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. 8304 _____ Холковский К.В

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург

2021

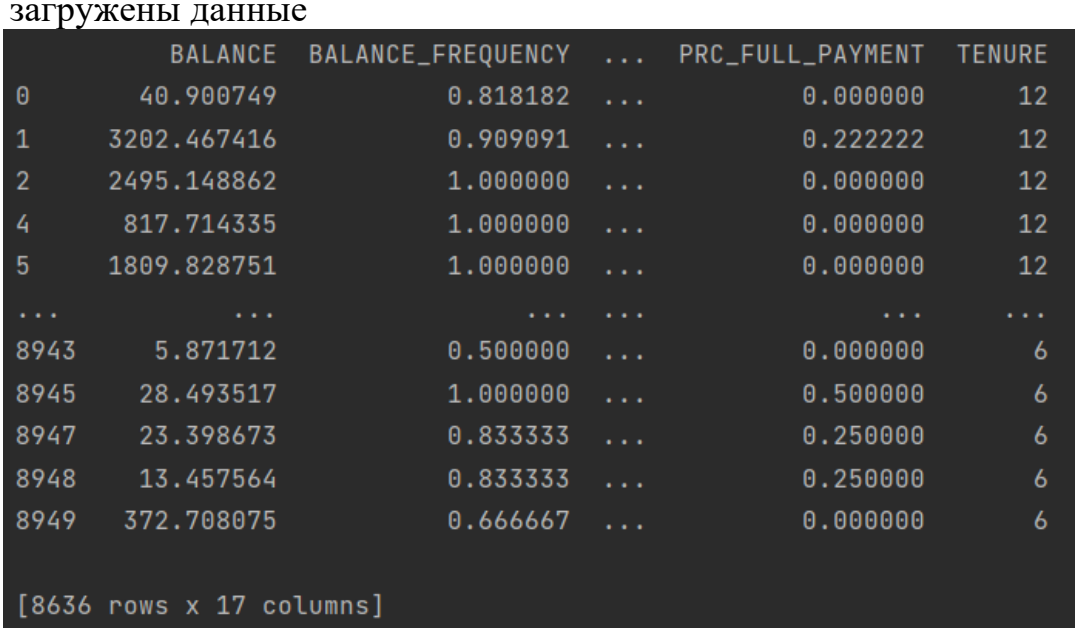
Цель работы

Ознакомиться с методами кластеризации модуля Sklearn.

Ход работы

1. Загрузка данных

Были загружены данные



	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]

Рис 1 – Загруженные данные

2. DBSCAN

1) Провели кластеризацию методом DBSCAN

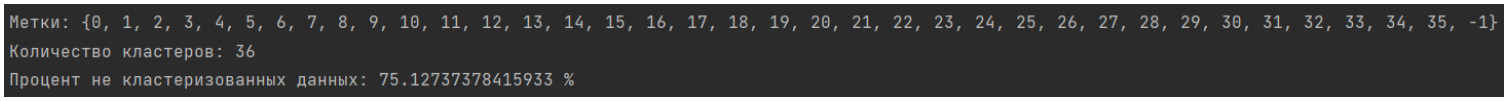


Рис 2 – Результат работы DBSCAN

Таблица 1 - Описание параметров DBSCAN

Параметр	Смысл параметра
eps	Максимальное расстояние между двумя элементами
min_samples	Количество выборок (или общий вес) в окрестности точки, которая будет считаться базовой точкой. Сюда входит и сама точка.
metric	Метрика для расчета расстояния.
metric_params	Параметры для метрики
algorithm	Алгоритм, который будет использоваться для вычисления

	точечных расстояний и поиска ближайших соседей.
leaf_size	Может повлиять на скорость построения и запрос, а также на объем памяти, необходимый для хранения дерева.
p	Степень метрики Минковского, которая будет использоваться для вычисления расстояния между точками.
n_jobs	Число процессов, чтобы распараллелить.

2) Был построен график количества кластеров и процента не кластеризованных наблюдений в зависимости от максимальной рассматриваемой дистанции между наблюдениями.

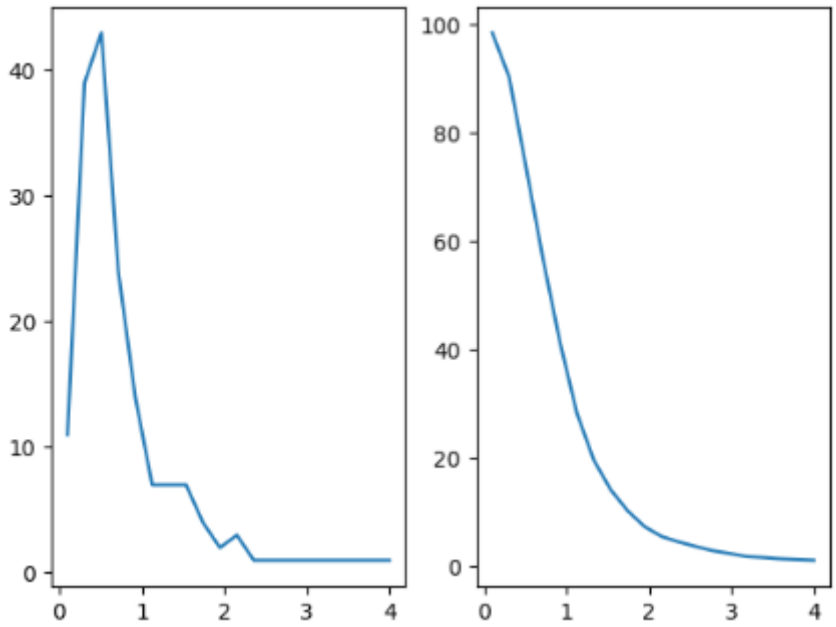


Рис 3 – Графики кол-ва кластеров и процента невнесенных данных от максимального расстояния.

3) Был построен график количества кластеров и процента не кластеризованных наблюдений в зависимости от максимальной рассматриваемой дистанции между наблюдениями.

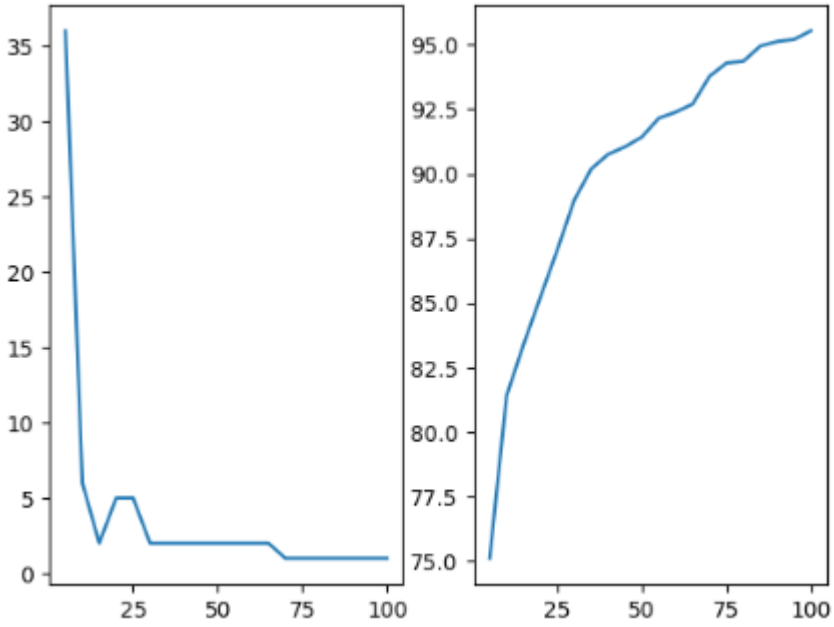


Рис 4 - Графики кол-ва кластеров и процента невнесенных данных от минимального числа элементов в кластере.

4) Были определены значения параметров, при котором количество кластеров равно 5, а процент некластеризованых данных равен 6. Eps=2.05, min_samples=3.

5) Были визуализированы результаты работы DBSCAN для параметров из пункта 4.

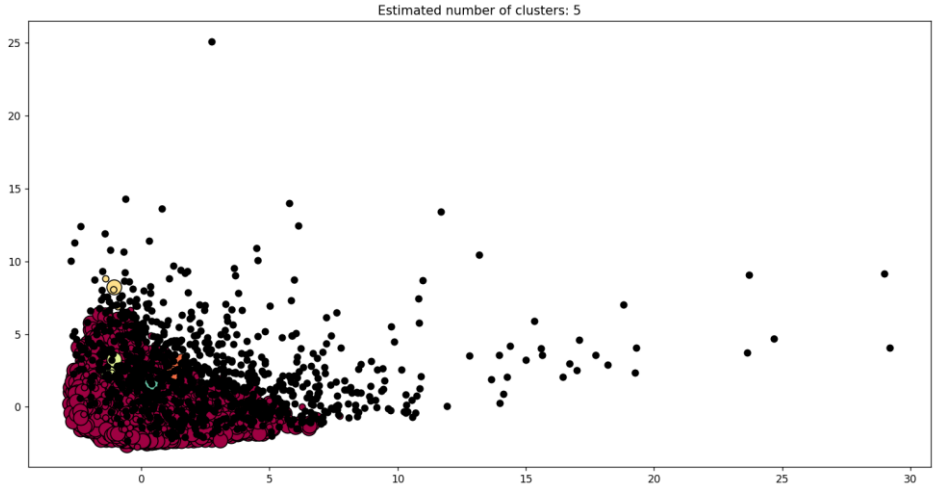


Рис 5 – Результат DBSCAN

3. **OPTICS**

Таблица 2 - Параметры OPTICS

Параметр	**Смысл параметра**
min_samples	Количество выборок в окрестности для точки, которая будет считаться базовой.
max_eps	Максимальное расстояние между двумя образцами, чтобы один считался соседним с другим.
metric	Метрика, используемая для вычисления расстояния.
p	Параметр для метрики Минковского из pairwise_distances.
metric_params	Дополнительные аргументы ключевого слова для метрической функции.
cluster_method	Метод извлечения, используемый для извлечения кластеров с использованием вычисленной достижимости и упорядочения.
eps	The maximum distance between two samples for one to be considered as in the neighborhood of the other.
xi	Определяет минимальную крутизну графика достижимости, которая составляет границу кластера.
predecessor_correction	Правильные кластеры в соответствии с предшественниками, рассчитанными OPTICS
min_cluster_size	Минимальное количество выборок в кластере OPTICS, выраженное в виде абсолютного числа или доли от количества выборок
algorithm	Алгоритм, используемый для вычисления ближайших соседей:
leaf_size	Размер листа передается в BallTree или KDTree.
memory	Используется для кеширования вывода вычисления дерева.
n_jobs	Количество параллельных заданий для поиска соседей.

Таблица 3 - Атрибуты OPTICS

Атрибут	Смысл атрибута
labels_	Кластерные метки для каждой точки в наборе данных, заданной для fit ()
reachability_	Расстояния достижимости на выборку, индексированные по порядку объектов.
ordering_	Кластерный упорядоченный список выборочных индексов.
core_distances_	Расстояние, на котором каждый образец становится центральной точкой, индексируется по порядку объектов.
predecessor_	Точка, из которой была получена выборка, проиндексированная по порядку объектов.
cluster_hierarchy_	Список кластеров в виде [начало, конец] в каждой строке, включая все индексы.
n_features_in_	Количество деталей, видимых во время посадки.
feature_names_in_	Названия особенностей, замеченных во время посадки. Определяется только тогда, когда X имеет имена функций, которые являются строками.

1) Были найдены параметры OPTICS соответствующие результатам как в 4 пункте DBSCAN. max_eps = 2.05, min_samples = 3.

2) Был визуализирован результат и построен график достижимости.

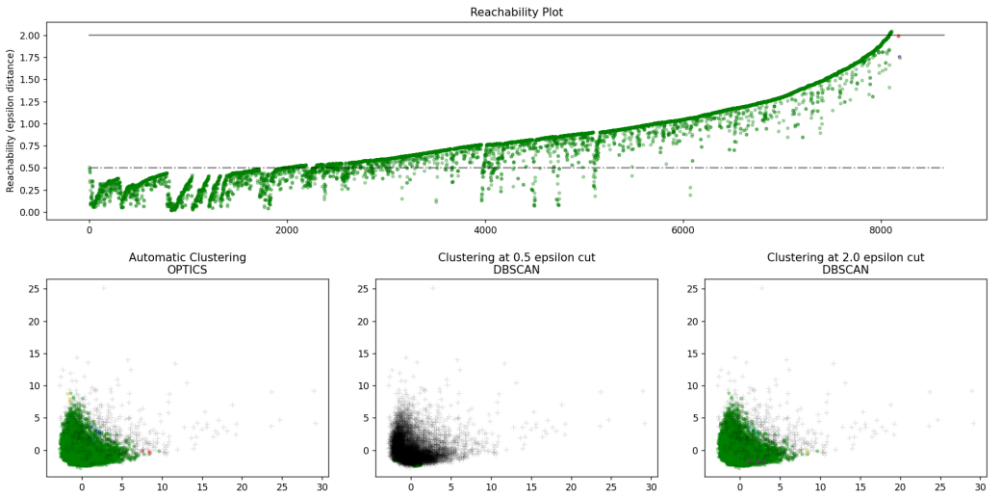


Рис 6 – Результат выполнения

braycurtis	$\sum \lvert u_i - v_i \rvert / \sum \lvert u_i + v_i \rvert$
chebyshev	$\max_i \lvert u_i - v_i \rvert.$
canberra	$d(u, v) = \sum_i \frac{\lvert u_i - v_i \rvert}{\lvert u_i \rvert + \lvert v_i \rvert}.$
russellrao	$\frac{n - c_{TT}}{n}$
manhattan	$\sum_i \lvert u_i - v_i \rvert.$

Вывод

Ознакомились с методами кластеризации модуля Sklearn.