Мартынов Д.А.

МПИ-20-4-2

Лабораторная работа 3

В работе производится сравнение трех методов кластерного анализа на разных датасетах. В качестве первого датасета были выбраны Ирисы Фишера, второй датасет генерировался автоматически со следующим набором параметров: *количество элементов, количество признаков, количество классов, стандартное отклонение*. Алгоритмы кластеризации запускались на сгенерированном датасете с разным набором параметров для реализации следующих сценариев запуска:

* линейно разделимые множества (с расстоянием между группами в 10^3 раз больше, чем диаметр группы)
* линейно разделимые множества (группы расположены близко или касаются друг друга)
* линейно неразделимое множество (средняя площадь пересечения классов 10-20%)
* линейно неразделимое множество (средняя площадь пересечения классов 50-70%)

Дендрограммы строятся для агломеративного (иерархического) метода. На них изображены расстояния между объединяемыми кластерами на каждом этапе работы алгоритма. Позволяют определить, каким должен быть параметр threshold, чтобы правильно установить количество кластеров, если таковое не задано изначально. Если расстояние между объединенными кластерами оказалось большим, то, скорее всего, это были 2 разных кластера.

В первой части отчета приведены графики только для некоторых сценариев запуска, чтобы не показывать почти одинаковые результаты, разность между которыми на графике увидеть не представляется возможным.

Сравнительная таблица с метриками качества для всех сценариев запуска приведена во второй части отчета.

***adjusted\_rand\_score –*** Индекс adjusted\_rand оценивает, насколько много из тех пар элементов, которые находились в одном классе, и тех пар элементов, которые находились в разных классах, сохранили это состояние после кластеризации алгоритмом.

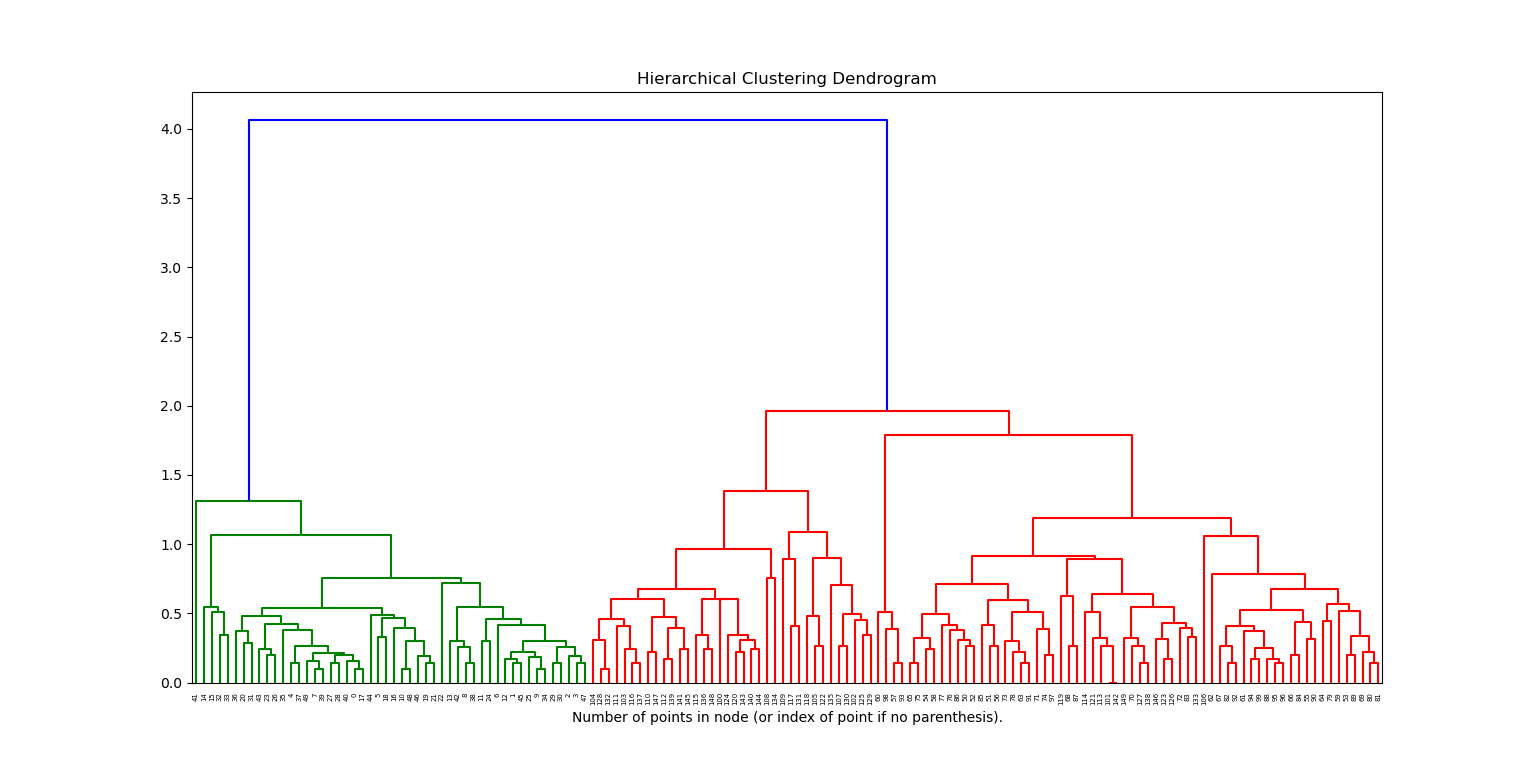
***jaccard\_score -*** Индекс Жаккара похож на Индекс Rand, только не учитывает пары элементов находящиеся в разные классах и разных кластерах.

***fowlkes\_mallows\_score -*** Индекс Фоулкса – Мэллова используется для определения сходства между двумя кластерами.

***f1\_score -*** F-мера представляет собой гармоническое среднее между точностью (precision) и полнотой (recall).

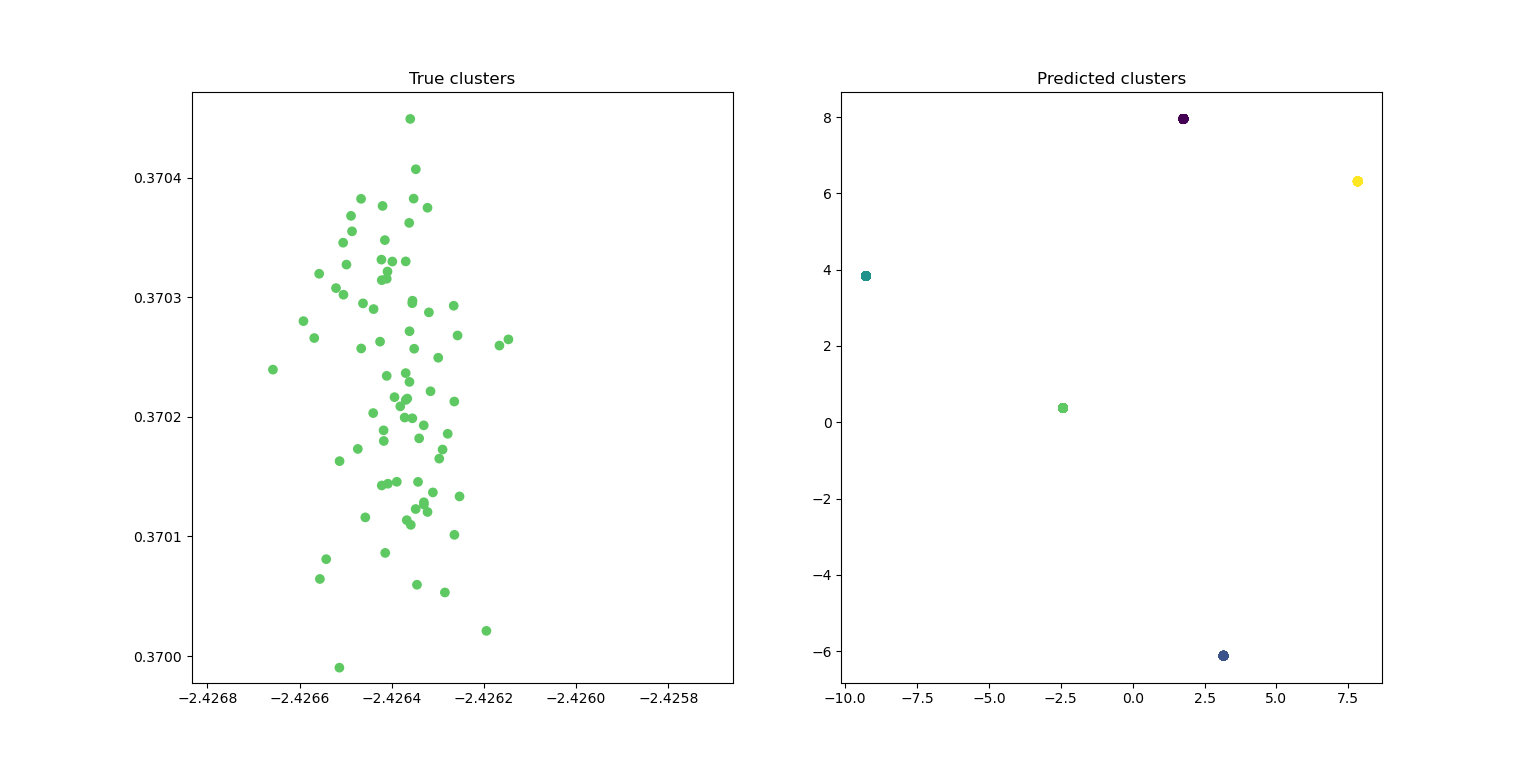
**Датасет Iris**

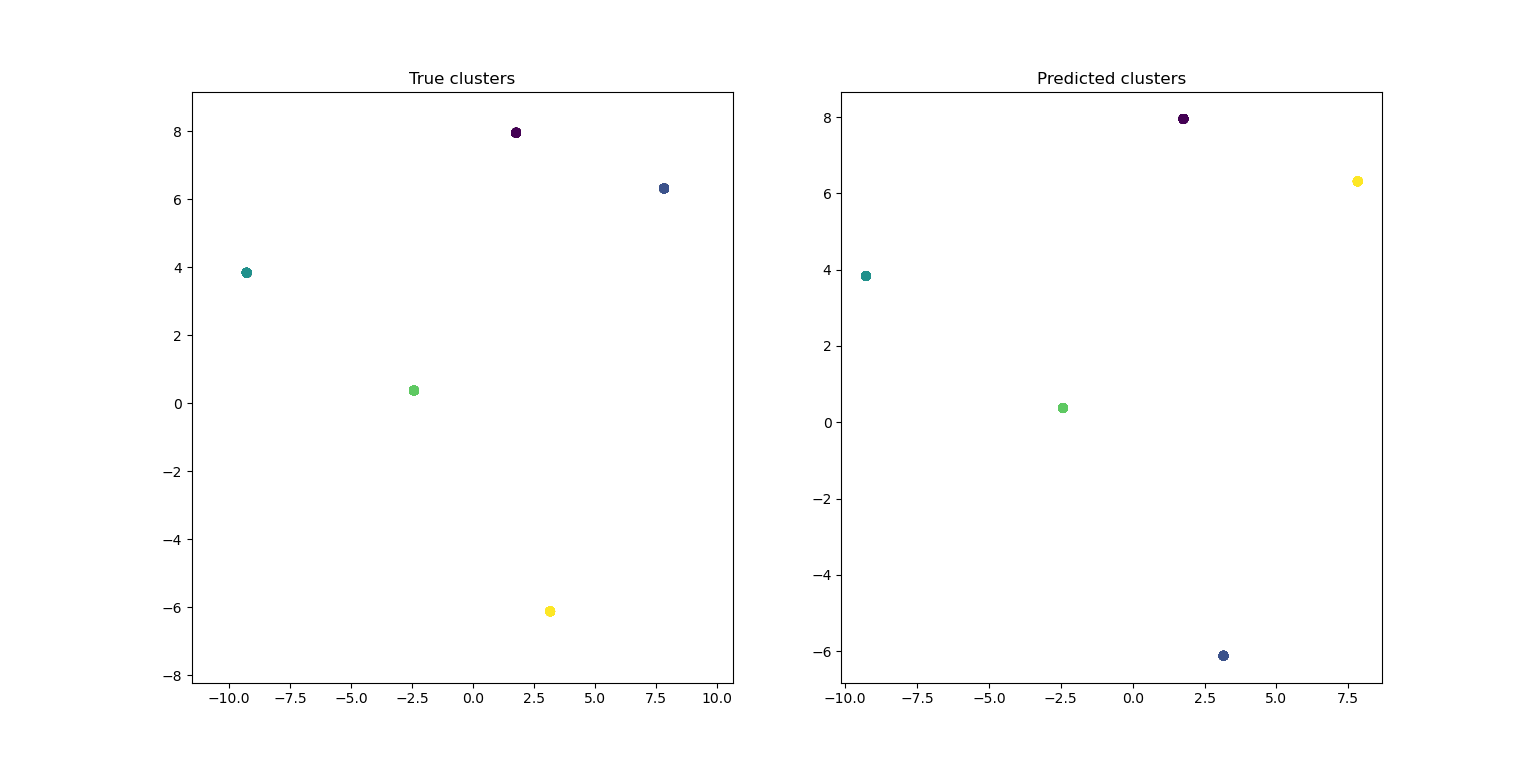
Дендрограмма для алгомеративного метода с евклидовой метрикой



**Сгенерированный датасет**

1. Расстояние между кластерами

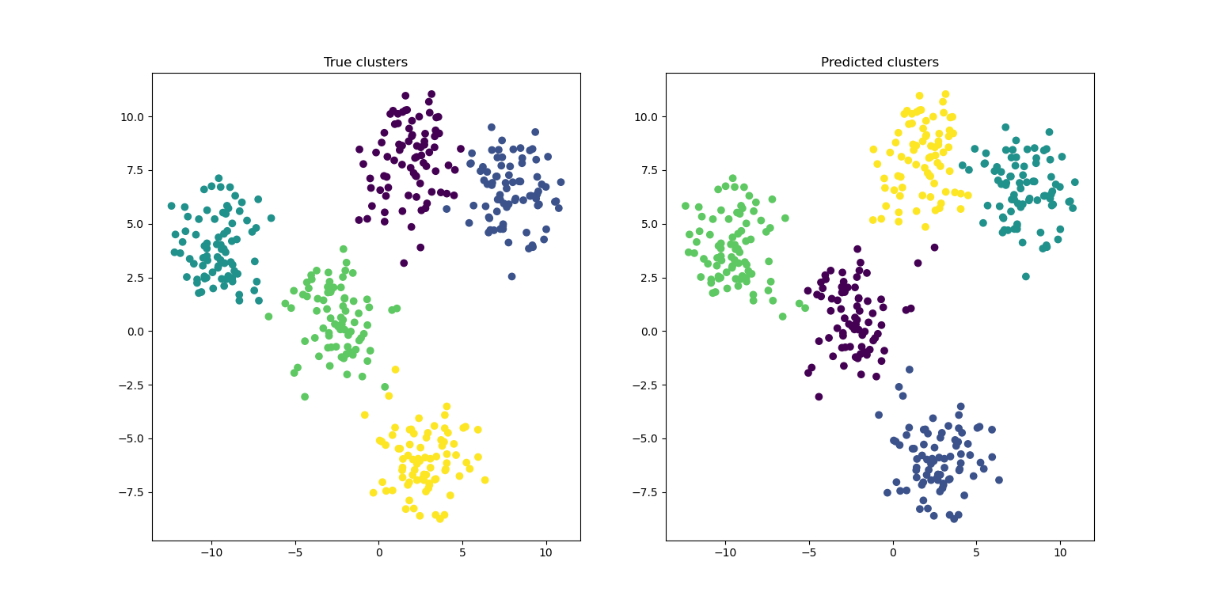


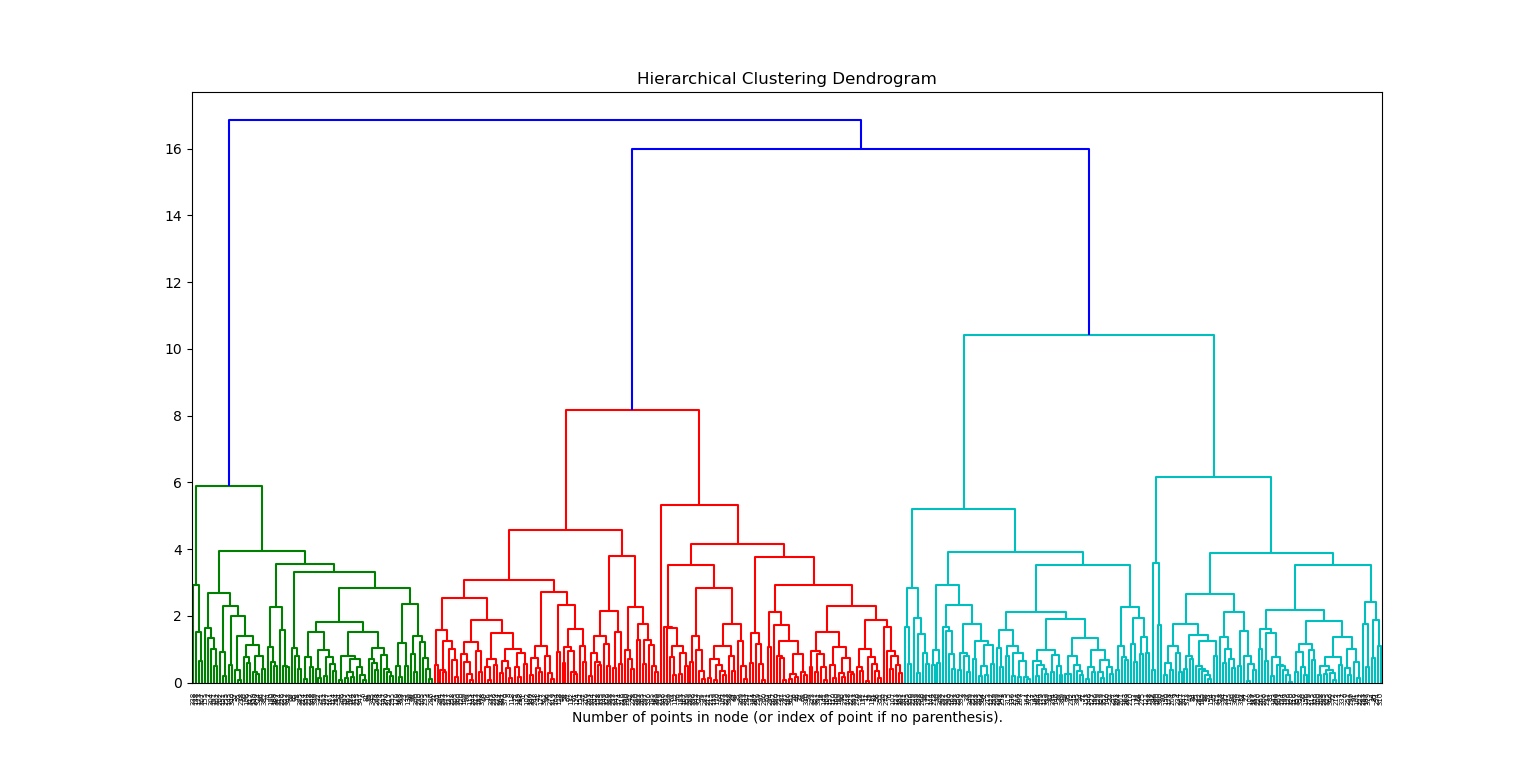


Результаты не сильно разнятся, но агломеративный метод показал себя чуть лучше. Приведенные графики демонстрируют соответствие требованию «расстояние между кластерами должно быть в раз больше диаметра кластера» и кластеризацию при помощи k-means соответственно.

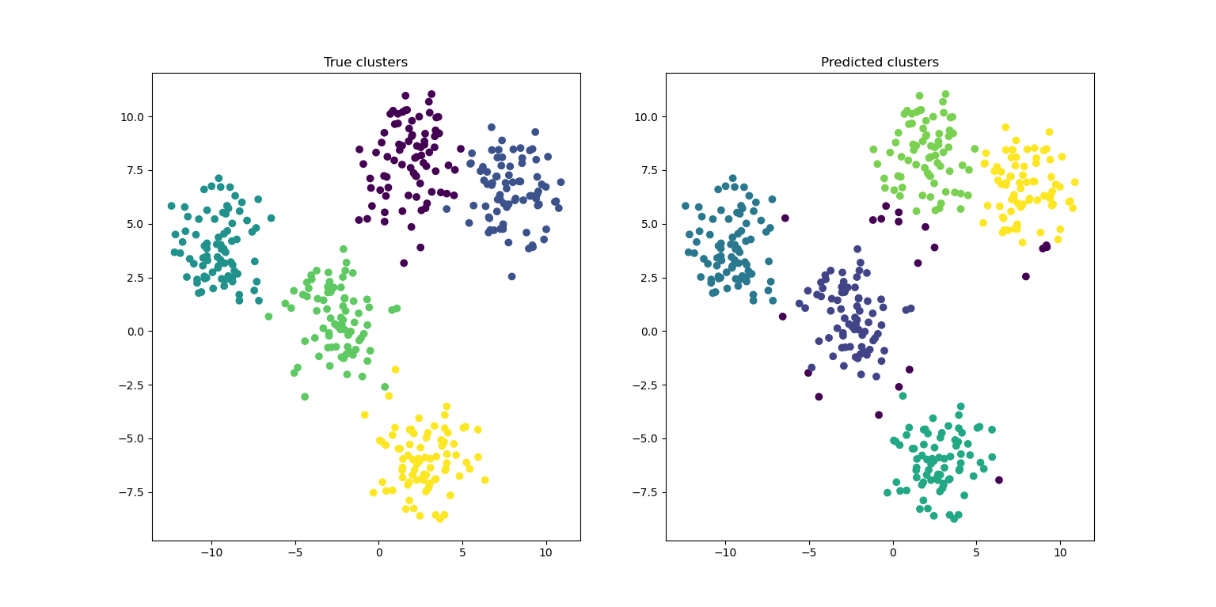
1. Кластеры линейно разделимы, расстояние минимально

На графиках результаты для агломеративного метода с евклидовой метрикой.





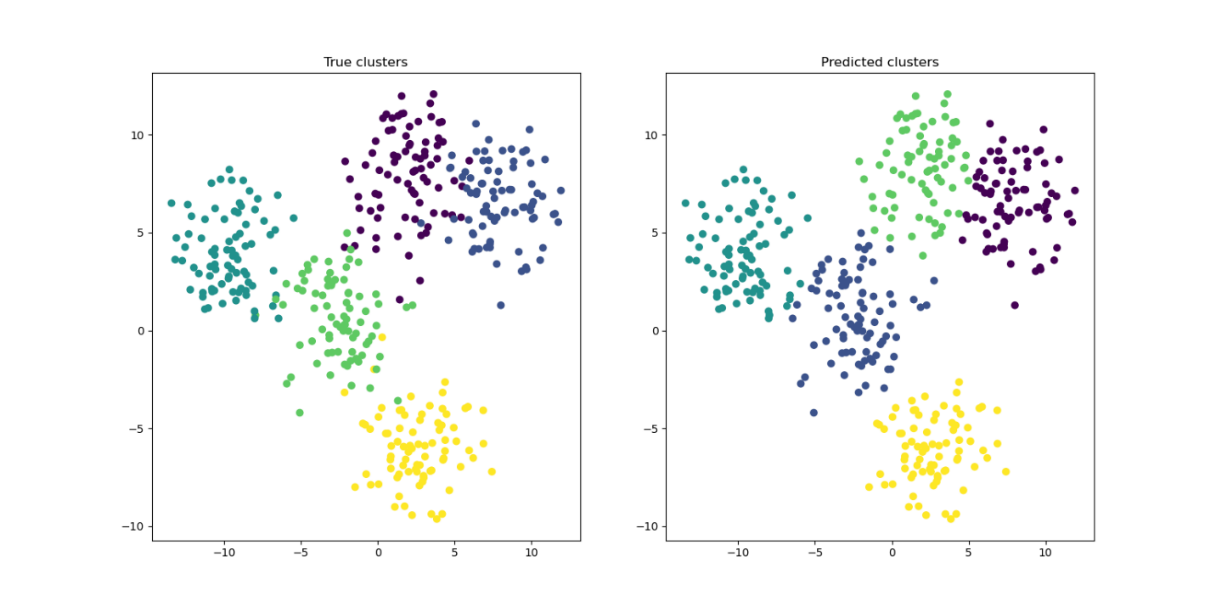
И для DBSCAN с параметрами (min\_samples=30, eps=2.03)

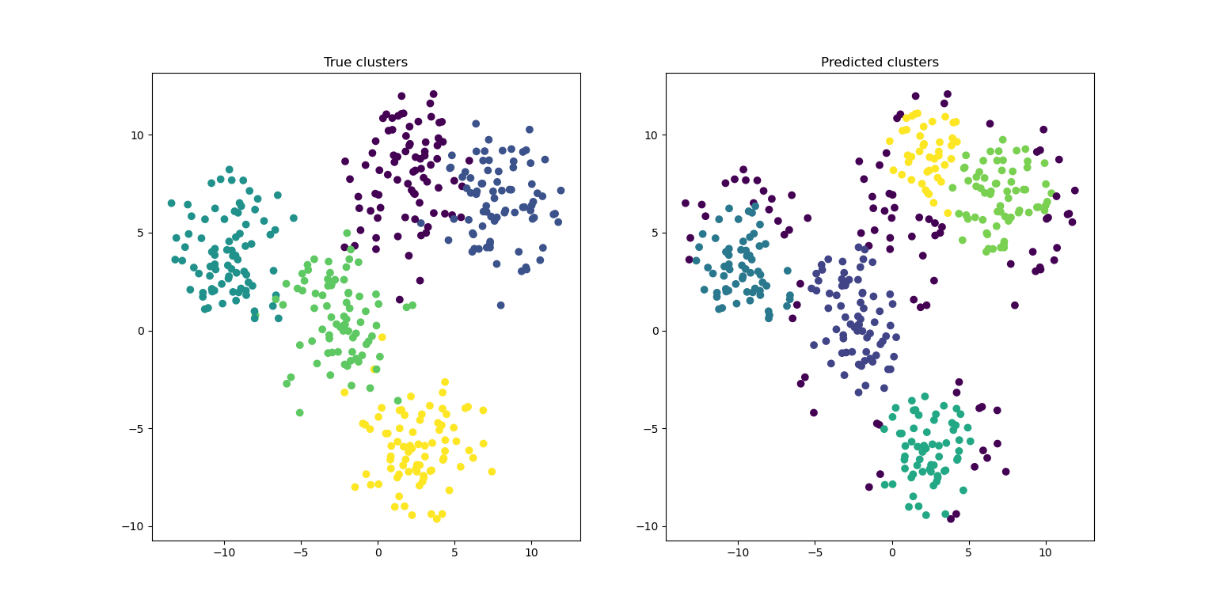


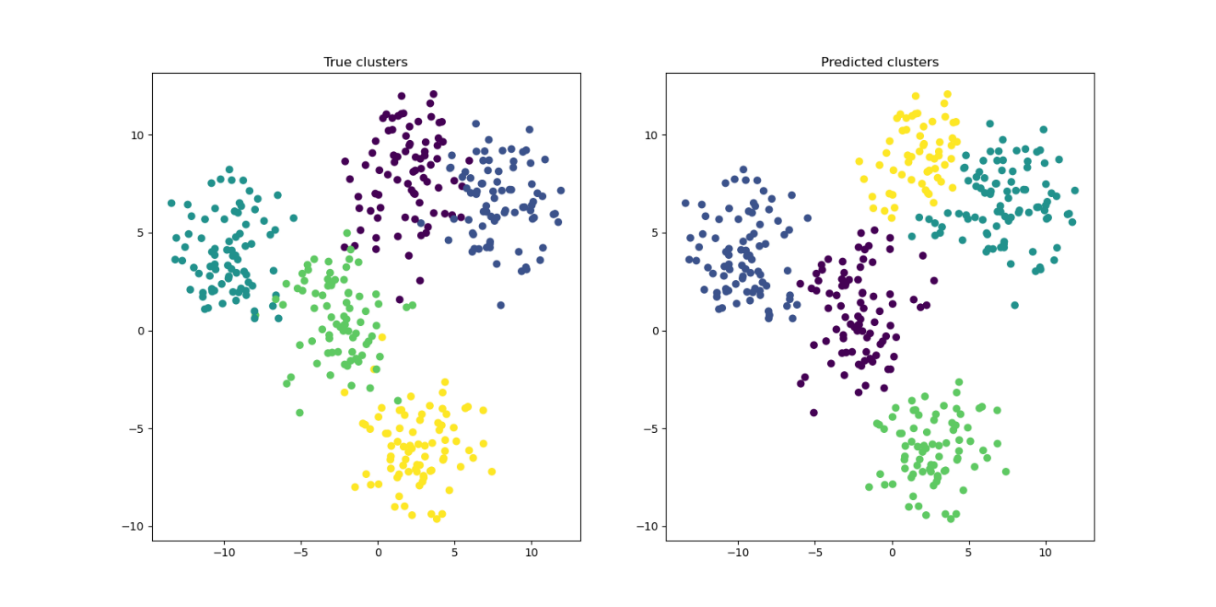
1. Кластеры линейно неразделимы, пересечение 10-20%

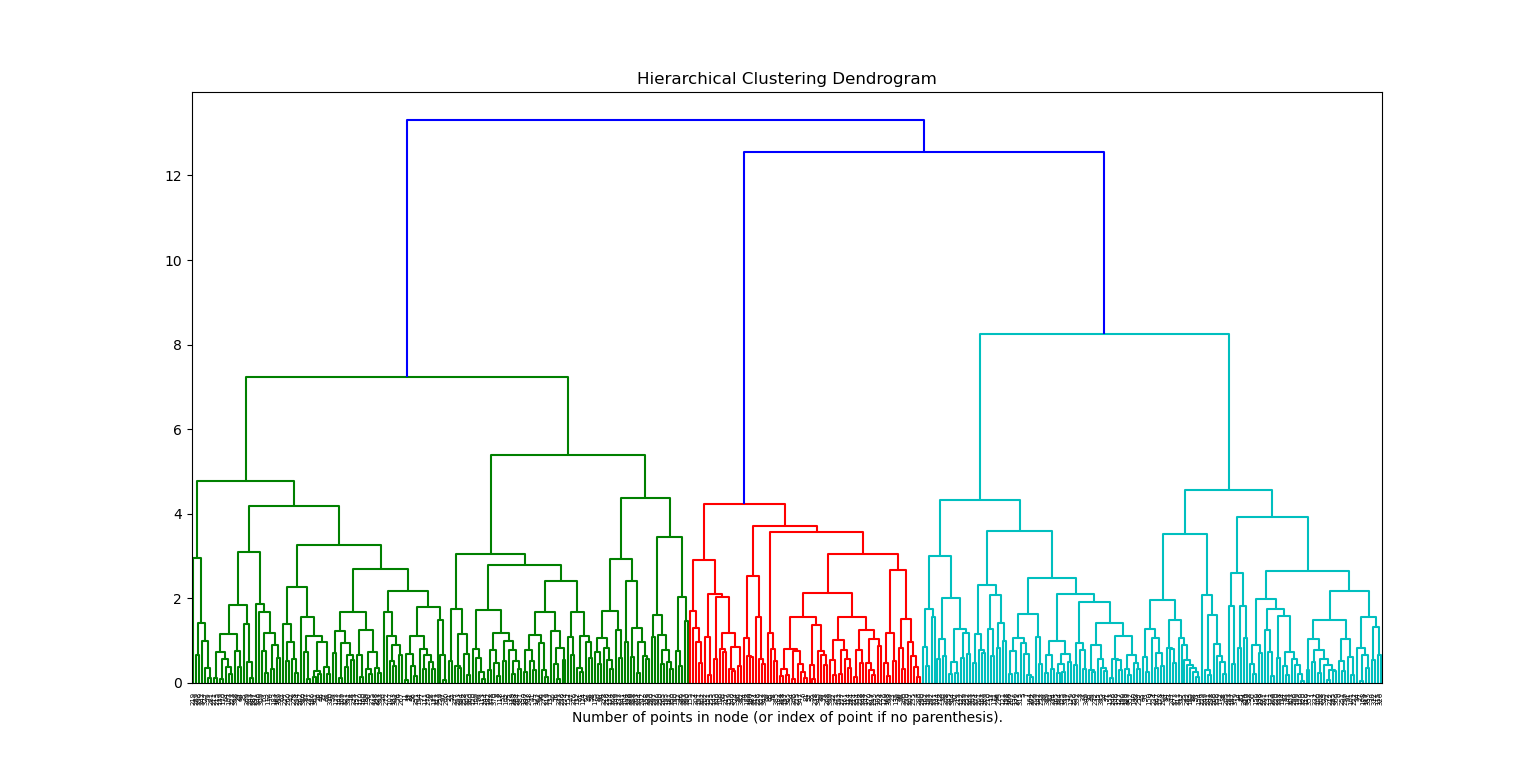
Для DBSCAN лучший результат для этого датасета был получен со следующими параметрами (min\_samples=28, eps=2.02). Уменьшение min\_samples ведет к тому, что метод теряет один из кластеров, объединяя два в один. Значения больше 30 приводят к тому же результату. Eps ниже 2 и больше 2.2 ухудшает результат, в конечном итоге приводя к тому, что метод совсем не может разделить элементы датасета, выдавая все данные за мусор (noise).

На графиках: результаты для k-means, DBSCAN, агломеративного метода с евклидовой метрикой, дендрограмма для метода с евклидовой метрикой.





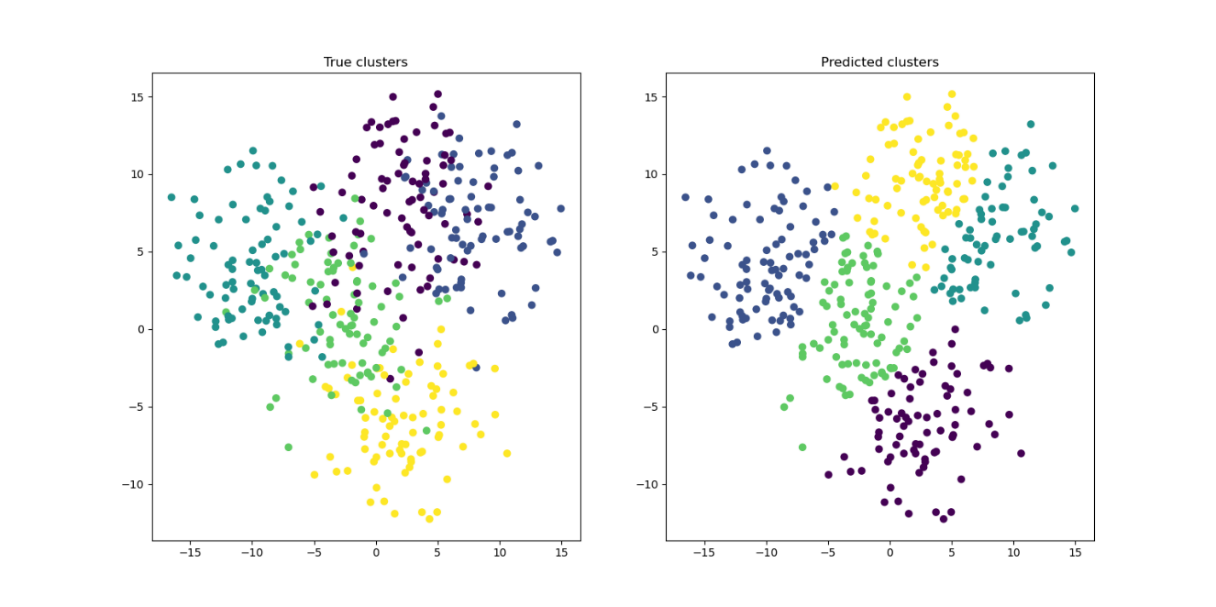


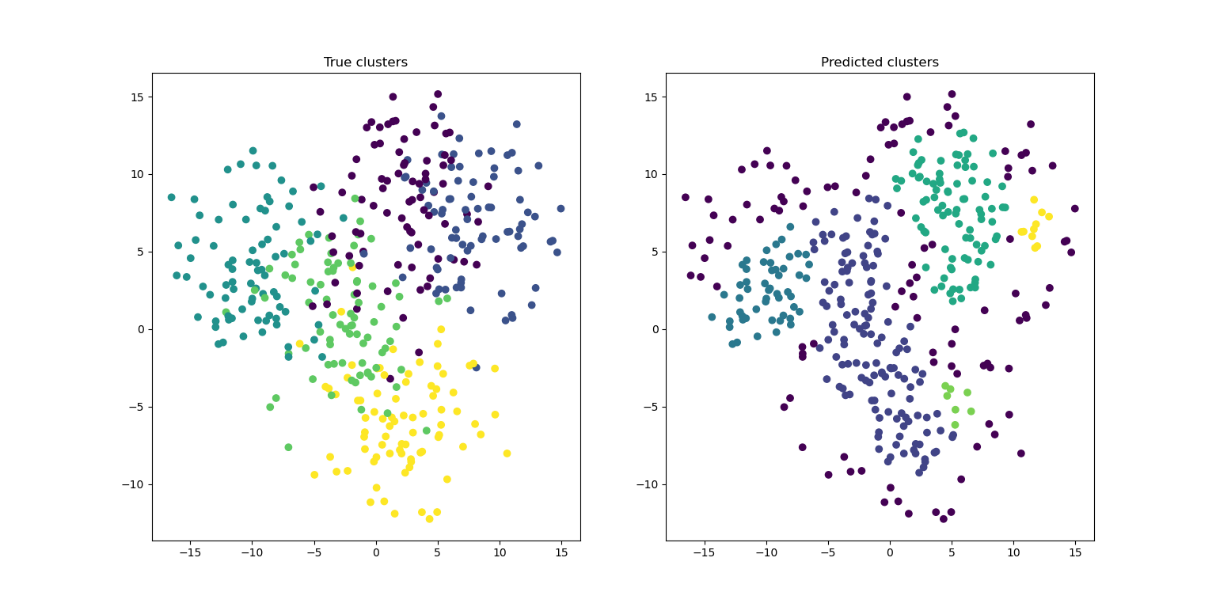


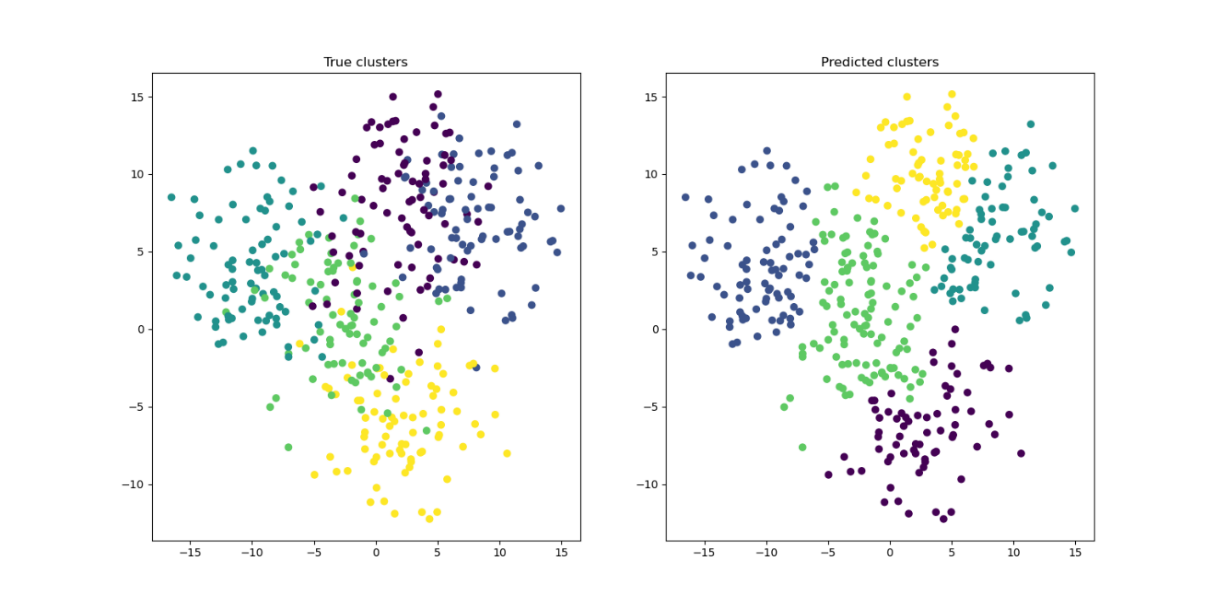
1. Кластеры линейно неразделимы, пересечение 50-70%

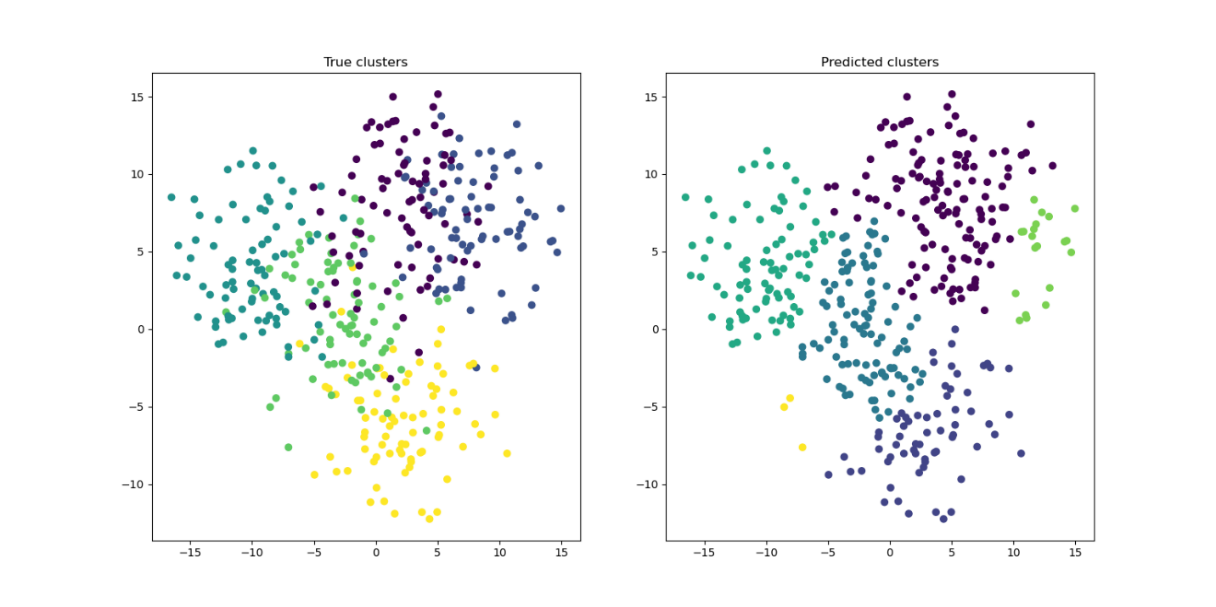
На дендрограмме для метода с манхэттенской метрикой отчетливее видны кластеры, как их разделяет алгоритм, но результат такой же плохой, как у метода с евклидовой метрикой. Для DBSCAN параметры запуска такие: (min\_samples=10, eps=1.8). С ними метод показал хоть какие-то результаты. Значения параметра eps ниже 1.5 приводили к тому, что метод не определял ни одного кластера (весь датасет – шум); аналогично для значений 2 и выше. Значение параметра min\_samples варьировалось от 10 до 20 с разными eps, но самый удачный результат получился для 10.

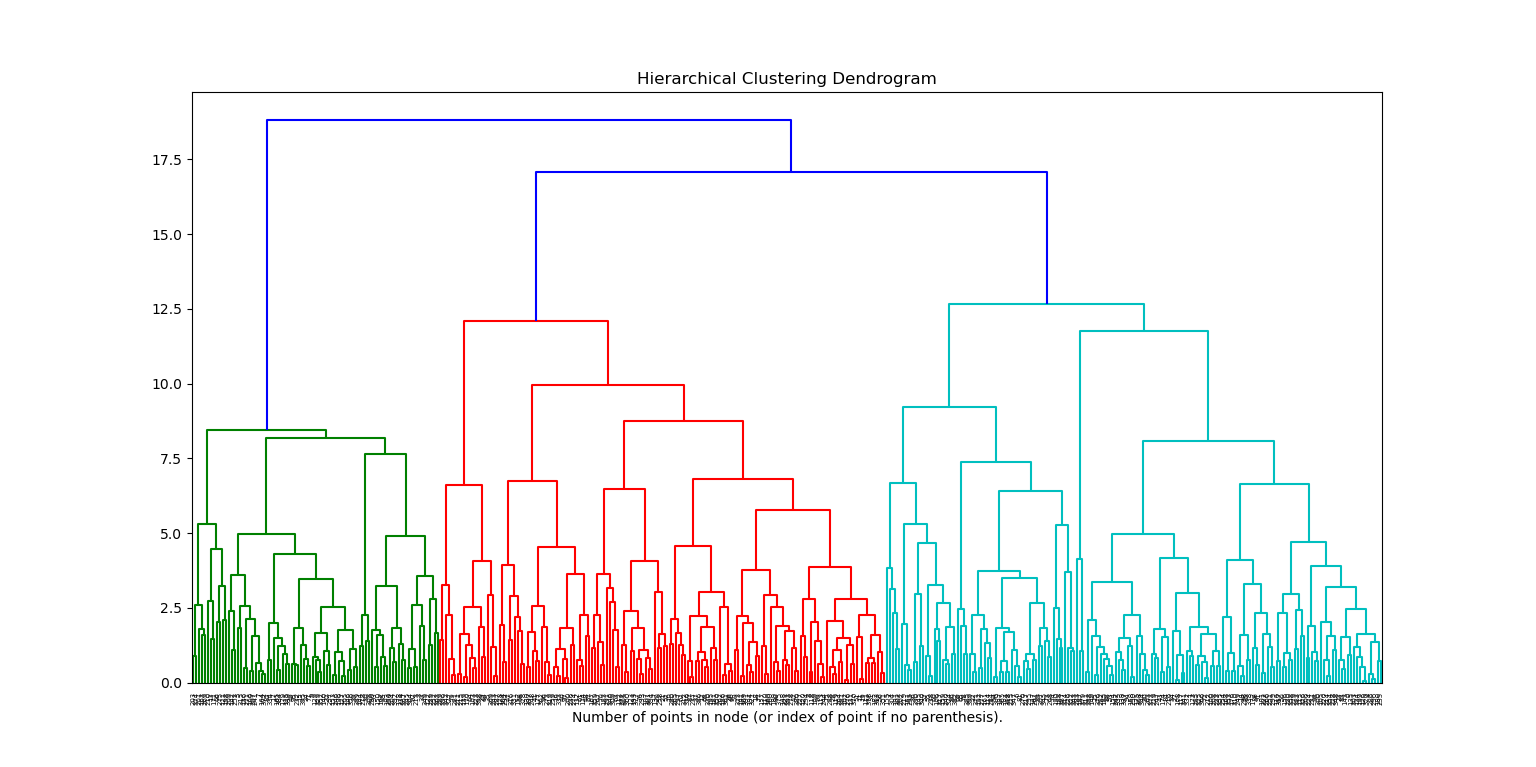
На графиках: результаты для k-means, DBSCAN, агломеративного метода с евклидовой метрикой, агломеративного метода с манхэттенской метрикой и дендрограммы для методов с обеими метриками.

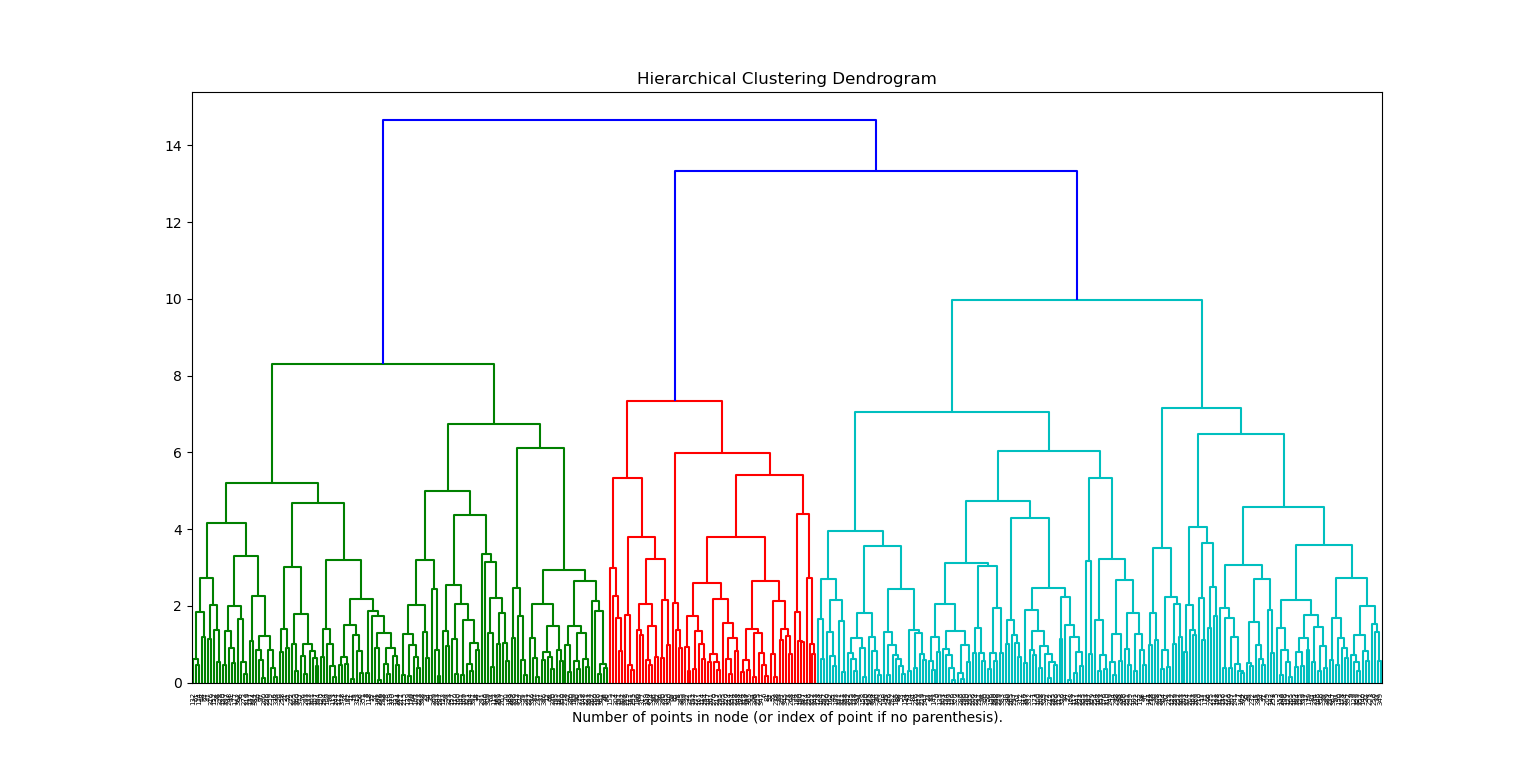












Сравнительная таблица с метриками качества для всех сценариев запуска

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Методы | k-means | Agglomerative (euclidean) | Agglomerative (manhattan) | DBSCAN |
| IRIS | ***adjusted\_rand\_score:*** *0.7302382722834697*  ***jaccard\_score:*** *0.13636363636363635*  ***fowlkes\_mallows\_score:*** *0.8208080729114153*  ***f1\_score:***  *0.24* | ***adjusted\_rand\_score:*** *0.7591987071071522*  ***jaccard\_score:*** *0.13636363636363635*  ***fowlkes\_mallows\_score:*** *0.8407289157574822*  ***f1\_score:***  *0.24* | ***adjusted\_rand\_score:*** *0.744526430873875*  ***jaccard\_score:*** *0.13636363636363635*  ***fowlkes\_mallows\_score:*** *0.8306394195437843*  ***f1\_score:***  *0.24* | ***Clusters:*** *2*  ***Noise points:*** *4*  ***adjusted\_rand\_score:*** *0.5609099839129877*  ***jaccard\_score:***  *0.5*  ***fowlkes\_mallows\_score:*** *0.7567359993078075*  ***f1\_score:*** *0.6666666666666666* |
| Generated dataset | | | |  |
| 10^3 | ***adjusted\_rand\_score:***  *1.0*  ***jaccard\_score:*** *0.42857142857142855*  ***fowlkes\_mallows\_score:*** *1.0*  ***f1\_score:***  *0.6* | ***adjusted\_rand\_score:***  *1.0*  ***jaccard\_score:*** *0.1111111111111111*  ***fowlkes\_mallows\_score:***  *1.0*  ***f1\_score:***  *0.20000000000000004* | ***adjusted\_rand\_score****:*  *1.0*  ***jaccard\_score:***  *0.0*  ***fowlkes\_mallows\_score:*** *1.0*  ***f1\_score:***  *0.0* | ***Clusters:*** *5*  ***Noise points:*** *0*  ***adjusted\_rand\_score:***  *1.0*  ***jaccard\_score:*** *0.1111111111111111*  ***fowlkes\_mallows\_score****:*  *1.0*  ***f1\_score:*** *0.20000000000000004* |
| Рядом | ***adjusted\_rand\_score:*** *0.9750236371937734*  ***jaccard\_score:*** *0.24610591900311526*  ***fowlkes\_mallows\_score:*** *0.9799689888138144*  ***f1\_score:***  *0.395* | ***adjusted\_rand\_score:*** *0.9631689856740275*  ***jaccard\_score:*** *0.006289308176100629*  ***fowlkes\_mallows\_score:*** *0.9704676649524663*  ***f1\_score:*** *0.012500000000000002* | ***adjusted\_rand\_score:*** *0.938825319686041*  ***jaccard\_score:*** *0.006289308176100629*  ***fowlkes\_mallows\_score:*** *0.950944249806295*  ***f1\_score:*** *0.012500000000000002* | ***Clusters:*** *5*  ***Noise points:*** *20*  ***adjusted\_rand\_score:*** *0.9271035543316661*  ***jaccard\_score:***  *0.0*  ***fowlkes\_mallows\_score:*** *0.9418309363575439*  ***f1\_score:***  *0.0* |
| 10% | ***adjusted\_rand\_score:*** *0.8742551256849869*  ***jaccard\_score:***  *0.25*  ***fowlkes\_mallows\_score:*** *0.8991751858838986*  ***f1\_score:*** *0.4000000000000001* | ***adjusted\_rand\_score:*** *0.8527609671350769*  ***jaccard\_score:*** *0.24031007751937986*  ***fowlkes\_mallows\_score:*** *0.8819501627503695*  ***f1\_score:***  *0.3875* | ***adjusted\_rand\_score:*** *0.8395854164671829*  ***jaccard\_score:*** *0.011378002528445006*  ***fowlkes\_mallows\_score:*** *0.8717291099203036*  ***f1\_score:***  *0.0225* | ***Clusters:*** *5*  ***Noise points:*** *94*  ***adjusted\_rand\_score:*** *0.5631572268060075*  ***jaccard\_score:*** *0.0012515644555694619*  ***fowlkes\_mallows\_score:*** *0.6453256343713013*  ***f1\_score:***  *0.0025* |
| 50% | ***adjusted\_rand\_score:*** *0.5381089892432729*  ***jaccard\_score:*** *0.08108108108108109*  ***fowlkes\_mallows\_score:*** *0.6298762410884104*  ***f1\_score:***  *0.15* | ***adjusted\_rand\_score:*** *0.5180856601634966*  ***jaccard\_score:*** *0.08843537414965986*  ***fowlkes\_mallows\_score:*** *0.6154882577940028*  ***f1\_score:***  *0.1625* | ***adjusted\_rand\_score:*** *0.48264175823084043*  ***jaccard\_score:*** *0.11576011157601115*  ***fowlkes\_mallows\_score:*** *0.5968062322409557*  ***f1\_score:***  *0.2075* | ***Clusters:*** *5*  ***Noise points:*** *108*  ***adjusted\_rand\_score:*** *0.23776721187890354*  ***jaccard\_score:*** *0.02040816326530612*  ***fowlkes\_mallows\_score:*** *0.4094296217835113*  ***f1\_score:***  *0.04* |

**Еще 2 теста на других датасетах**

Судя по результатам проведенных тестов, DBSCAN хорошо подходит для датасетов, в которых нет пересечения кластеров. При этом кластеры могут быть разной формы. Для проверки сгенерируем 2 месяца и попробуем разделить их на 2 кластера.

***Clusters:*** *2*

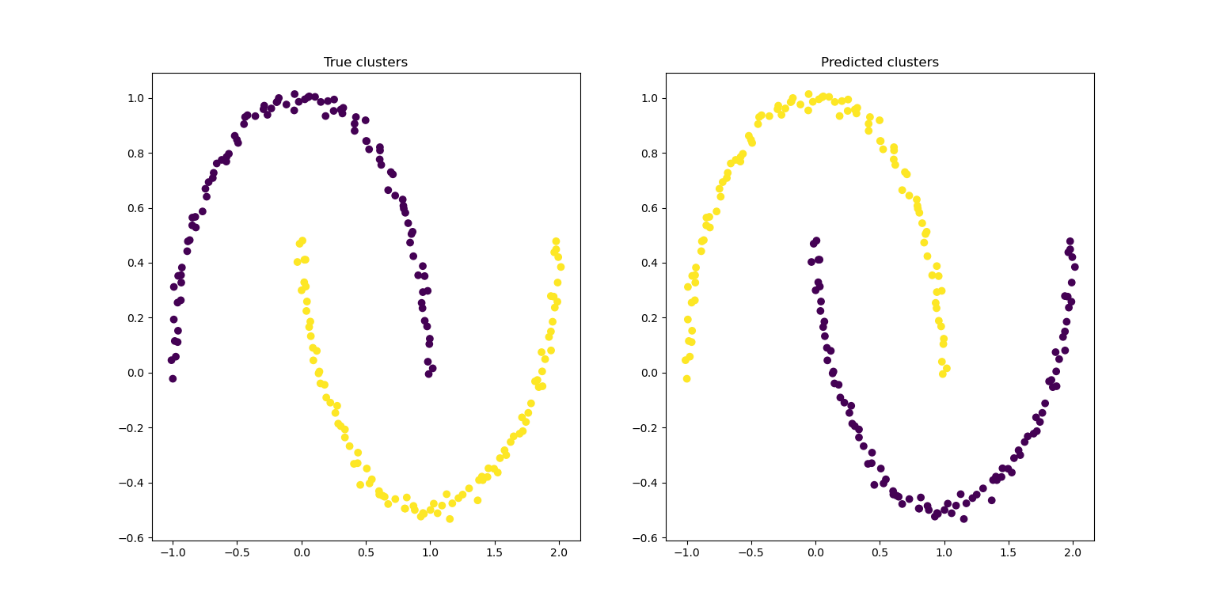
***Noise points:*** *0*

***adjusted\_rand\_score:*** *1.0*

***jaccard\_score:*** *0.0*

***fowlkes\_mallows\_score:*** *1.0*

***f1\_score:*** *0.0*



200 элементов

***Clusters:*** *2*

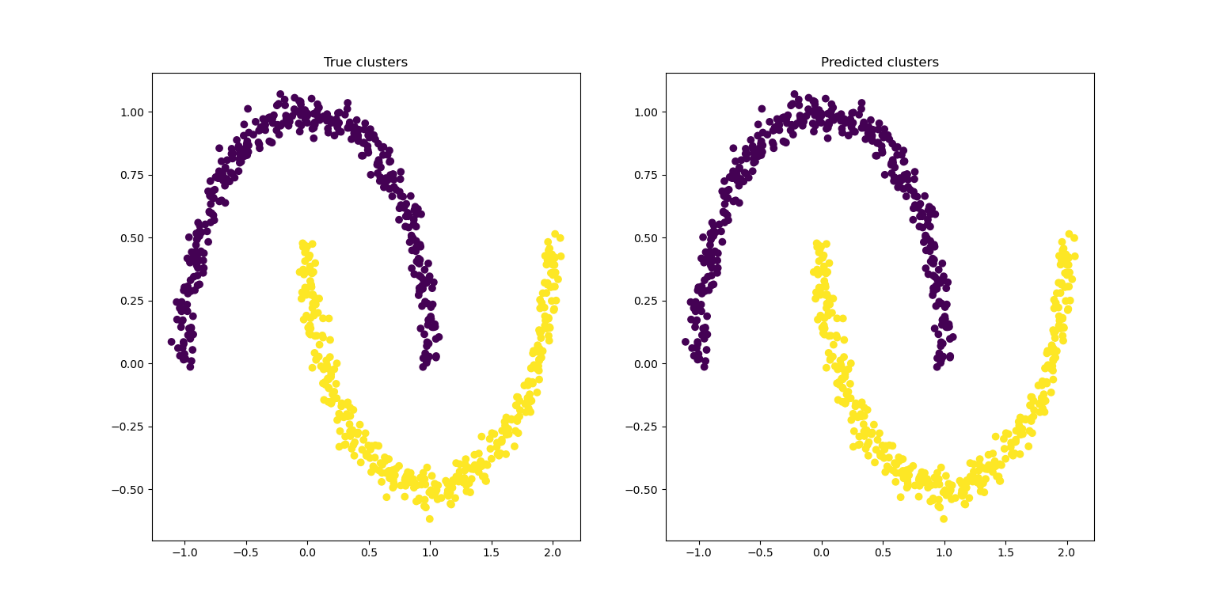
***Noise points:*** *0*

***adjusted\_rand\_score:*** *1.0*

***jaccard\_score:*** *1.0*

***fowlkes\_mallows\_score:*** *1.0*

***f1\_score:*** *1.0*



750 элементов

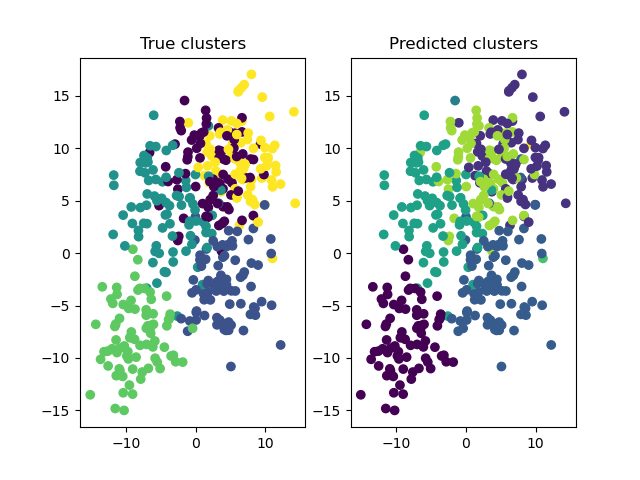
Пример с датасетом с 7 признаками. На этом датасете DBSCAN не смог выделить ни один кластер ни в одном запуске. Запускался он с евклидовой и манхэттенской метриками, при этом варьировались параметры. Агломеративный же метод смог выделить кластеры. Для него получилось построить очень хорошую дендрограмму, а также получить приемлемые результаты.

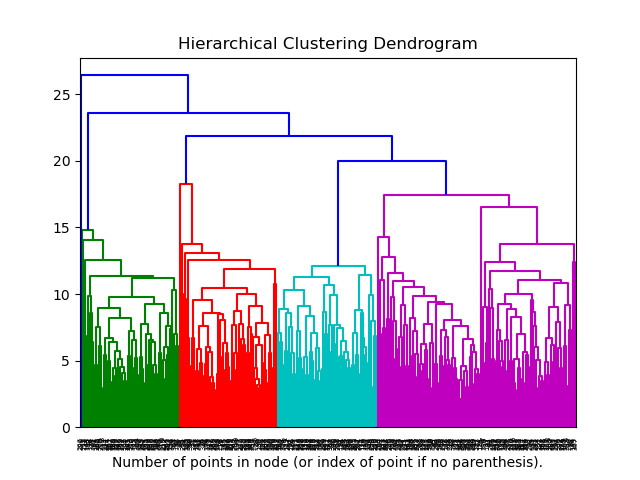
***adjusted\_rand\_score:*** *0.9475973819320783*

***jaccard\_score:*** *0.0*

***fowlkes\_mallows\_score:*** *0.9579232912541293*

***f1\_score:*** *0.0*





**Выводы**

Агломеративный метод с евклидовой метрикой показал себя неплохо по сравнению с k-means, а вот при использовании манхэттенской метрики результаты получаются хуже во всех тестах с двумерными данными. Если количество признаков увеличить, то алгомеративный метод будет выдавать хорошие результаты. DBCSAN работает хорошо даже с необычными по форме (визуально) данными, но только до тех пор, пока не появляется пересечение кластеров, и, чем больше процент пересечения, тем хуже он работает, вплоть до того, что совсем не может выделить даже 1 кластер. Не помогает ему и увеличение количества признаков, судя по результатам для датасета iris и сгенерированного датасета с 7 признаками.