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

ОТЧЕТ

по лабораторной работе № 3
по дисциплине «Машинное обучение»
Тема: «Кластеризация»

Студент гр. 0308 Сабурова Е.А.

Преподаватель Гатауллин Р.И.

Санкт-Петербург

2023

Оглавление

1. Цель работы

Получения и закрепления навыков предобработки данных и применения методов машинного обучения для решения задач кластеризации.

2. Задание

1. Обучение моделей и подбор параметров (где применимо):
    a. метод К-средних
    b. DBSCAN
    c. Иерархическая кластеризация
2. Оценка моделей
    a. Экспертная оценка
    b. Сравнение разбиения на классы с помощью кластеризации с реальными.
    c. Визуализация предсказанных значений

3. Краткая теоретическая информация по теме работы

Кластеризация — распределение данных на группы, группировка множества объектов на подмножества (кластеры) таким образом, чтобы объекты из одного кластера были более похожи друг на друга, чем на объекты из других кластеров по какому-либо критерию.

Кластеризировать объекты можно по разным алгоритмам. Чаще всего используют следующие:

- на основе центра тяжести треугольника;
- на базе подключения;
- сокращения размерности;
- плотности (основанные на пространственной кластеризации);
- вероятностные;
- машинное обучение, в том числе нейронные сети.

Алгоритмы кластеризации используются в биологии, социологии и информационных технологиях.

Коэффициент силуэта в кластеризации — это величина, которая позволяет оценить степень соответствия построенной кластерной структуры обучающим данным, т.е. оценить качество кластеризации.

Иными словами, коэффициент силуэта показывает, насколько каждый объект «похож» на другие объекты в том кластере, в который он был распределен в процессе кластеризации, и «не похож» на объекты из других кластеров.

Силуэт кластера — метод графического представления результатов кластеризации, с помощью которого можно визуально оценить качество построенной кластерной модели.

Алгоритмы моделей машинного обучения, использованные в работе

Алгоритм K-средних

Алгоритм k-средних разделяет множество X из N строк в K кластеров, каждый из которых описывается его средним μj. Среднее также называют центроидом. Центры не относятся к множеству X.

После этого, алгоритм пытается минимизировать “инерционность” или же квадратичное отклонение точек кластеров от центров этих кластеров

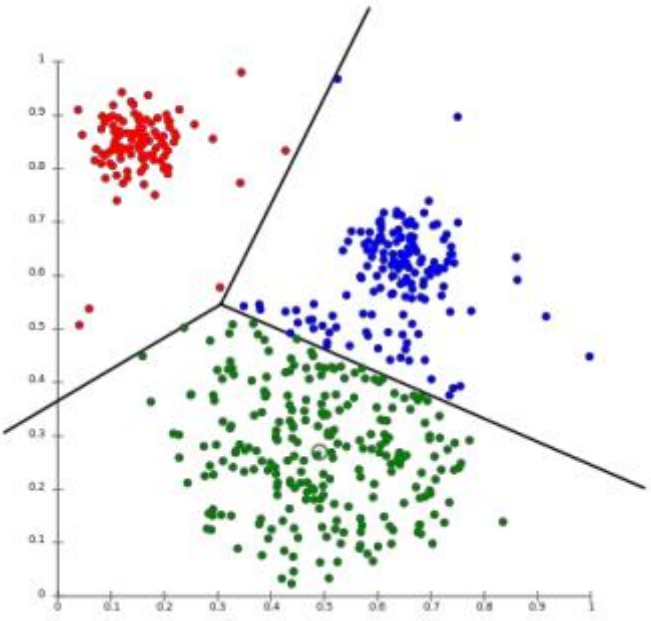

Рисунок 1 – изображение работы k-средних

Недостатки алгоритма:

1. Число кластеров надо знать заранее.

2. Результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен. Классический вариант подразумевает случайный выбор кластеров, что очень часто являлось источником погрешности.

3. Инерция предполагает, что кластеры являются выпуклыми и изотропными, что не всегда так. Она плохо реагирует на вытянутые кластеры или многообразия неправильной формы.

DBSCAN (Density-based spatial clustering of applications with noise)

Если дан набор точек в некотором пространстве, алгоритм группирует вместе точки, которые тесно расположены (точки со многими близкими соседями), помечая как выбросы точки, которые находятся одиноко в областях с малой плотностью (ближайшие соседи которых лежат далеко)

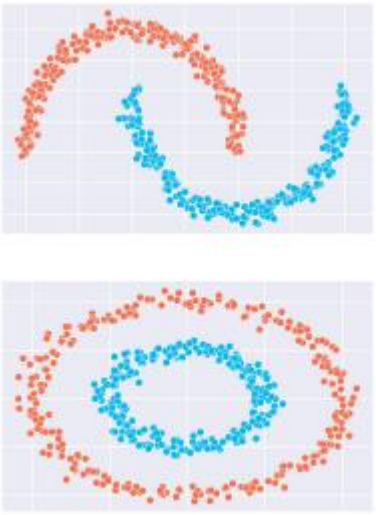

Рисунок 2 – изображение DBSCAN

Иерархическая кластеризация

Иерархическая кластеризация - это общее семейство алгоритмов кластеризации, которые строят кластеры путем последовательного слияния или разделения. Такая иерархия кластеров представляется в виде дерева (или дендрограммы). Корнем дерева является уникальный кластер, в котором собраны все образцы, листьями - кластеры, содержащие только один образец

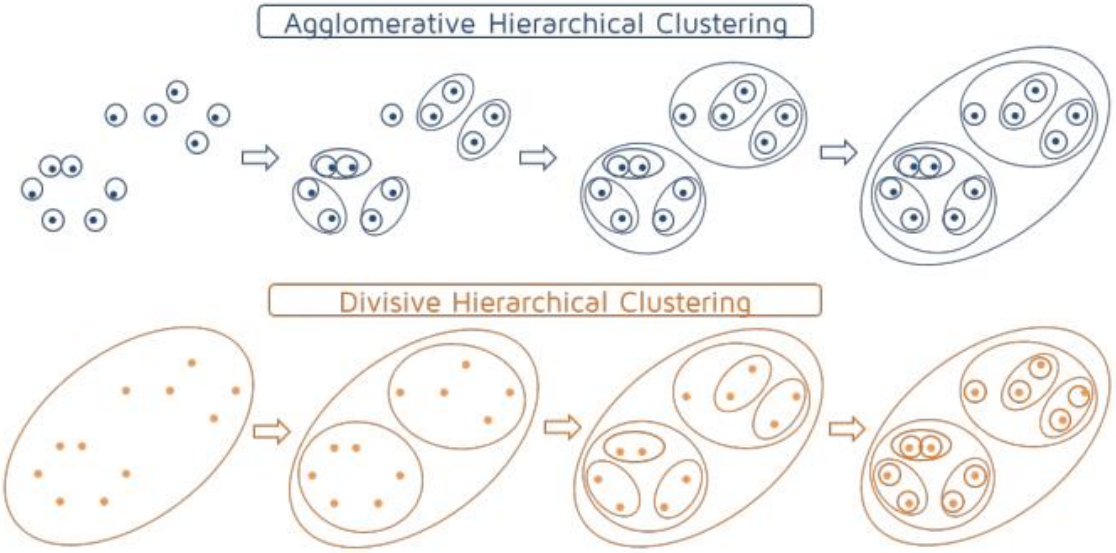

Рисунок 3 – изображение иерархической кластеризации

4. Ход работы

В лабораторной работе рассматривается кластеризация для набора данных о погоде в Индии. Цель работы – проверить, делятся ли данные на кластеры, отражающие погоду в определенную фазу луны.

Визуализация исходных данных представлена на Рисунке 1.

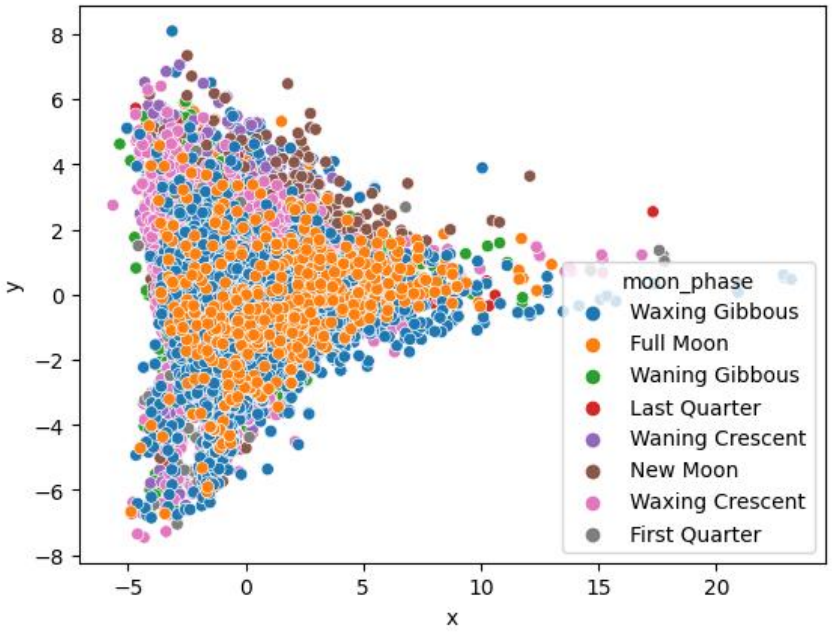

Рисунок 4 – Визуализация исходных данных

KMeans

В качестве параметров для обучения изменялся только параметр

количество кластеров. Количество кластеров = 8 (по количеству фаз луны).

Результаты оценки алгоритма по метрикам представлены на рисунке.

```
homogeneity_score = 0.08121369811032282
completeness_score = 0.08280654160503725
v_measure_score = 0.08200238559755746
adjusted_rand_score = 0.04562984019395719
adjusted_mutual_info_score = 0.08132977191411205
silhouette_score = -0.05656872924342385
```

Для экспертной оценки выводилась основная статистическая информация о датасете, сгруппированном по размеченным кластерам. Результаты представлены на рисунке.

	air_quality_PM10								wind_mph		...	temperature_celsius		wind_degree					
	count	mean	std	min	25%	50%	75%	max	count	mean	...	75%	max	count	mean	std	min	25%	50%	75%	max
kmeans																					
0	1432.0	15.287709	13.582812	1.2	6.100	11.10	20.000	93.7	1432.0	2.840363	...	24.3	30.0	1432.0	127.132682	84.163964	1.0	67.00	113.0	173.0	360.0
1	4570.0	21.584114	18.558702	0.7	7.200	16.50	31.400	157.6	4570.0	5.000088	...	26.0	32.4	4570.0	234.762363	81.359571	1.0	213.00	254.0	283.0	360.0
2	3885.0	41.892716	20.916811	3.1	25.400	39.90	55.900	133.6	3885.0	4.958636	...	26.6	33.0	3885.0	142.086744	88.823935	1.0	73.00	122.0	210.0	360.0
3	248.0	340.590726	187.421243	55.2	214.175	317.30	419.300	1043.7	248.0	3.934274	...	28.0	34.2	248.0	137.423387	113.726687	10.0	38.25	109.0	230.0	360.0
4	775.0	45.559097	29.543074	0.7	18.950	44.10	66.300	158.7	775.0	3.656129	...	15.7	21.5	775.0	118.958710	104.954307	1.0	45.00	71.0	165.0	359.0
5	1112.0	51.507284	44.366585	1.4	18.775	37.45	70.725	229.6	1112.0	7.084083	...	31.2	38.3	1112.0	258.282374	72.721058	2.0	247.00	275.0	301.0	359.0
6	3302.0	137.374258	56.378228	50.6	93.225	125.40	167.750	374.4	3302.0	4.018686	...	27.7	36.3	3302.0	148.016354	104.033005	1.0	63.00	117.0	242.0	360.0
7	2731.0	17.160857	19.565889	0.8	5.300	11.50	22.000	233.2	2731.0	11.578945	...	26.6	35.1	2731.0	232.635665	64.168604	1.0	223.00	252.0	270.0	355.0

В данной таблице основной информацией можно считать количество элементов в кластере, т.к. остальные параметры, к сожалению, оказались не информативными (во всех кластерах примерно одинаковые). Если посмотреть на оригинальное разбиение на рисунке, то можно сделать вывод, что распределения похожи:

```
Waxing Crescent    3837
Waxing Gibbous     3828
Waning Crescent    3283
Waning Gibbous     3275
Full Moon          1644
New Moon           1094
First Quarter       549
Last Quarter        545
Name: count, dtype: int64
```

Однако, визуализация признаков не совпадает с исходной:

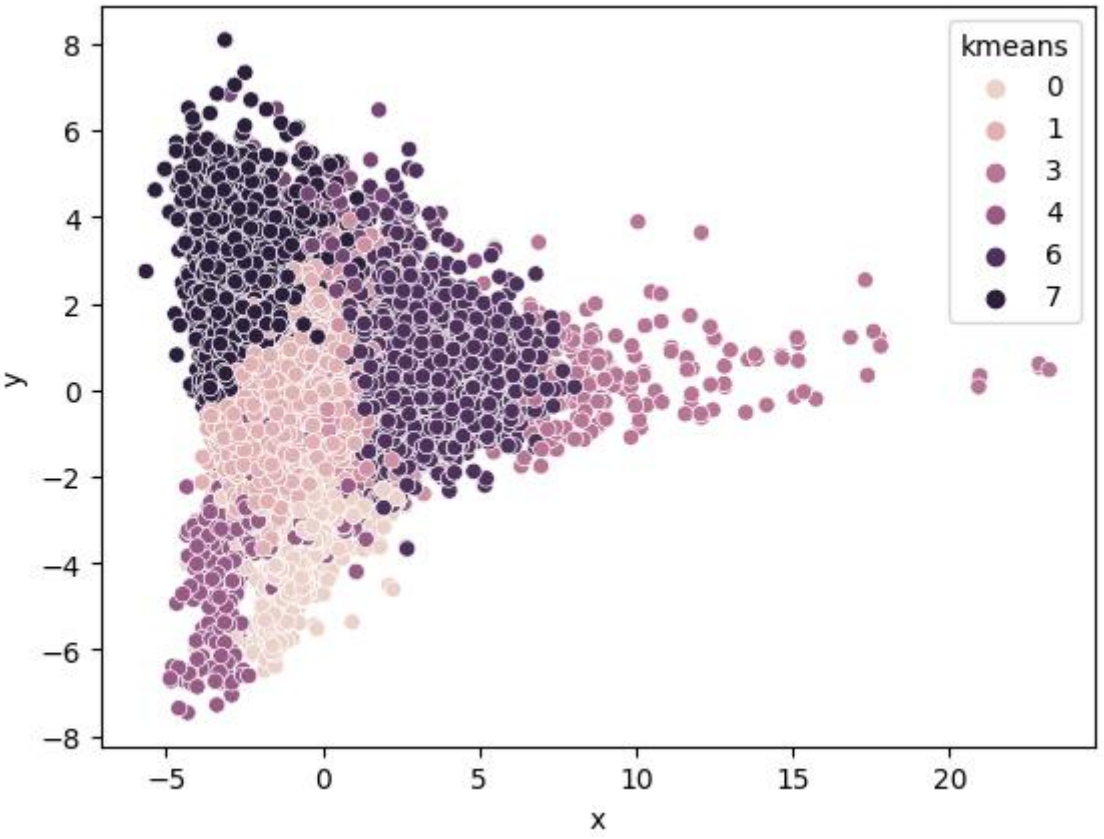

Рисунок 5 – визуализация признаков kmeans

Итого: можно сделать вывод, что при разбиении на кластеры, используя алгоритм kmeans, кластеры не будут совпадать с фазой луны.

DBSCAN

Гиперпараметры для обучения модели:

- eps = 0.3 (максимальное расстояние между кластерами)
- минимальное количество объектов в кластере - 3

Обучение происходило не на всех данных, а на фрагменте, содержащем 90% данных, т.к. вычислительных мощностей ноутбука не хватило для анализа всего датасета.

В результате обучения получились следующие результаты:

- Предполагаемое количество кластеров: 38
- Предполагаемое количество шумов: 16135

Если проанализировать таблицу со статистическими результатами,

представленную на рисунке, можно сделать вывод, что большинство точек алгоритм посчитал за выбросы, а остальные сгруппировал в 38 кластеров по 3-4 элемента в каждом. Скорее всего эту информацию нельзя как-то интерпретировать.

	air_quality_PM10								wind_mph		...	temperature_celsius		wind_degree						
	count	mean	std	min	25%	50%	75%	max	count	mean	...	75%	max	count	mean	std	min	25%	50%	75%	max
dbscan																					
-1.0	16135.0	46.707698	5.815968e+01	0.7	11.6	28.1	59.7	1043.7	16135.0	5.876207	...	27.0	38.3	16135.0	186.176883	97.781994	1.0	92.0	212.0	267.0	360.0
0.0	4.0	251.800000	0.000000e+00	251.8	251.8	251.8	251.8	251.8	4.0	2.200000	...	30.0	30.0	4.0	10.000000	0.000000	10.0	10.0	10.0	10.0	10.0
1.0	3.0	48.400000	0.000000e+00	48.4	48.4	48.4	48.4	48.4	3.0	4.300000	...	27.0	27.0	3.0	253.000000	0.000000	253.0	253.0	253.0	253.0	253.0
2.0	3.0	17.800000	0.000000e+00	17.8	17.8	17.8	17.8	17.8	3.0	5.100000	...	25.5	25.5	3.0	257.000000	0.000000	257.0	257.0	257.0	257.0	257.0
3.0	3.0	77.300000	0.000000e+00	77.3	77.3	77.3	77.3	77.3	3.0	2.200000	...	13.0	13.0	3.0	28.000000	0.000000	28.0	28.0	28.0	28.0	28.0
4.0	3.0	32.700000	0.000000e+00	32.7	32.7	32.7	32.7	32.7	3.0	6.300000	...	25.2	25.2	3.0	254.000000	0.000000	254.0	254.0	254.0	254.0	254.0
5.0	3.0	14.300000	2.175584e-15	14.3	14.3	14.3	14.3	14.3	3.0	2.200000	...	19.7	19.7	3.0	69.000000	0.000000	69.0	69.0	69.0	69.0	69.0
6.0	3.0	50.200000	8.702336e-15	50.2	50.2	50.2	50.2	50.2	3.0	2.200000	...	13.6	13.6	3.0	12.000000	0.000000	12.0	12.0	12.0	12.0	12.0
7.0	3.0	3.900000	0.000000e+00	3.9	3.9	3.9	3.9	3.9	3.0	2.200000	...	21.0	21.0	3.0	331.000000	0.000000	331.0	331.0	331.0	331.0	331.0
8.0	3.0	55.700000	8.702336e-15	55.7	55.7	55.7	55.7	55.7	3.0	3.800000	...	28.0	28.0	3.0	180.000000	0.000000	180.0	180.0	180.0	180.0	180.0
9.0	3.0	67.400000	0.000000e+00	67.4	67.4	67.4	67.4	67.4	3.0	2.200000	...	13.4	13.4	3.0	343.000000	0.000000	343.0	343.0	343.0	343.0	343.0
10.0	3.0	4.100000	0.000000e+00	4.1	4.1	4.1	4.1	4.1	3.0	2.200000	...	21.3	21.3	3.0	350.000000	0.000000	350.0	350.0	350.0	350.0	350.0
11.0	3.0	97.600000	1.740467e-14	97.6	97.6	97.6	97.6	97.6	3.0	2.200000	...	14.2	14.2	3.0	22.000000	0.000000	22.0	22.0	22.0	22.0	22.0
12.0	3.0	15.800000	2.175584e-15	15.8	15.8	15.8	15.8	15.8	3.0	5.800000	...	24.4	24.4	3.0	271.000000	0.000000	271.0	271.0	271.0	271.0	271.0
13.0	3.0	15.500000	0.000000e+00	15.5	15.5	15.5	15.5	15.5	3.0	2.200000	...	19.8	19.8	3.0	77.000000	0.000000	77.0	77.0	77.0	77.0	77.0
14.0	3.0	20.500000	0.000000e+00	20.5	20.5	20.5	20.5	20.5	3.0	2.200000	...	21.3	21.3	3.0	17.000000	0.000000	17.0	17.0	17.0	17.0	17.0
15.0	3.0	43.800000	8.702336e-15	43.8	43.8	43.8	43.8	43.8	3.0	7.600000	...	25.0	25.0	3.0	236.000000	0.000000	236.0	236.0	236.0	236.0	236.0
16.0	3.0	7.400000	1.087792e-15	7.4	7.4	7.4	7.4	7.4	3.0	2.200000	...	20.1	20.1	3.0	342.000000	0.000000	342.0	342.0	342.0	342.0	342.0
17.0	3.0	12.700000	2.175584e-15	12.7	12.7	12.7	12.7	12.7	3.0	2.200000	...	20.0	20.0	3.0	8.000000	0.000000	8.0	8.0	8.0	8.0	8.0

Оценка по метрикам:

```
homogeneity_score = 0.06864011631273441
completeness_score = 0.19783611673693718
v_measure_score = 0.10191898848368838
adjusted_rand_score = 4.799107886884161e-05
adjusted_mutual_info_score = 0.053418383803754944
silhouette_score = -0.6322164301771409
```

Визуализация результатов

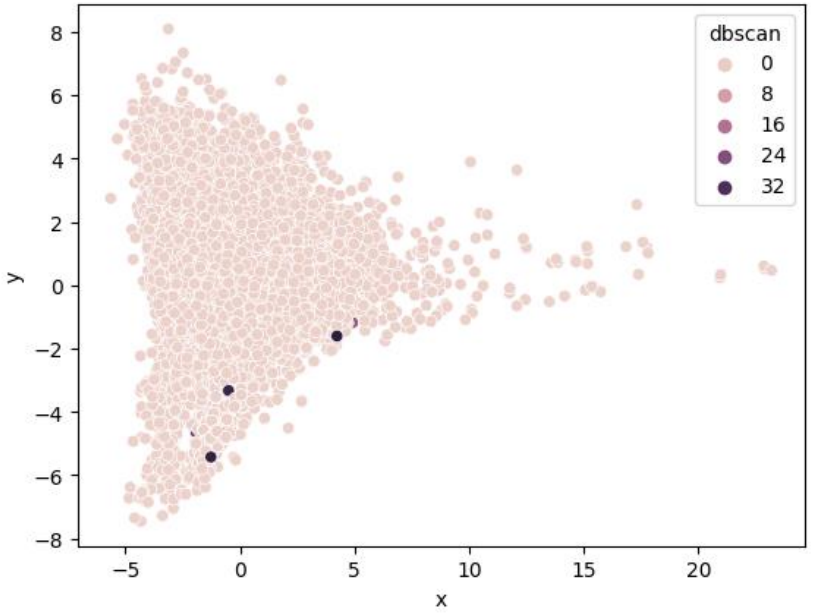

Рисунок 6 – визуализация признаков DBSCAN

Иерархическая кластеризация

Гиперпараметры для обучения модели:

• количество кластеров = 8 (по количеству фаз луны);

• способ измерения расстояния между центроидами – метод Варда.

Основные статистические данные по каждому признаку кластера представлены на рисунке.

	air_quality_PM10								wind_mph		...	temperature_celsius		wind_degree						
	count	mean	std	min	25%	50%	75%	max	count	mean	...	75%	max	count	mean	std	min	25%	50%	75%	max
hierarh																				
0	2009.0	174.134893	110.034818	21.2	105.2	154.60	213.800	1043.7	2009.0	3.808661	...	27.10	34.2	2009.0	147.454455	113.913489	1.0	50.0	109.0	259.0	360.0
1	3075.0	24.202699	24.905755	0.8	7.4	15.70	32.200	205.1	3075.0	3.711089	...	26.00	32.0	3075.0	157.582114	97.619033	1.0	75.0	147.0	240.0	360.0
2	483.0	20.210145	19.089807	1.1	6.0	15.00	26.950	158.4	483.0	7.024845	...	25.05	31.0	483.0	189.650104	101.374085	1.0	86.0	233.0	270.0	360.0
3	1528.0	61.584228	56.060811	1.4	20.5	42.25	91.225	323.7	1528.0	6.125589	...	30.50	35.3	1528.0	229.008508	93.695262	2.0	179.0	263.0	298.0	360.0
4	890.0	50.238539	33.034068	0.7	23.7	47.70	72.800	175.3	890.0	3.641011	...	16.80	26.3	890.0	113.066292	102.453474	1.0	45.0	68.0	138.0	359.0
5	3743.0	67.143708	42.922438	3.1	33.4	57.30	90.100	257.1	3743.0	5.115469	...	27.20	38.3	3743.0	140.136789	84.934958	1.0	77.0	120.0	196.0	359.0
6	3128.0	29.479795	21.499206	0.8	13.9	25.40	40.425	168.4	3128.0	5.080850	...	26.40	33.1	3128.0	234.833760	84.435039	1.0	216.0	256.0	286.0	360.0
7	3199.0	13.893435	14.457656	0.7	4.3	9.20	18.500	159.1	3199.0	10.210660	...	26.10	32.0	3199.0	238.336668	56.007449	10.0	226.0	253.0	270.0	355.0

Распределение, также как и в kmeans, по количеству объектов в кластере похоже на исходные данные.

Результаты по метрикам:

```
homogeneity_score = 0.07419540886762205
completeness_score = 0.08707658496614461
v_measure_score = 0.08012157189572255
adjusted_rand_score = 0.0335980718588023
adjusted_mutual_info_score = 0.07931116677641882
silhouette_score = -0.04882367098256186
```

Визуализация

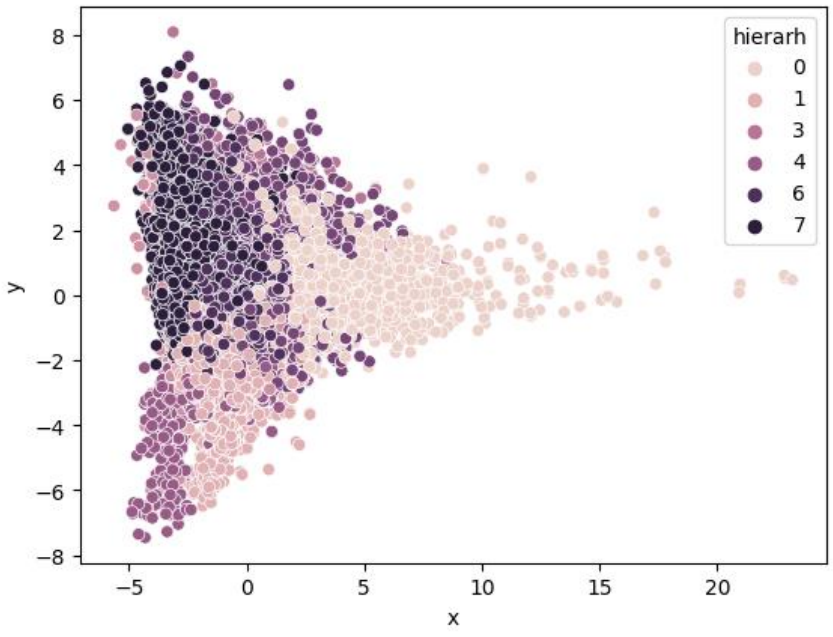

Рисунок 7 – визуализация признаков иерархической кластеризации

Выводы

В результате выполнения работы был проведен кластерный анализ данных о погоде в Индии. Анализ показал, что погода в Индии и фаза луны в день измерений не имеют зависимостей. Кроме того, метод DBSCAN оказался вовсе неприменим для этого набора данных, так как его результаты больше относятся к выбросам.