdata>. Данные представлены в виде csv файла. Датасет содержит пропущенные значения
2. Создан Python скрипт. Загружены данные в датафрейм, убрав столбец с метками и откинув наблюдения с пропущенными значениями



	BALANCE	BALANCE_FREQUENCY	...	PRC_FULL_PAYMENT	TENURE
0	40.900749	0.818182	...	0.000000	12
1	3202.467416	0.909091	...	0.222222	12
2	2495.148862	1.000000	...	0.000000	12
4	817.714335	1.000000	...	0.000000	12
5	1809.828751	1.000000	...	0.000000	12
...	...	...	...	...	...
8943	5.871712	0.500000	...	0.000000	6
8945	28.493517	1.000000	...	0.500000	6
8947	23.398673	0.833333	...	0.250000	6
8948	13.457564	0.833333	...	0.250000	6
8949	372.708075	0.666667	...	0.000000	6

[8636 rows x 17 columns]

Рисунок 1 — Исходные данные

3. Проведем кластеризацию методов k-средних. Так как разные признаки лежат в разных шкалах, то стандартизируем данные.
4. Проведем кластеризацию методов DBSCAN при параметрах по умолчанию. Выведем метки кластеров, количество кластеров, а также процент наблюдений, которые кластеризовать не удалось

{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

0.7512737378415933

5. Параметры DBSCAN:

- `eps` – радиус окрестности основной точки
- `min_samples` – минимальное число точек в окрестности, чтобы считать ее основной
- `metric, metric_params` – метрика вычисления расстояния
- `algorithm` - {‘auto’, ‘ball\_tree’, ‘kd\_tree’, ‘brute’ } – алгоритм поиска соседей
- `p` – степень метрики Минковского

6. Построен график количества кластеров и процента не кластеризованных наблюдений в зависимости от максимальной рассматриваемой дистанции между наблюдениями. Минимальное значение количества точек образующих, кластер оставлено по умолчанию

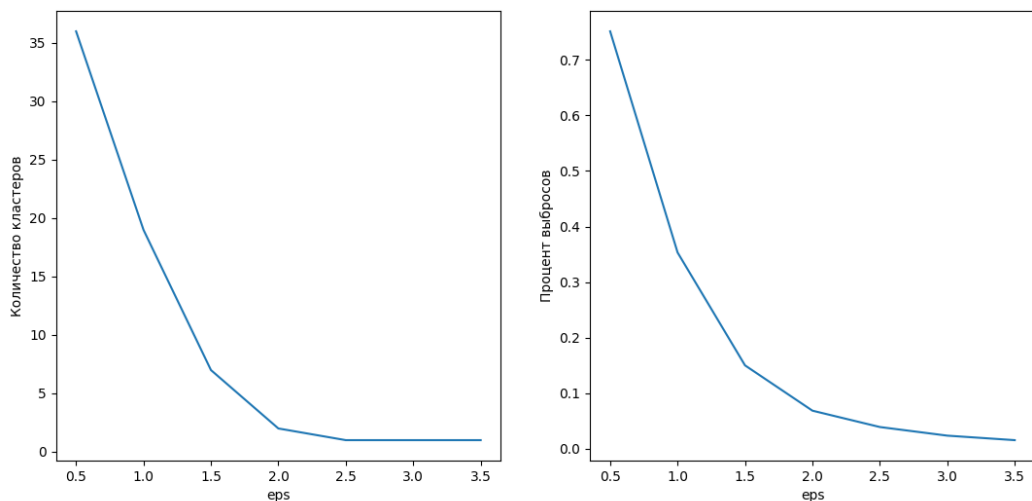


Рисунок 1 — Количество кластеров и процент выбросов от `eps`

7. Построен график количества кластеров и процента не кластеризованных наблюдений в зависимости от минимального значения количества точек, образующих кластер. Максимальную рассматриваемую дистанцию между наблюдениями оставлено по умолчанию

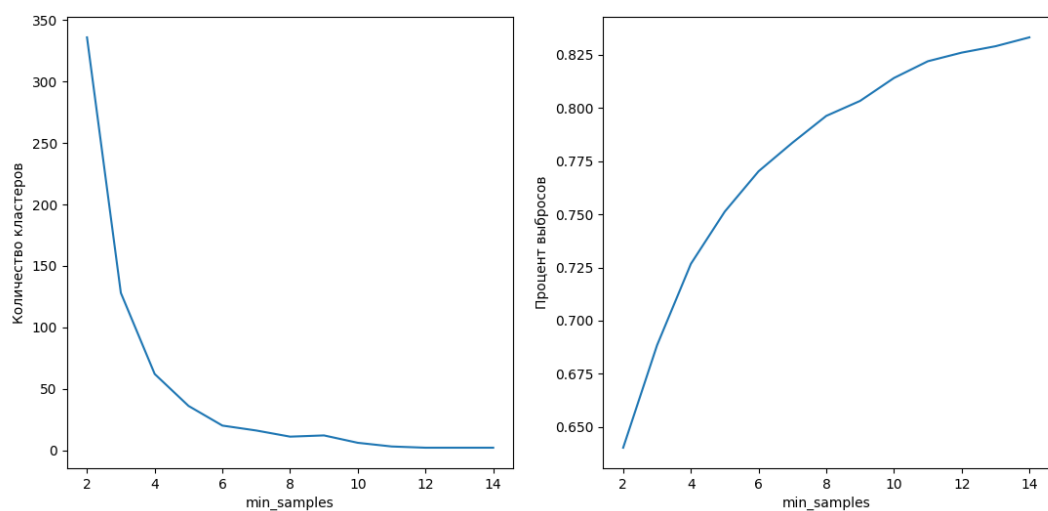


Рисунок 2 — Количество кластеров и процент выбросов от min\_samples

8. Определено значения параметров, при котором количество кластеров получается от 5 до 7, и процент не кластеризованных наблюдений не превышает 12%

<b>samples</b>	<b>eps</b>	<b>count of clusters</b>	<b>percent</b>	<b>is valid</b>
1	1.5	1106	0.0	False
1	1.6	941	0.0	False
1	1.7	793	0.0	False
1	1.8	677	0.0	False
1	1.9	585	0.0	False
1	2.0	518	0.0	False
1	2.1	457	0.0	False
1	2.2	397	0.0	False
1	2.3	348	0.0	False
1	2.4	315	0.0	False
2	1.5	70	0.12	False
2	1.6	66	0.101	False
2	1.7	58	0.085	False
2	1.8	48	0.073	False
2	1.9	36	0.064	False
2	2.0	36	0.056	False
2	2.1	32	0.049	False
2	2.2	30	0.043	False
2	2.3	32	0.037	False
2	2.4	29	0.033	False
3	1.5	21	0.131	False
3	1.6	17	0.113	False
3	1.7	13	0.096	False
3	1.8	11	0.082	False

3	1.9	8	0.07	False
3	2.0	6	0.063	True
3	2.1	9	0.055	False
3	2.2	12	0.047	False
3	2.3	11	0.042	False
3	2.4	8	0.038	False

9. Понижена размерность данных до 2 при используя метод главных компонент. Визуализированы результаты кластеризации, полученные в пункте 6 (метки должны быть получены на данных до уменьшения размерности)

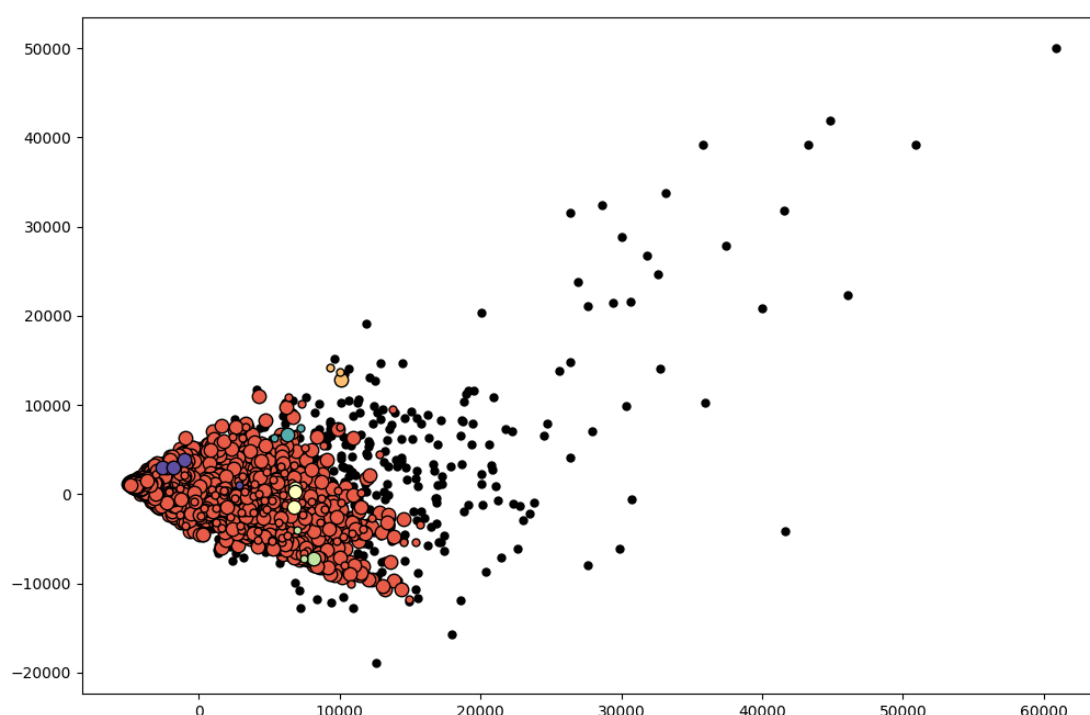


Рисунок 3 — Результат кластеризации

10. Описаны параметры метода OPTICS, а также какими атрибутами он обладает

- `max_eps` – максимальное расстояние между двумя наблюдениями, чтобы один считался соседним с другим
- `min_samples` – минимальное число точек в окрестности, чтобы считать ее основной
- `metric, metric_params` – метрика вычисления расстояния
- `cluster_method` - метод извлечения кластеров – `{‘xi’, ‘dbscan’}`

- algorithm - {‘auto’, ‘ball\_tree’, ‘kd\_tree’, ‘brute’ } – алгоритм поиска соседей
- p – степень метрики Минковского

11. Результаты кластеризации методом OPTICS близки к результатам DBSCAN при cluster\_method=dbscan, max\_eps=2, min\_samples=3.

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

12. Визуализирован полученный результат, а также построен график достижимости (reachability plot)

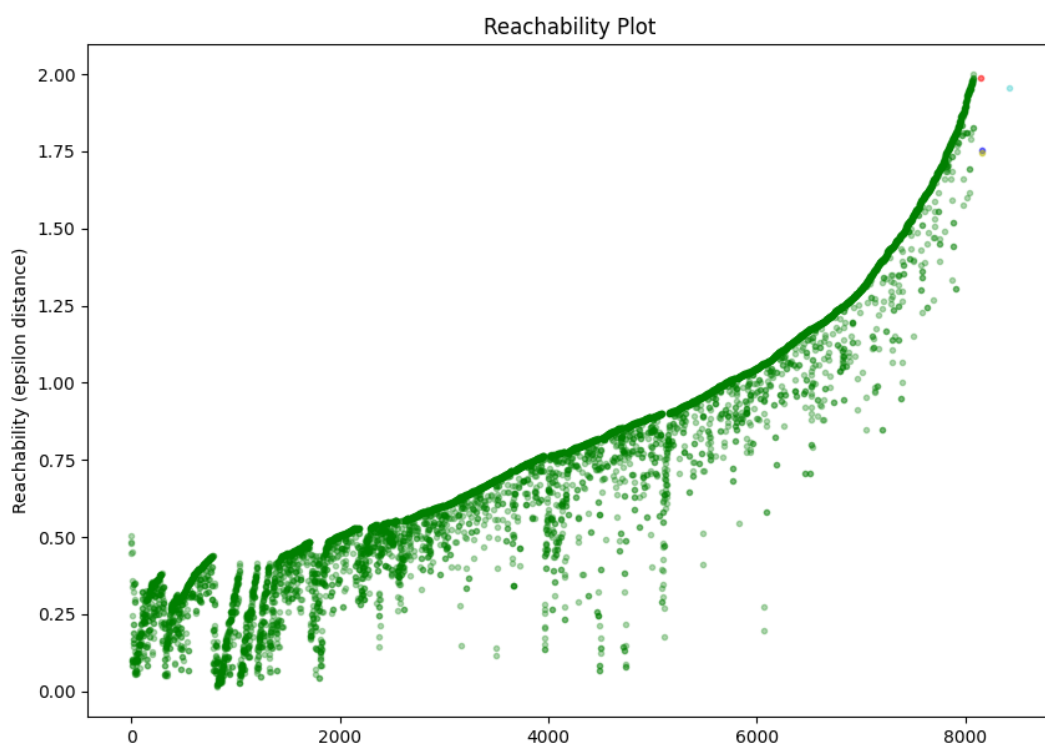


Рисунок 4 — reachable plot

13. Исследована работа метода OPTICS с использованием различных метрик.

Метрика	min_samples	max_eps	num_of_clusters	not_classified
---------	-------------	---------	-----------------	----------------

manhattan	5	1	30	0.74
		10	1	0.01
	10	1	6	0.81
		10	1	0.01
euclidean	5	1	19	0.36
		10	1	0.0
	10	1	3	0.42
		10	1	0.0
canberra	5	1	18	0.5
		10	1	0.0
	10	1	19	0.89
		10	1	0.0
Braycurtis	5	1	1	0.0
		10	1	0.0
	10	1	1	0.0
		10	1	0.0
chebyshev	5	1	3	0.1
		10	1	0.0
	10	1	1	0.12
		10	10	0.0

## Вывод

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