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

ОТЧЕТ

по лабораторной работе №5
по дисциплине «Машинное обучение»
Тема: Кластеризация (k-средних, иерархическая)

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://archive.ics.uci.edu/ml/datasets/iris. Данные представлены в виде data файла. Данные представляют собой информацию о трех классах цветов
2. Создать Python скрипт. Загрузить данные в датафрейм

	0	1	2	3	4
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
..	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

[150 rows x 5 columns]

Рисунок 1 — Исходные данные

3. Проведем кластеризацию методов k-средних
4. Получим центры кластеров и определим какие наблюдения в какой кластер попали
5. Построим результаты классификации для признаков попарно (1 и 2, 2 и 3, 3 и 4)

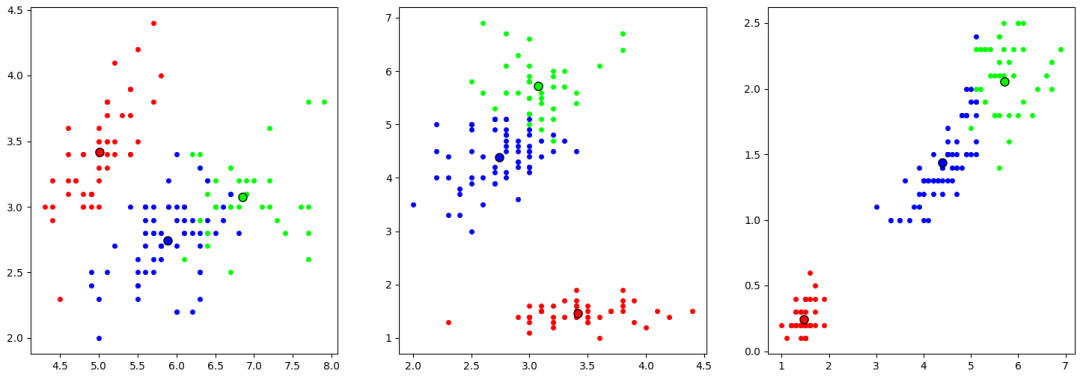

Рисунок 2 — Попарные результаты

Исходя из рисунка видно, что наилучшее разделение произошло по признакам 3, 4.

6. Уменьшена размерность данных до 2 используя метод главных компонент и нарисована карта для всей области значений, на которой каждый кластер занимает определенную область со своим цветом

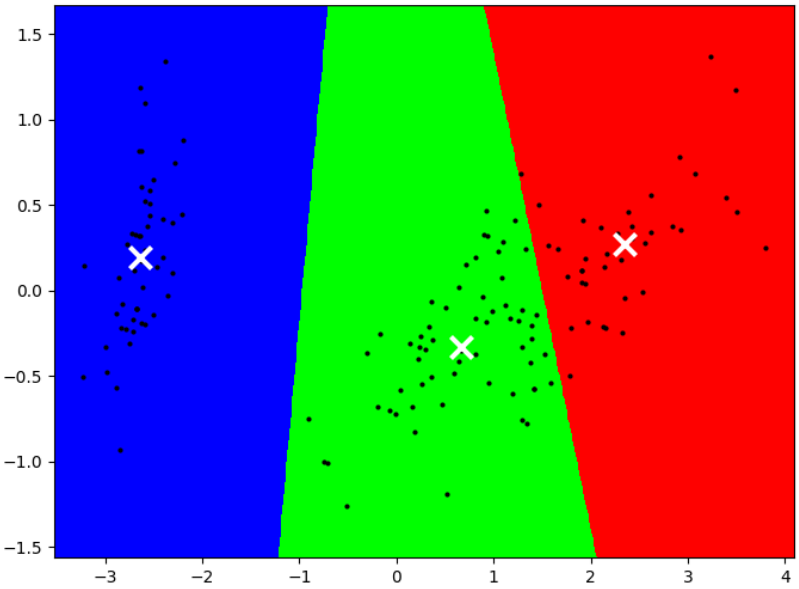

Рисунок 2 — Карта области значений с уменьшением размерности

7. Исследована работа алгоритма k-средних при различных параметрах init. Сначала нужно было выполнить несколько раз с параметров 'random', затем для вручную выбранных точек

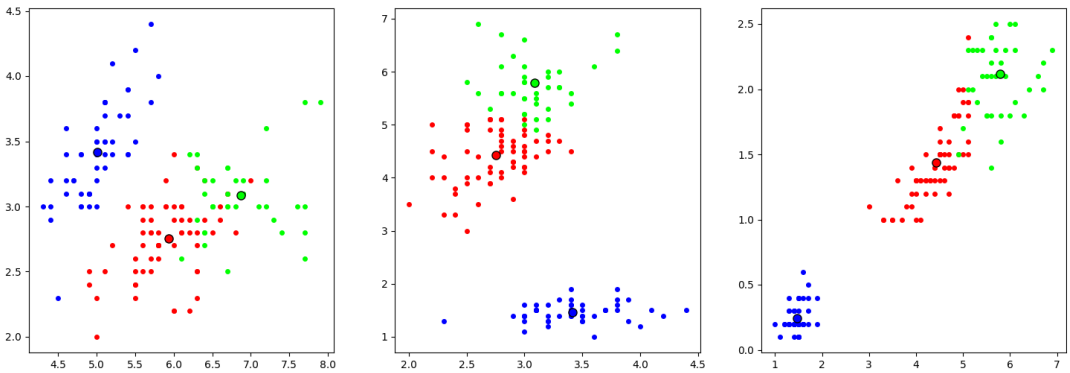

Рисунок 3 — init='random', n_clusters=3, max_iter=1

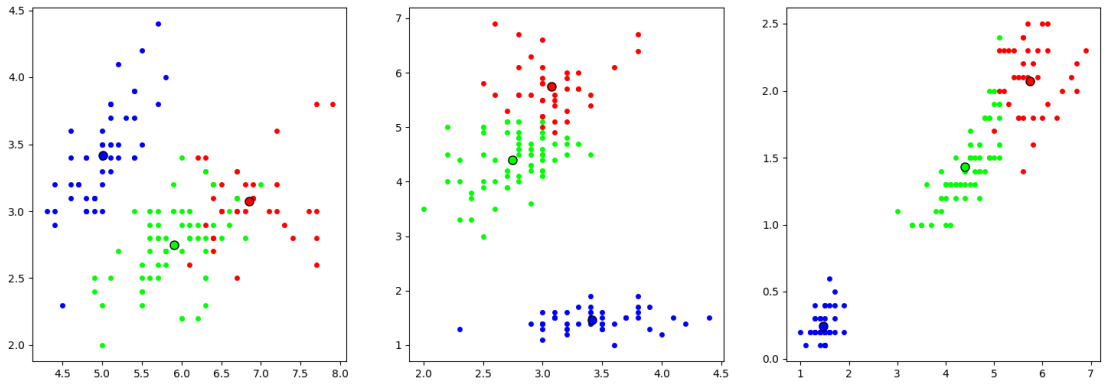

Рисунок 4 — init='random', n_clusters=3, max_iter=100

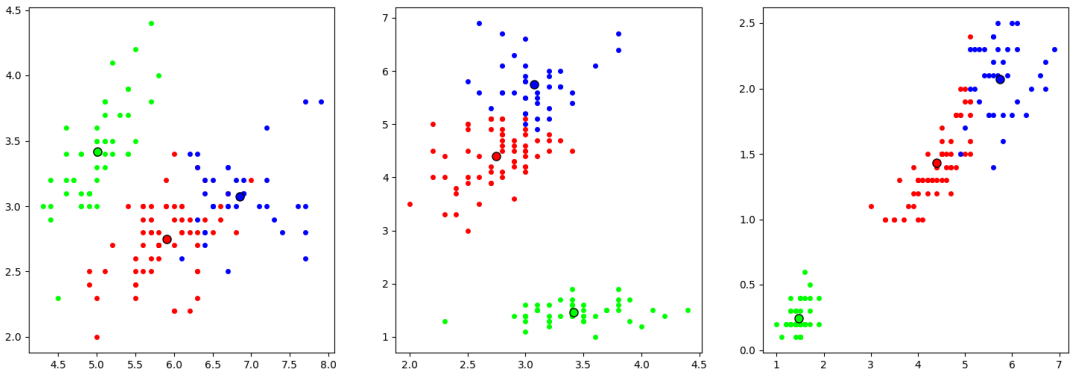

Рисунок 5 — init='random', n_clusters=3, max_iter=500

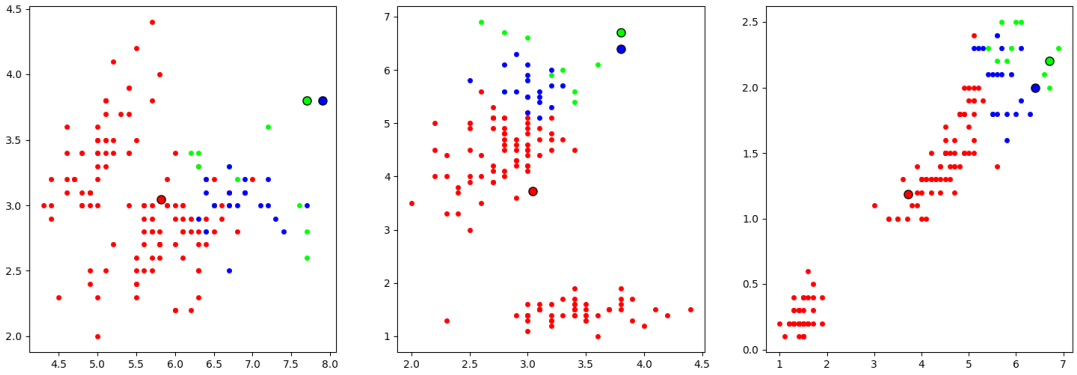

Рисунок 6 — init=[[0,0,0,0], [0,0,0,0], [0,0,0,0]], n_clusters=3, max_iter=1

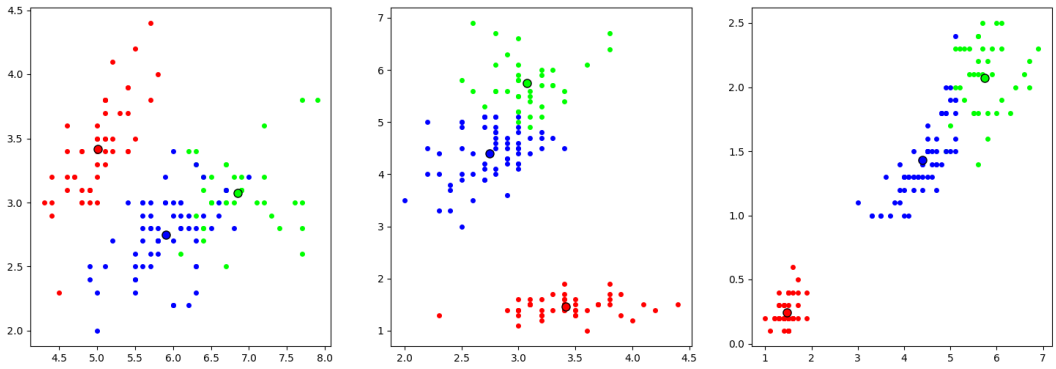

Рисунок 7 — init=[[0,0,0,0], [0,0,0,0], [0,0,0,0]], n_clusters=3, max_iter=100

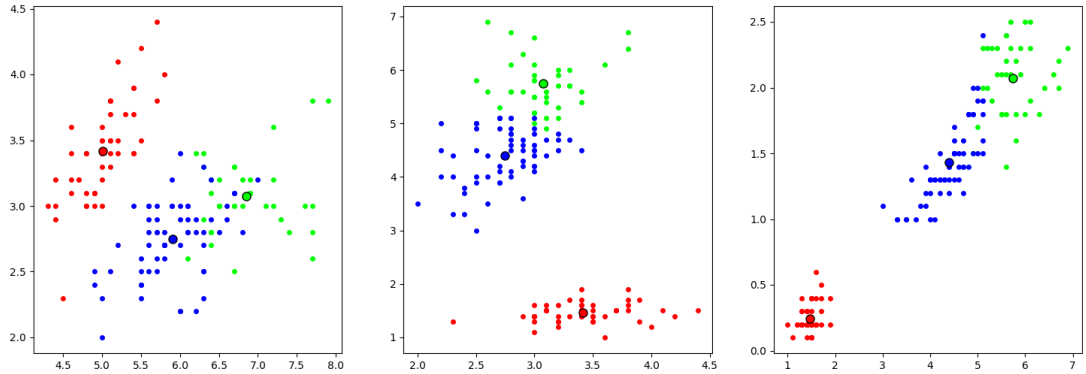

Рисунок 8 — init=[[0,0,0,0], [0,0,0,0], [0,0,0,0]], n_clusters=3, max_iter=500

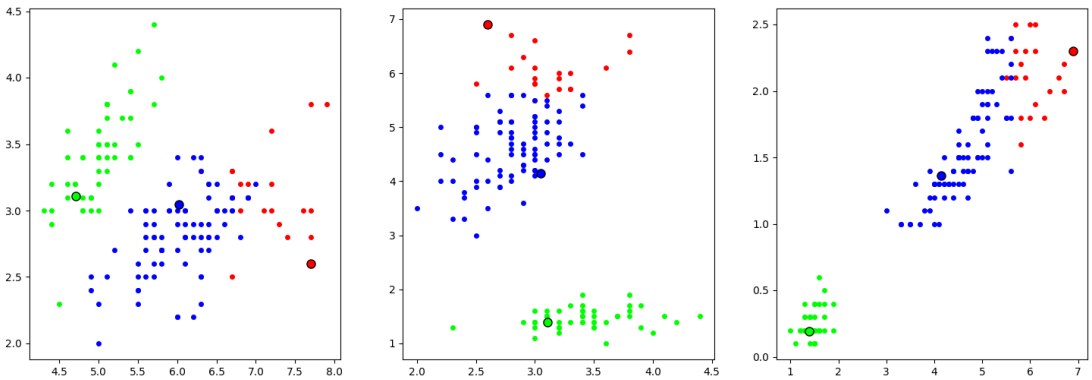

Рисунок 9 — init=[[1,1,1,1], [2,2,2,2], [3,3,3,3]], n_clusters=3, max_iter=1

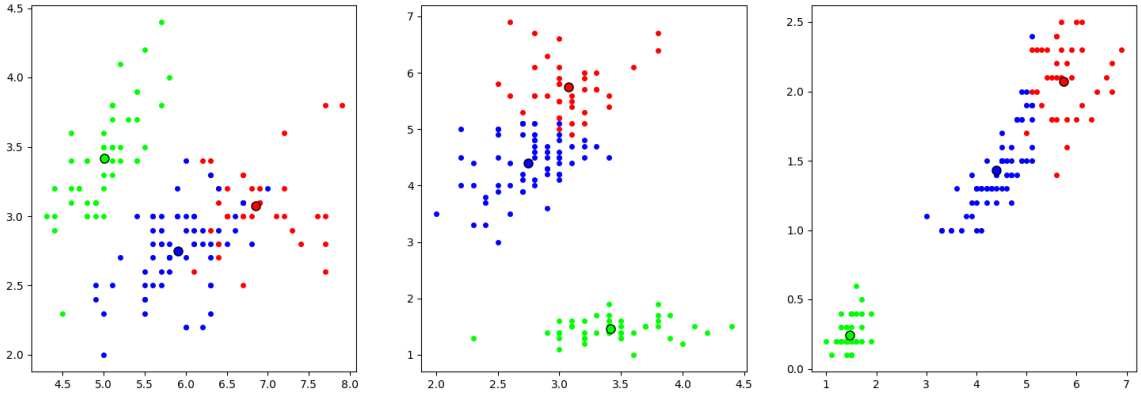

Рисунок 10 — init=[[1,1,1,1], [2,2,2,2], [3,3,3,3]], n_clusters=3, max_iter=100

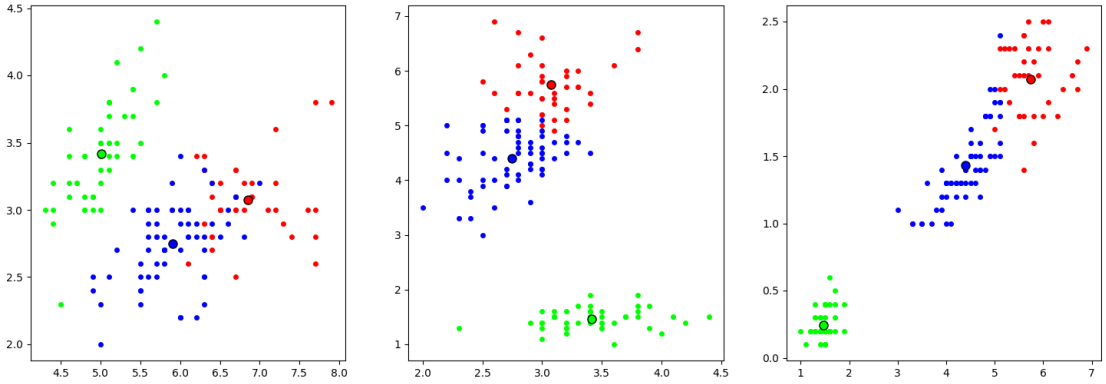

Рисунок 11 — init=[[1,1,1,1], [2,2,2,2], [3,3,3,3]], n_clusters=3, max_iter=500

Как можно заметить после init ручным и рандомным способами не привел к видимым изменениям центроид, за исключением случаев, когда максимальное число итераций – 1. В таком случае алгоритм просто не успевает отработать правильно.

8. Определено наилучшее количество методом локтя

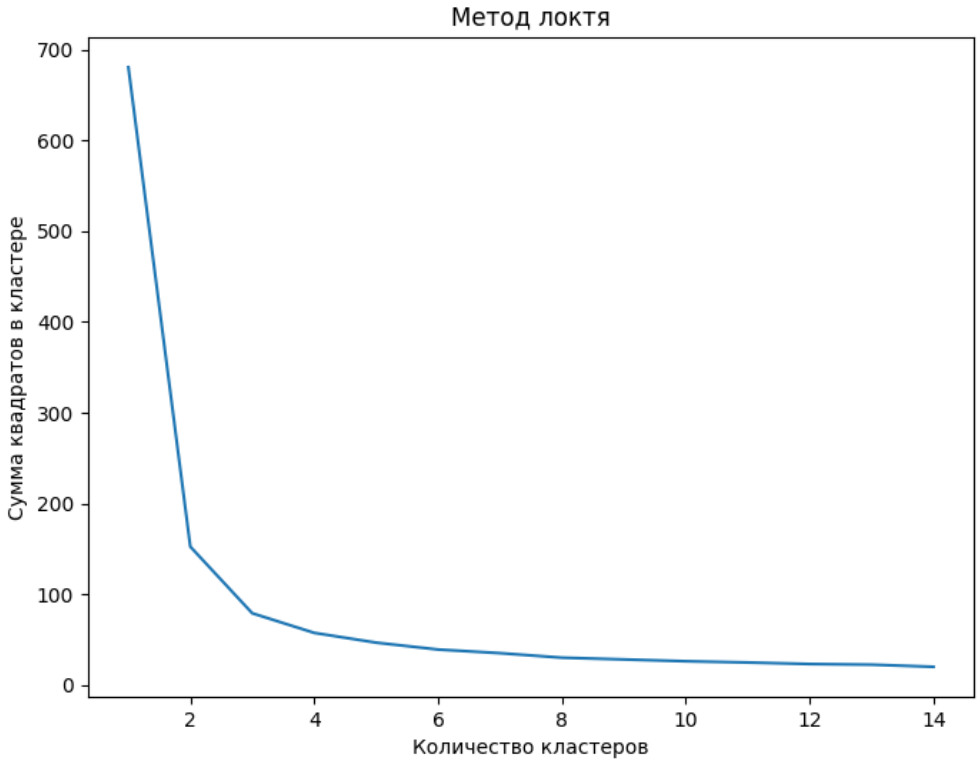

Рисунок 11 — Метод локтя

Как видно, точка изгиба – 2.

9. Проведена кластеризация используя пакетную кластеризацию k-средних.

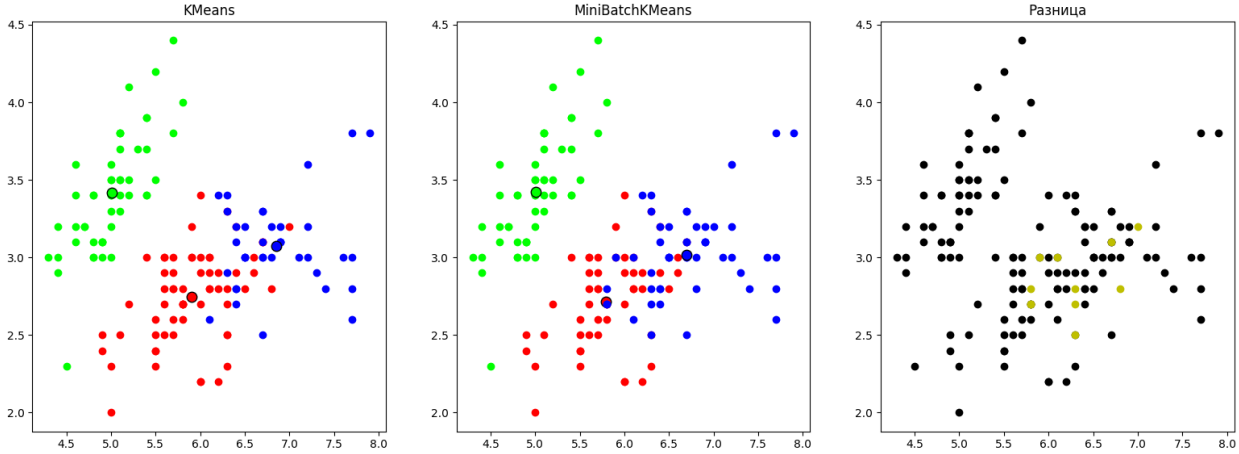

Рисунок 12 — KMeans и MiniBatchMeans

На 1-ом графике кластеризация KMeans, на 2-м – MiniBatchMeans, на 3-м разница между ними. Черным отмечены все совпадающие точки, желтым отличающиеся. Разница между методами в том, что MiniBatchMeans принимает постепенно данные пакетами, а не все сразу. Выигрыш в скорости работы, но потеря точности.

10. Проведем иерархическую кластеризацию на тех же данных

11.Отобразим результаты кластеризации

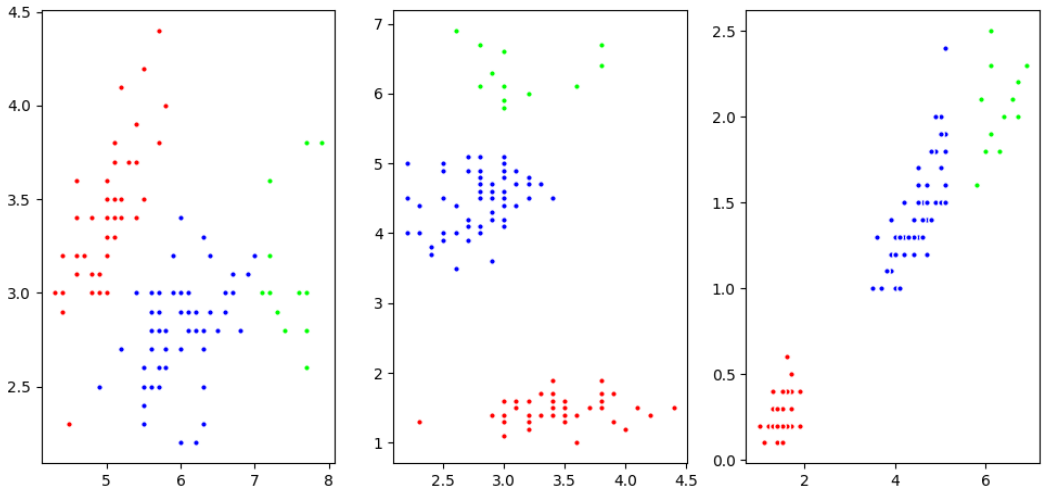

Рисунок 13 — Иерархическая кластеризация

Разница между методами в том, что при начальном состоянии каждая точка – кластер, по мере итерирования находятся ближайшие кластеры и сливаются по заданной метрике длины.

12.Проведены исследование для различного размера кластеров (от 2 до 5). Приведены полученные результаты

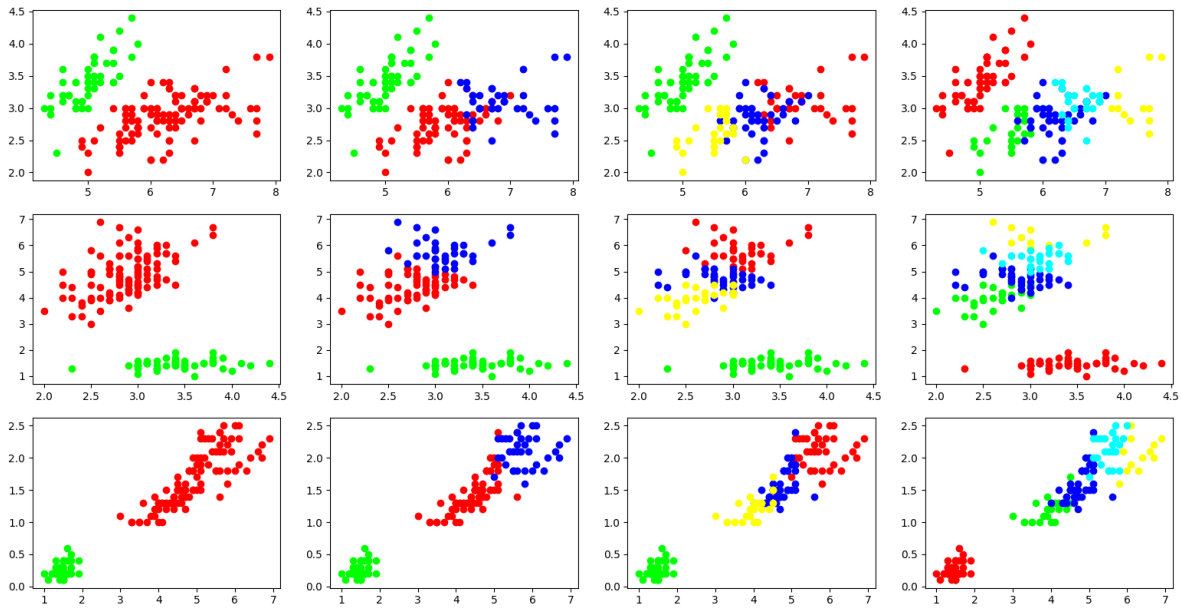

Рисунок 14 — Иерархическая кластеризация для разного числа кластеров

13.Нарисована дендограмма до уровня 6

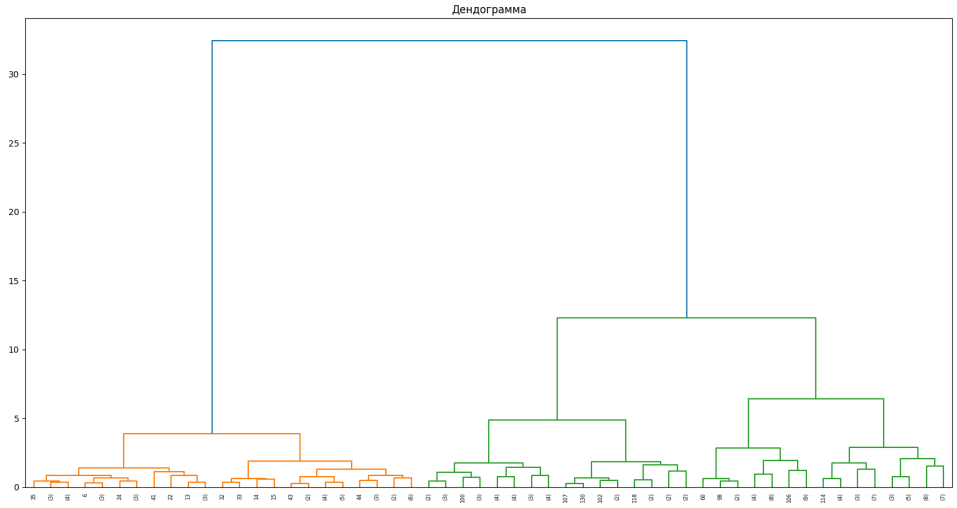

Рисунок 15 — Дендограмма

14. Сгенерированы случайные данные в виде двух колец

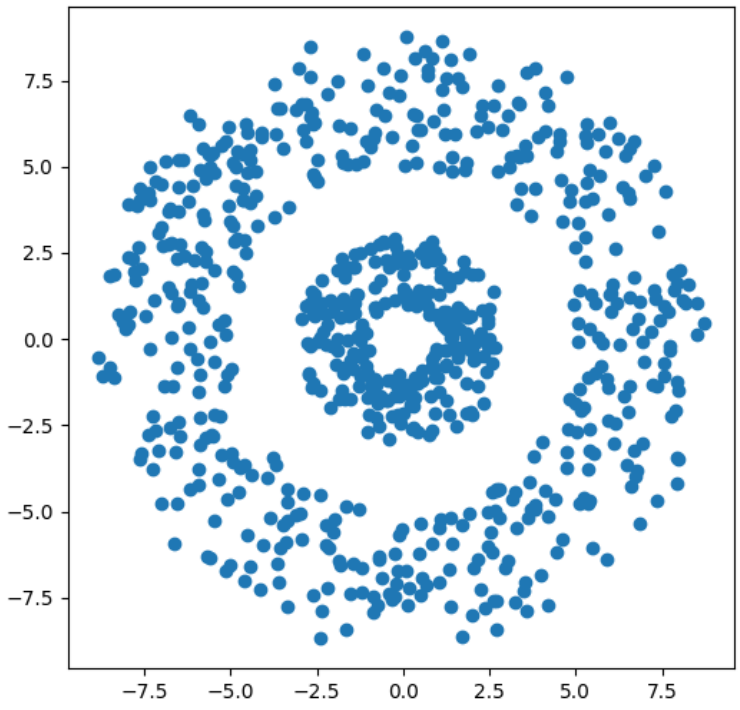

Рисунок 16 — Кольца

15.Проведена иерархическая кластеризация.

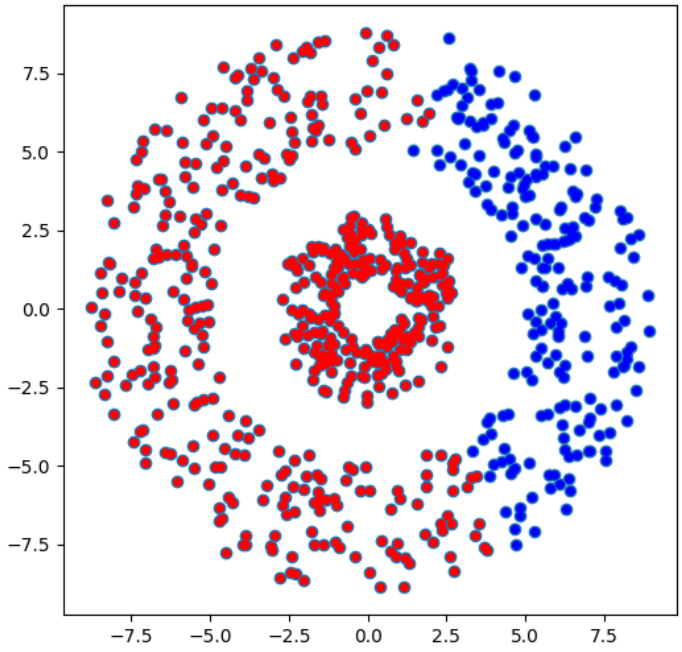

Рисунок 17 — Иерархическая кластеризация методом Уорда

16. Исследована кластеризация при всех параметрах linkage. Отображены и обоснованы полученные результаты

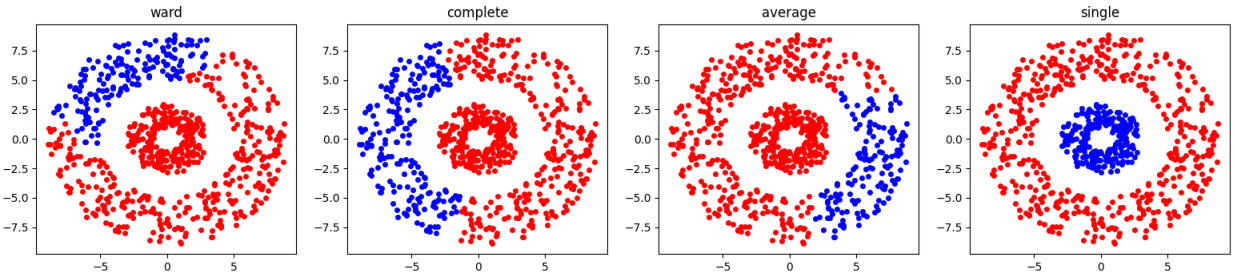

Рисунок 18 — Разные параметры linkage

- Ward – минимизация суммы квадратов разностей
- Complete – минимизация максимально расстояния
- Average – минимизация среднего расстояния
- Single – минимизация расстояния

Разделение колец произошло только при Single Link, т.к. расстояние между кластерами есть расстояние между ближайшими точками.

В иных случаях расстояние между кластерами, лежащими на разных кольцах меньше, чем между кластерами на одном кольце.

Вывод

В ходе выполнения данной лабораторной работы было выполнено ознакомление с методами кластеризации модуля Sklearn. Пакетный метод k-средних приводит к небольшим изменения в сравнении с полным k-средним. Метод иерархической кластеризации при правильной настройке может определить нелинейную зависимость между данными.