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

ОТЧЕТ

по лабораторной работе № 2

по дисциплине «Основы машинного обучения»

Тема: кластеризация

Вариант 134К

Студентка гр. 1304 _____ Хорошкова А.С.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург

2024

Задание.

Подготовить наборы данных.

Провести кластеризацию с помощью алгоритма K-Means.

Провести кластеризацию с помощью алгоритма DBSCAN.

Провести кластеризацию с помощью Иерархической кластеризации.

Изучить набор данных с большим количество признаков.

Выполнение работы.

1. Подготовка наборов данных

Были загружены данные из файлов lab2_blobs.csv, lab2_checker.csv и lab2_noisymoons.csv согласно варианту. Для загрузки была использована функция из первой лабораторной работы (представлена на Листинг 1.1).

Листинг 1.1

```
def upload_df(file_name, delete_first=True):
    df = pd.read_csv(file_name, delimiter=',')
    if delete_first:
        return df.drop(df.columns[[0]], axis=1)
    else:
        return df
```

С помощью метода print_df_data (Листинг 1.2) из первой лабораторной работы у всех датафреймов был вызван метод head, выводящий первые 5 строк датафрейма, и метод describe, выводящий характеристики датафрейма. Таким образом, была проверена корректность загрузки.

Листинг 1.2

```
def print_df_data(df):
    print("\n============== head ==============")
    print(df.head())
    print("\n============= describe ===============")
    print(df.describe())
```

Результат вывода представлен на Рисунок 1.

lab2_blobs.csv

========== head ==========

	x	y
0	-8.0267	-4.9731
1	-7.0422	-2.6454
2	8.9214	9.5679
3	1.0887	-0.2884
4	0.4739	-0.0737

========== describe ==========

	x	y
count	260.000000	260.000000
mean	0.128889	0.306402
std	5.277836	5.784841
min	-10.882000	-10.543400
25%	-6.125800	-4.001125
50%	1.917600	-0.603950
75%	4.040475	6.950550
max	8.921400	11.368800

lab2_checker.csv

========== head ==========

	x	y
0	4.0510	0.9697
1	7.5581	5.1224
2	2.8765	7.0870
3	3.8366	0.8614
4	4.2159	0.7742

========== describe ==========

	x	y
count	250.000000	250.000000
mean	3.646908	5.183671
std	2.286382	2.522758
min	-0.367000	-0.024100
25%	2.408200	4.282800
50%	3.322550	5.123200
75%	4.439800	6.553200
max	8.682800	9.940200

lab2_noisymoons.csv

========== head ==========

	x	y
0	-0.5237	0.8448
1	0.2002	0.9865
2	-0.4794	0.8188
3	0.5155	0.9515
4	1.8891	0.1536

========== describe ==========

	x	y
count	280.000000	280.000000
mean	0.420280	0.299863
std	0.831229	0.492804
min	-1.083500	-0.950400
25%	-0.168400	-0.083950
50%	0.451250	0.308050
75%	0.949200	0.772400
max	2.064800	1.145700

Рисунок 1 - вывод первых пяти строк датафреймов и описания датафреймов

С помощью функции scatterplot библиотеки seaborn построены диаграммы рассеяния для оценки формы трёх наборов данных, результаты представлены на Рисунок 2, Рисунок 3 и Рисунок 4.

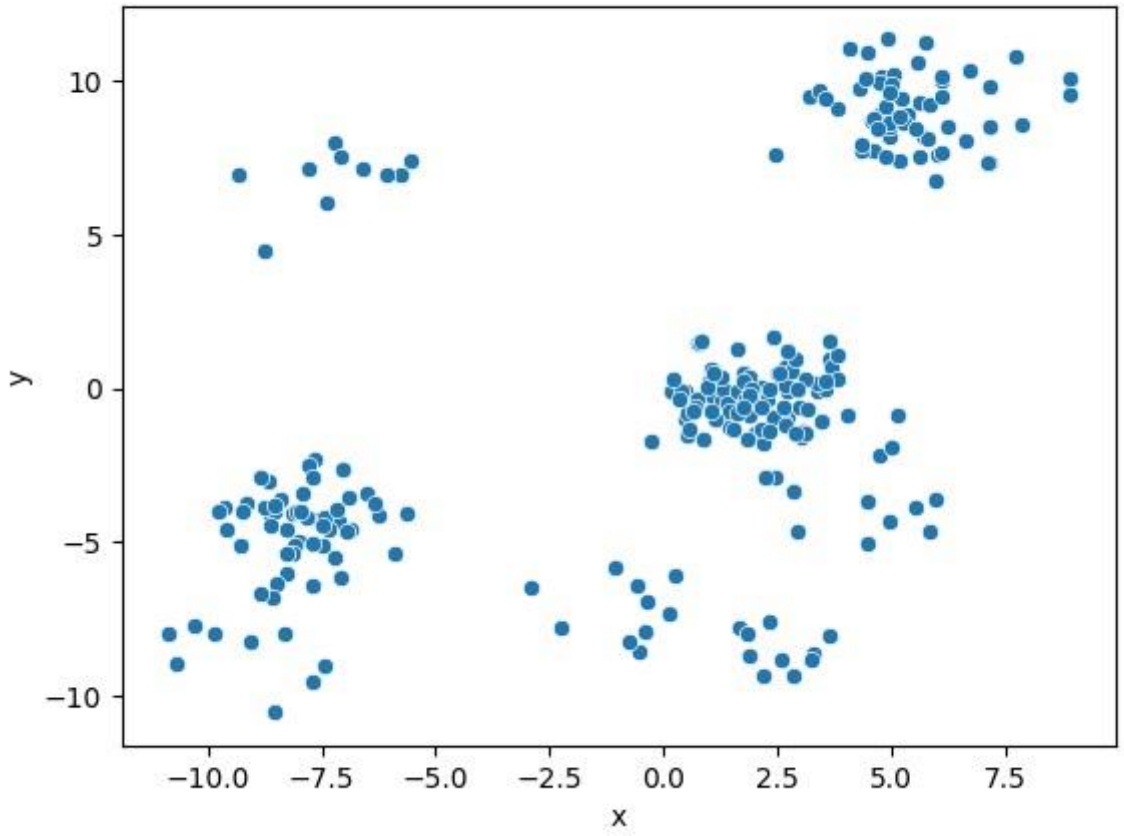


Рисунок 2 – диаграмма рассеяния для lab2_blobs.csv

На Рисунок 2 видна форма данных lab2_blobs.csv. График имеет четыре чётких скопления точек: слева сверху неплотное и относительно небольшое скопление, справа сверху плотное скопление, немного рассеянное по краям, снизу слева плотное, но имеющее отстающие от общего скопления точки снизу, и справа снизу самое крупное скопление точек, также имеющее отстающие плотно расположенные точки снизу, которые можно принять за отдельное скопление.

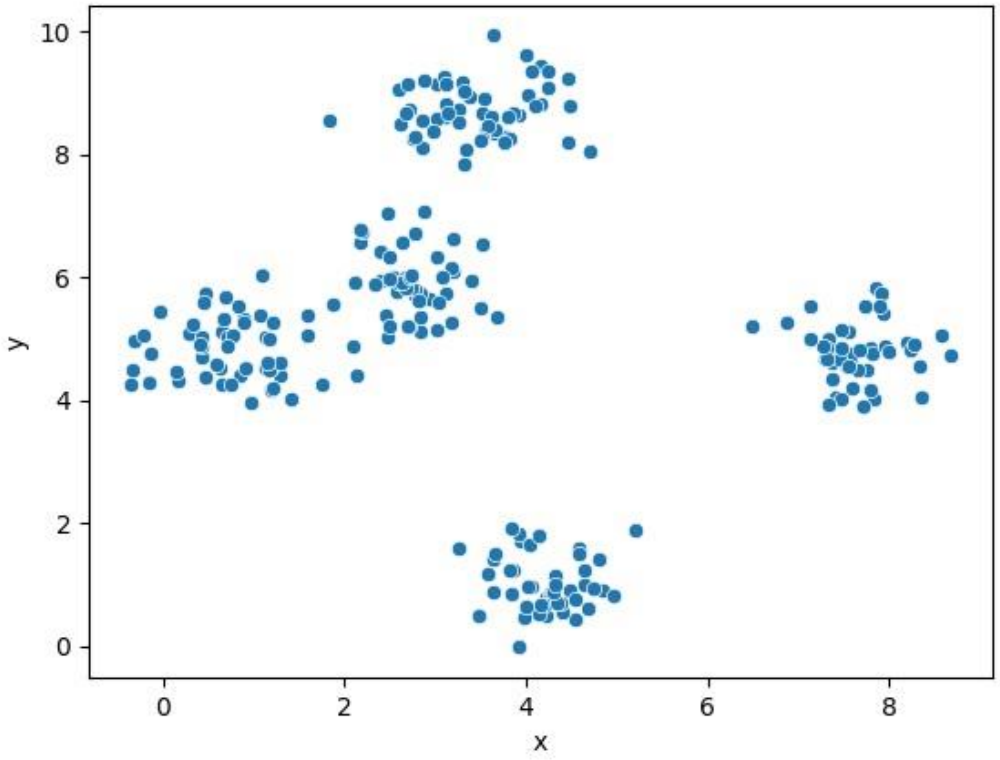


Рисунок 3 - графики диаграмма рассеяния для lab2_checker.csv

На Рисунок 3 видна форма данных lab2_checker.csv. График имеет два обособленных скопления точек снизу и крупное скопление точек сверху, состоящее из трёх визуально разделимых подгрупп.

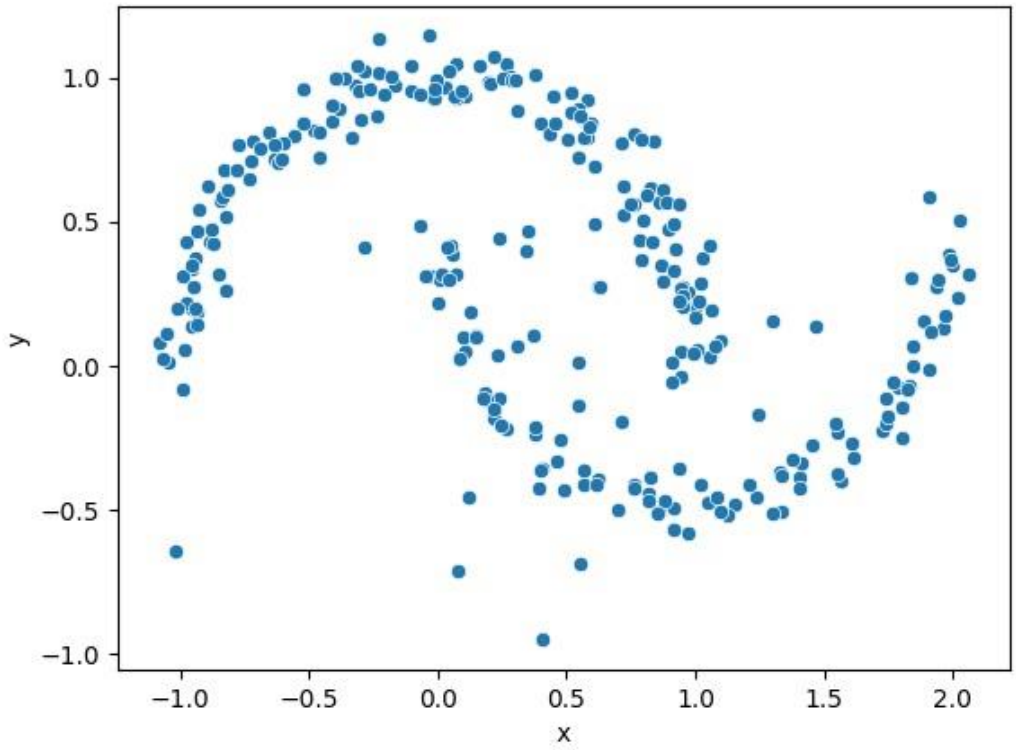


Рисунок 4 - диаграмма рассеяния для lab2_noisymoons.csv

На Рисунок 4 видна форма данных lab2_noisymoons.csv. График имеет два скопления точек в форме полумесяца, огибающих друг друга. Также данные имеют небольшое количество точек, отходящих от основных скоплений.

Для подготовки данных к последующей кластеризации была проведена их нормировка. Нормировка данных в данном случае необходима для того, чтобы каждый признак данных находился в конкретном диапазоне (в нашем случае от 0 до 1), таким образом, разный масштаб признаков не будет влиять на итог кластеризации. Функция, нормализующая набор данных, представлена на Листинг 1.3.

Листинг 1.3

```
def scale(df, columns):
    arr = df.to_numpy()
    scaler = MinMaxScaler()
    scaler.fit(arr)
    arr = scaler.transform(arr)
    return pd.DataFrame(arr, columns=columns)
```

На Рисунок 5, Рисунок 6 и Рисунок 7 представлены диаграммы рассеяния после нормировки для трёх наборов данных соответственно.

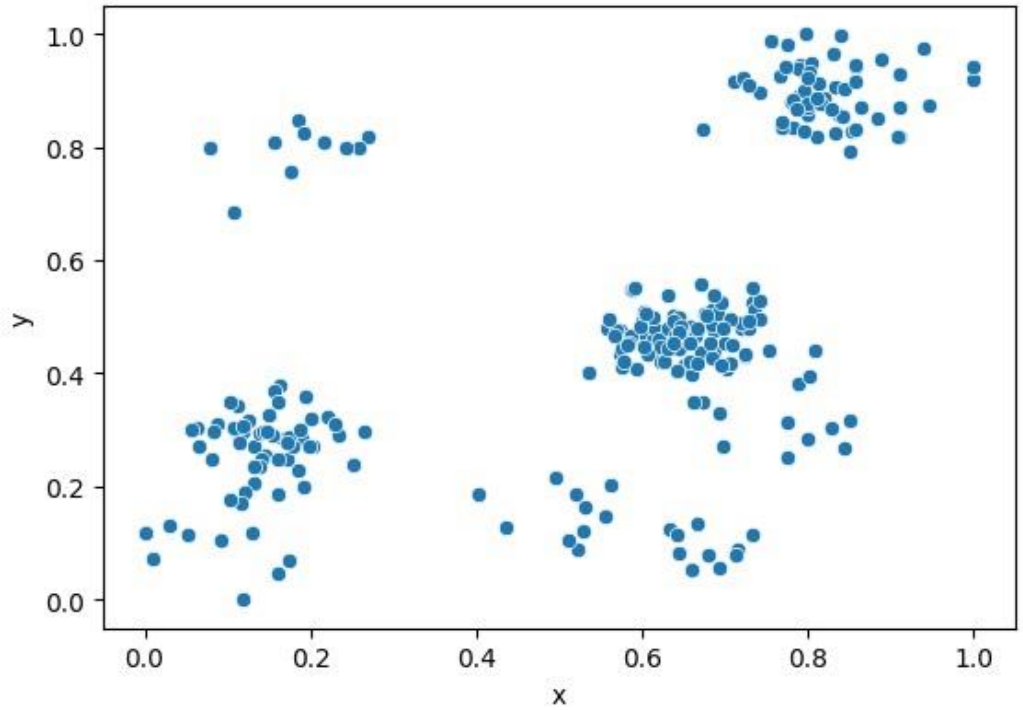


Рисунок 5 - диаграмма рассеяния для lab2_blobs.csv после нормировки

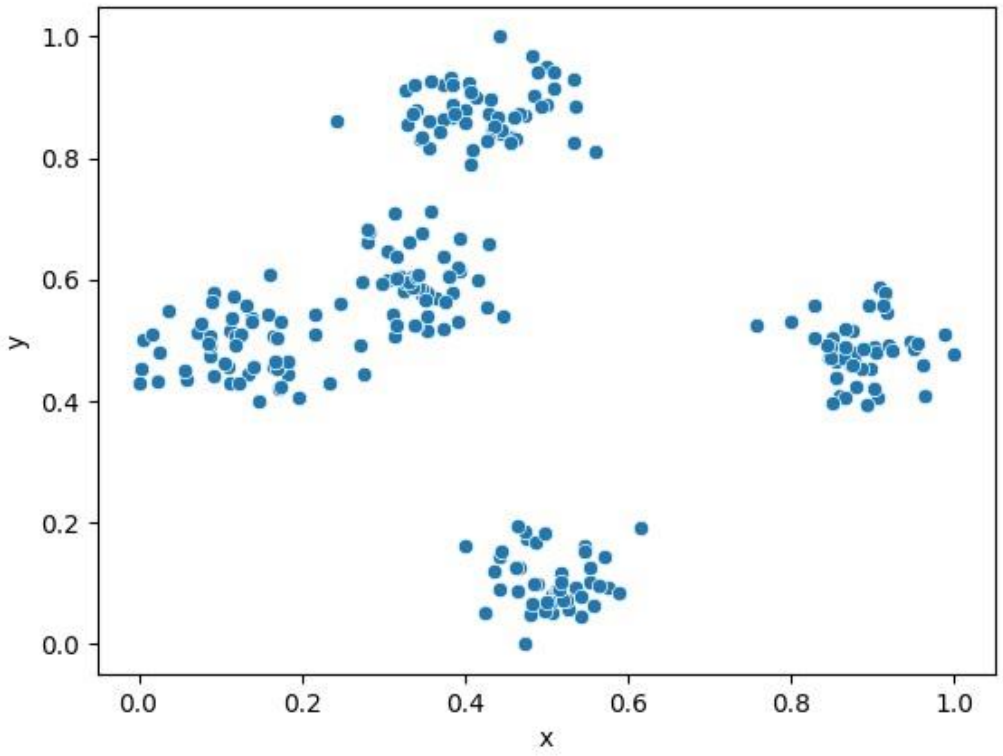


Рисунок 6 - графики диаграмма рассеяния для lab2_checker.csv после нормировки

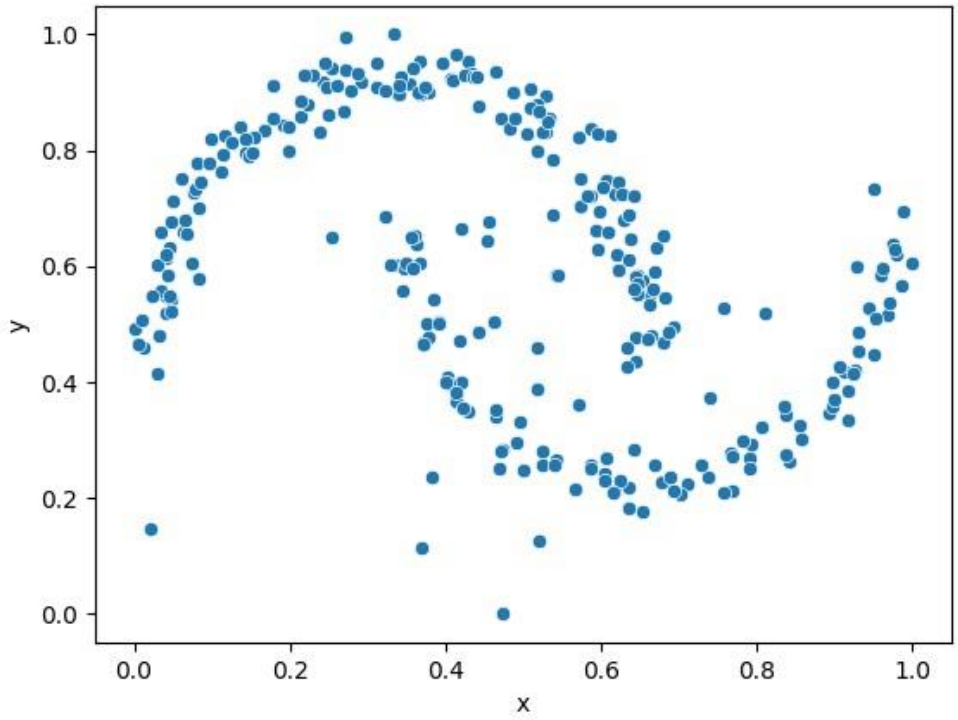


Рисунок 7 - диаграмма рассеяния для lab2_noisymoons.csv после нормировки

2. K-mean

С помощью функции show_elbow_method (Листинг 2.1) было проведено исследование оптимального количества кластеров методом локтя. На Рисунок 8,

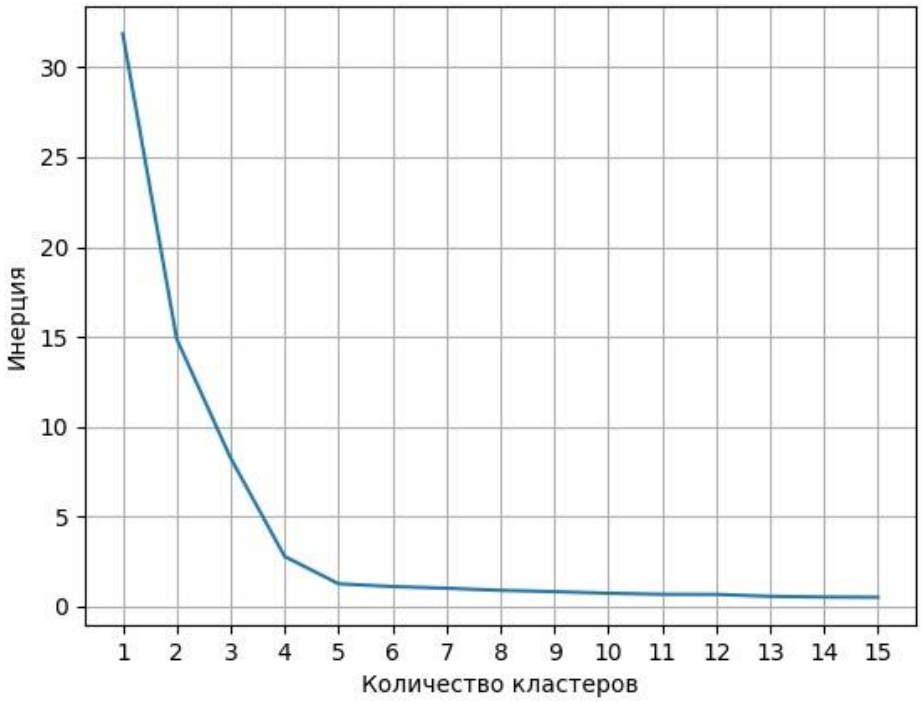


Рисунок 9 и Рисунок 10 представлены графики, помогающие определить оптимальное количество кластеров методом локтя для трёх наборов данных соответственно.

Листинг 2.1

```
def show_elbow_method(df, y=15):
    inert_list = []
    for i in range(y):
        temp_km = KMeans(i + 1, n_init=5)
        temp_km.fit(df)
        inert_list.append(temp_km.inertia_)

    plt.plot(list(range(1, y + 1)), inert_list)
    plt.ylabel('Инерция')
    plt.xlabel('Количество кластеров')
    plt.xticks(list(range(1, y + 1)))
    plt.grid()
    plt.show()
```

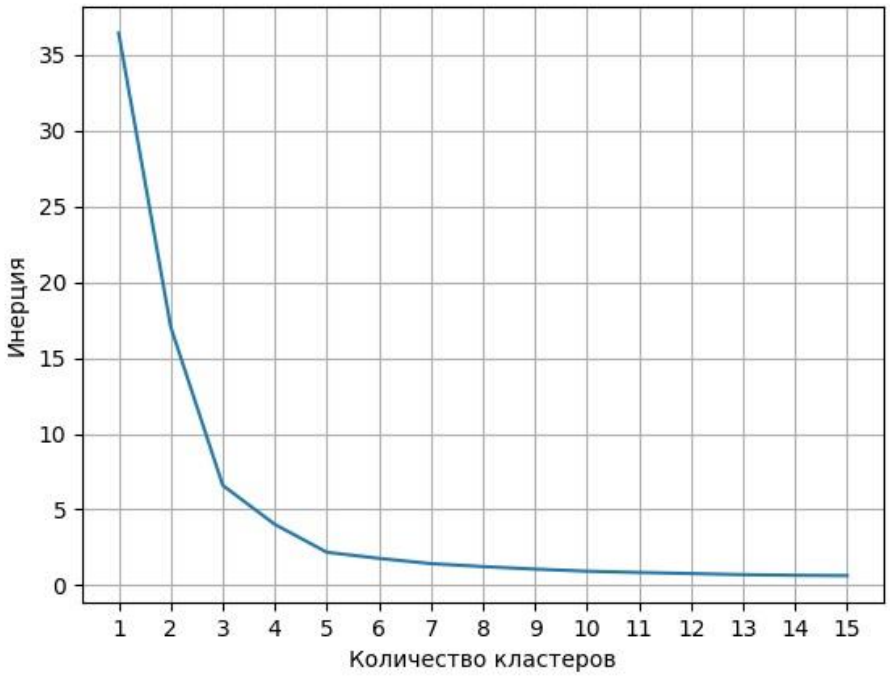


Рисунок 8 – метод локтя для lab2_blobs.csv

По Рисунок 8 видно, что по методу локтя оптимальное количество кластеров для lab2_blobs.csv 3-5.

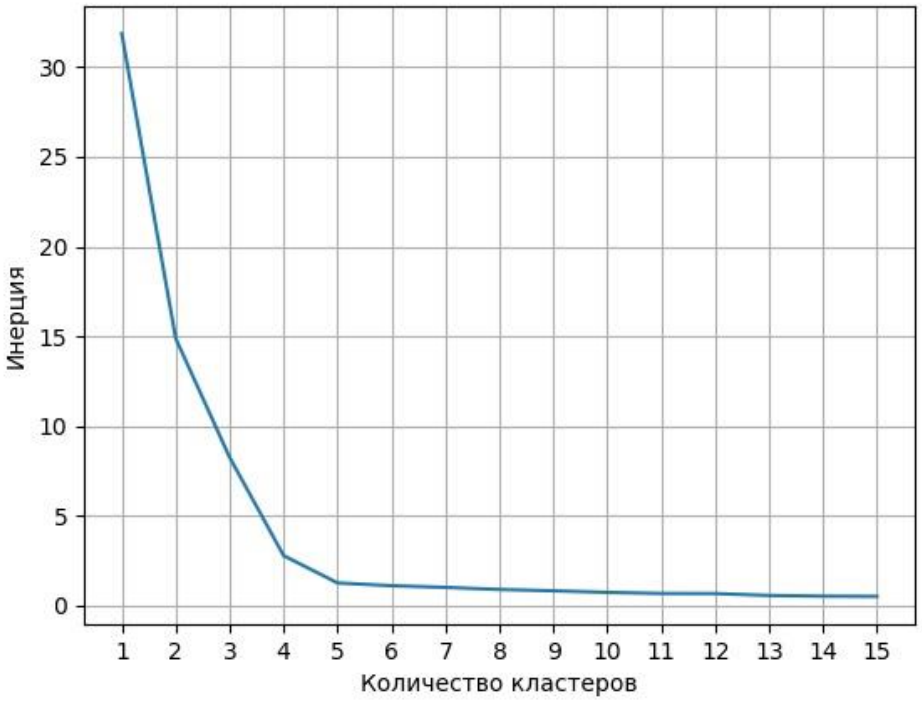
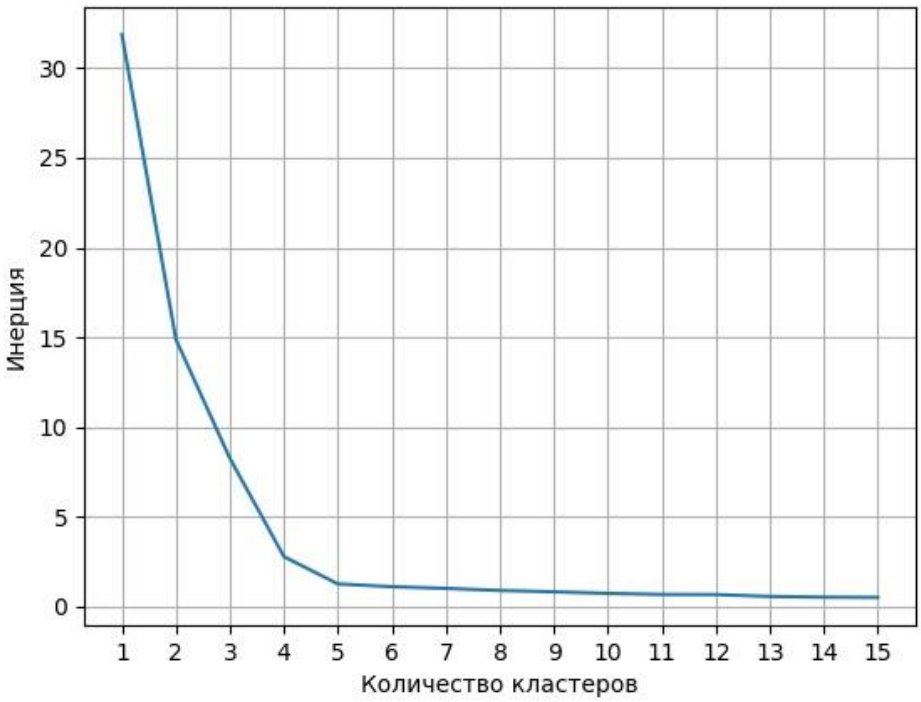


Рисунок 9 – метод локтя для lab2_checker.csv



По

Рисунок 9 видно, что по методу локтя оптимальное количество кластеров для lab2_checker.csv 4-5.

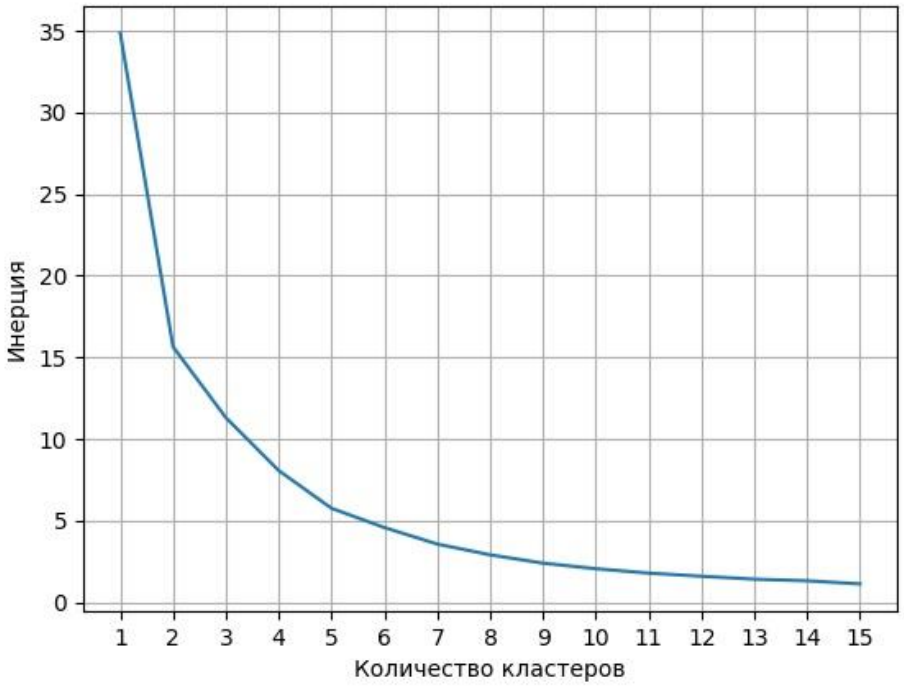


Рисунок 10 – метод локтя для lab2_noisymoons.csv

По Рисунок 10 видно, что по методу локтя оптимальное количество кластеров lab2_noisymoons.csv 5-8.

С помощью функции show_elbow_method (Листинг 2.2Листинг 2.1) было проведено исследование оптимального количества кластеров методом силуэта. На Рисунок 11, Рисунок 12 и Рисунок 13 представлены графики, помогающие определить оптимальное количество кластеров методом силуэта для трёх наборов данных соответственно.

Листинг 2.2

```
def show_silhouette_method(df, y=15):
    sil_list = []
    for i in range(1, y):
        temp_km = KMeans(i + 1, n_init=5)
        temp_clust = temp_km.fit_predict(df)
        sil_list.append(silhouette_score(df, temp_clust))

    plt.plot(list(range(2, y + 1)), sil_list)
    plt.ylabel('Среднее значение коэффициента силуэта')
    plt.xlabel('Количество кластеров')
    plt.xticks(list(range(1, y + 1)))
    plt.grid()
    plt.show()
```

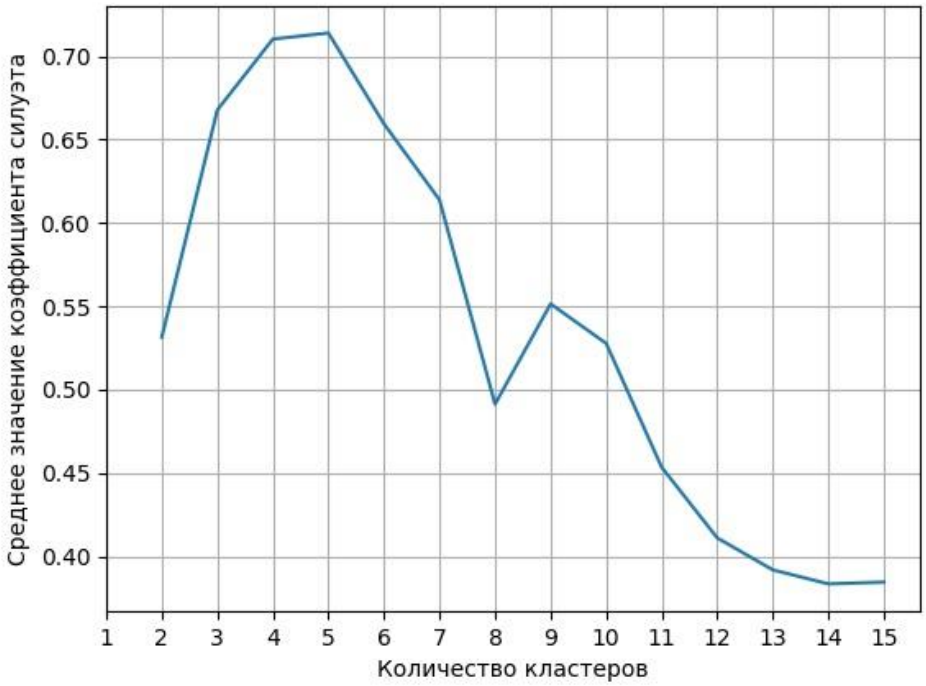


Рисунок 11 – метод силуэта для lab2_blobs.csv

По Рисунок 11 видно, что по методу силуэта оптимальным количеством кластеров для lab2_blobs.csv является 5.

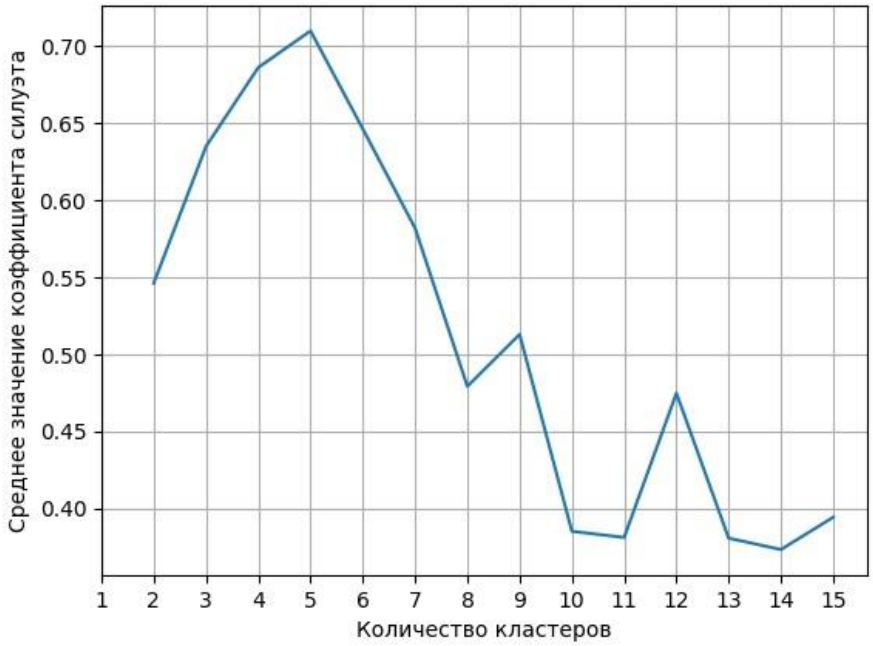


Рисунок 12 – метод силуэта для lab2_checker.csv

По Рисунок 12 видно, что по методу силуэта оптимальным количеством кластеров для lab2_checker.csv является 5.

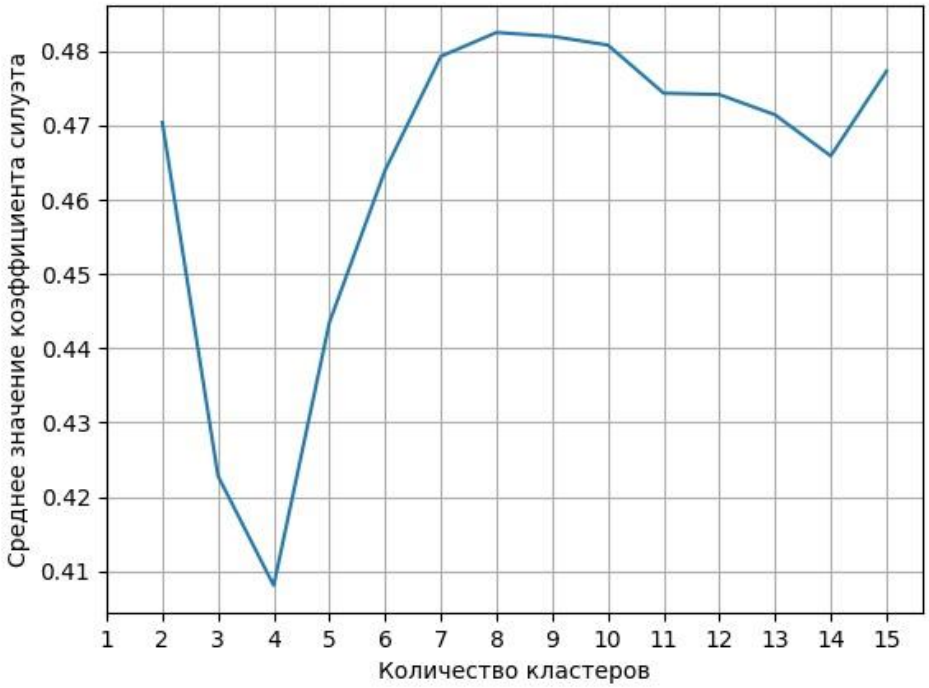


Рисунок 13 - метод силуэта для lab2_noisymoons.csv

По Рисунок 13 видно, что по методу силуэта оптимальным количеством кластеров lab2_noisymoons.csv является 8.

С помощью функции clusterize (Листинг 2.3) была проведена кластеризация алгоритмом K-means с выбранным количеством кластеров.

Листинг 2.3

```
def clusterize(df, clusters):
    norm_km = KMeans(n_clusters=clusters)
    norm_clust = norm_km.fit_predict(df)
    norm_pd = pd.DataFrame(df, columns=['x', 'y'])
    norm_pd['cluster'] = norm_clust
    return norm_pd, norm_km
```

Для данных lab2_blobs.csv и lab2_checker.csv было выбрано количество кластеров, равное 5, для данных lab2_noisymoons.csv было количество кластеров, равное 8.

Диаграммы рассеяния результатов кластеризации для трёх датасетов представлены на Рисунок 14,

Рисунок 15 и **Error! Reference source not found.**.

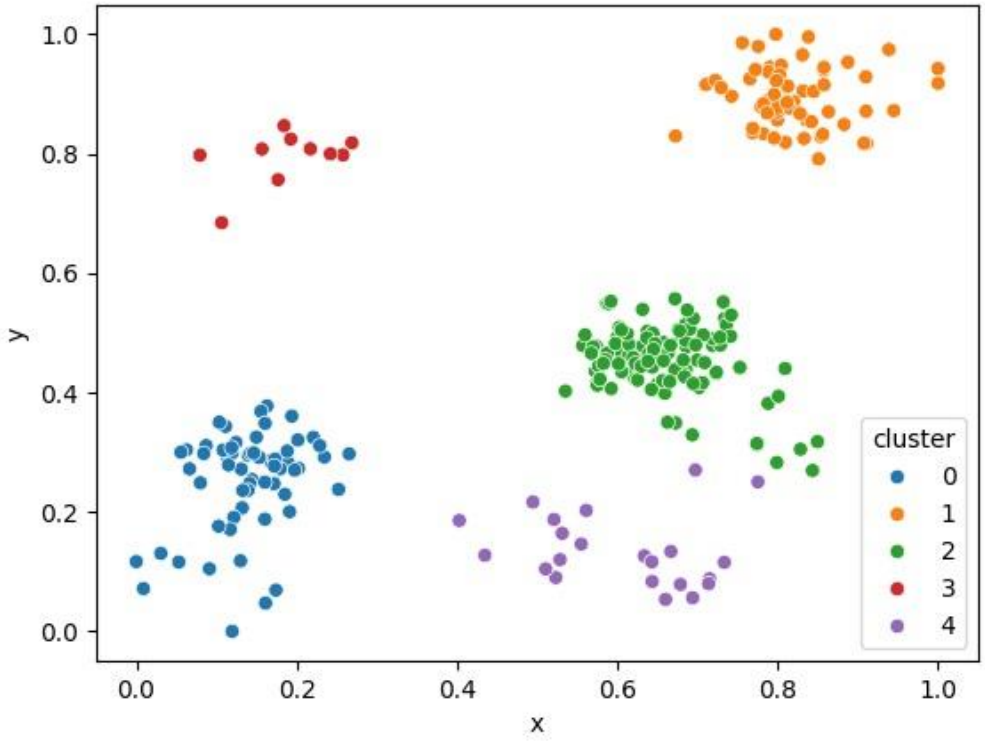


Рисунок 14 - диаграмма рассеяния результатов кластеризации на 5 кластеров для lab2_blobs.csv методом k-means

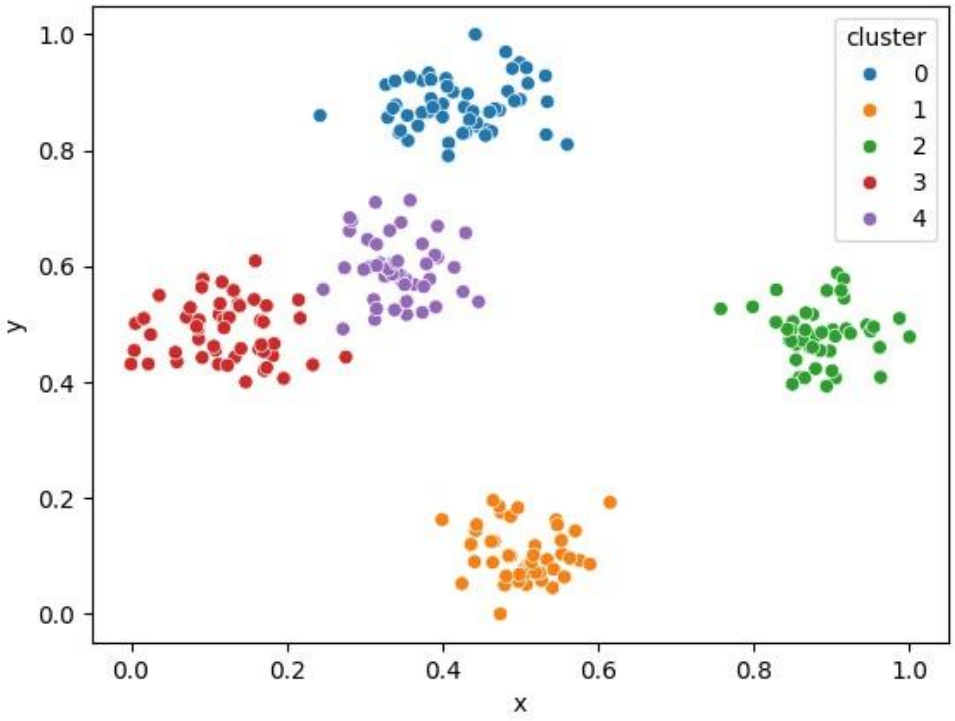


Рисунок 15 - диаграмма рассеяния результатов кластеризации на 5 кластеров для lab2_checker.csv методом k-means

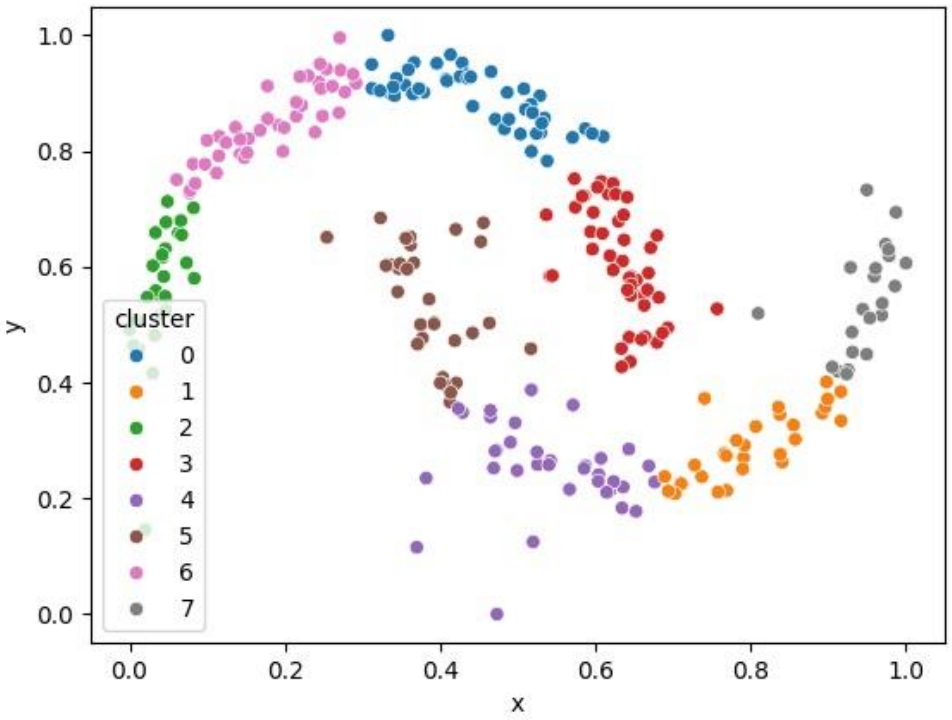


Рисунок 16 – диаграмма рассеяния результатов кластеризации на 8 кластеров для lab2_noisymoons.csv методом k-means

Была построена диаграмма Вороного для результатов кластеризации. На диаграмме были отмечены центроиды полученных кластеров. Результаты для трёх наборов данных представлены на Рисунок 17, Рисунок 18 и Рисунок 19.

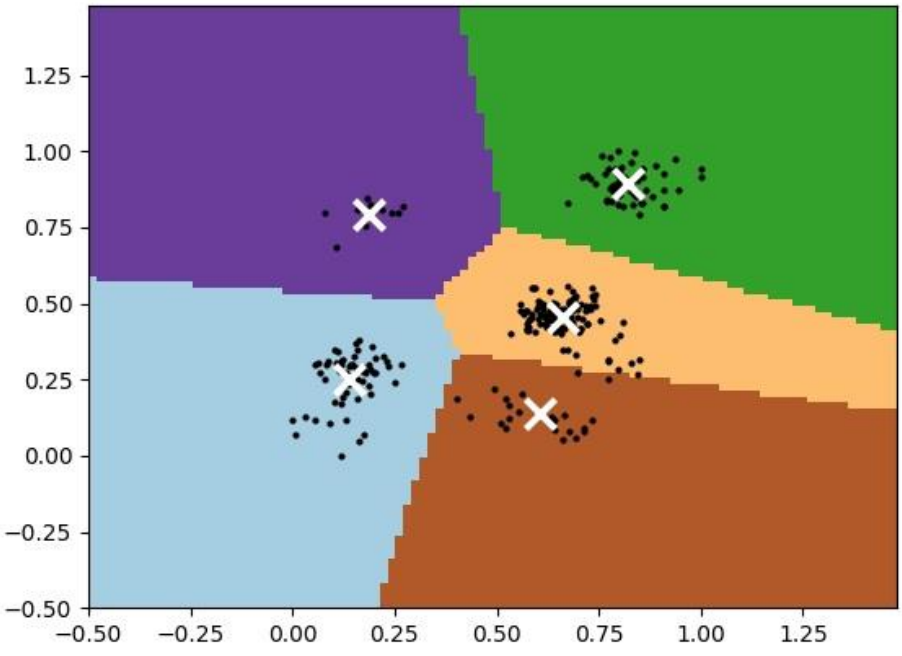


Рисунок 17 – диаграмма Вороного для lab2_blobs.csv

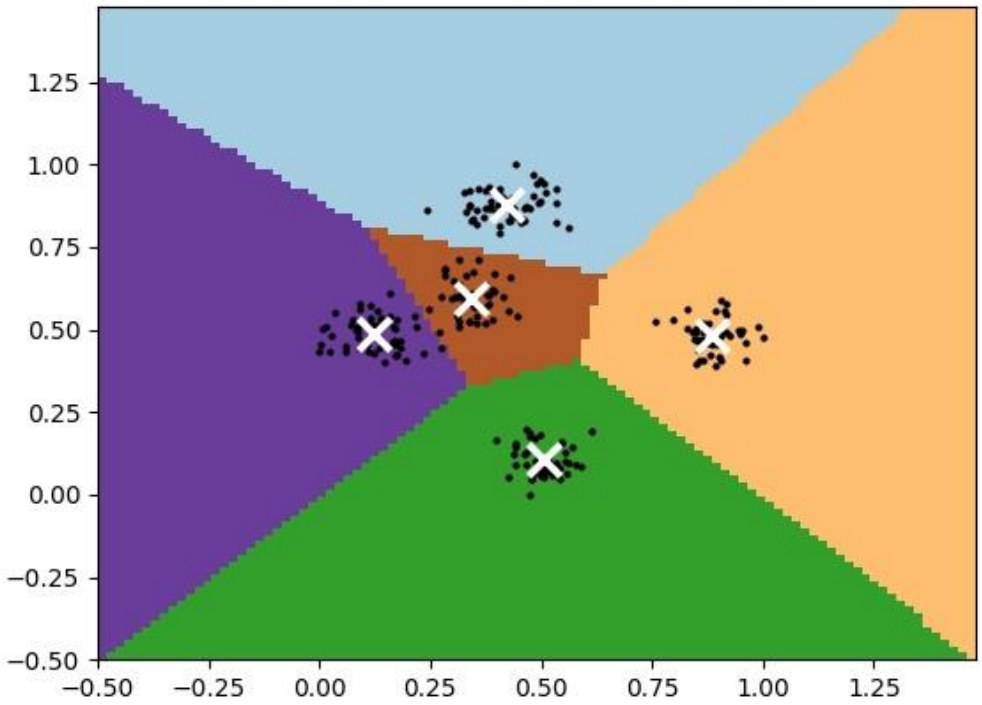


Рисунок 18 – диаграмма Вороного для lab2_checker.csv

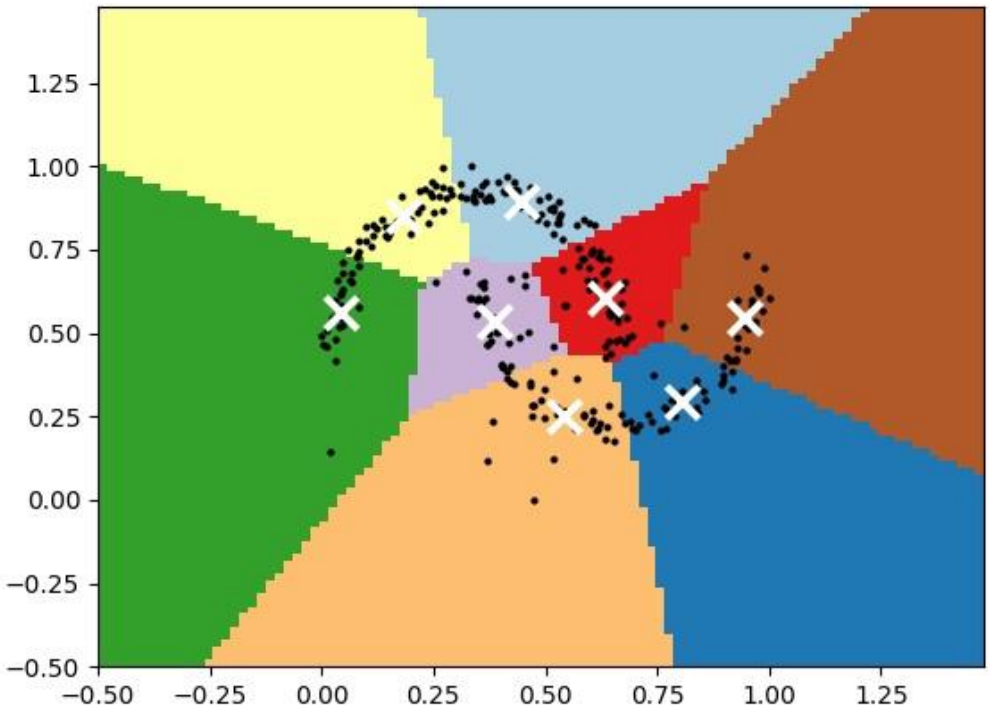


Рисунок 19 - диаграмма Вороного для lab2_noisymoons.csv

Для каждого признака была построена диаграмма “box-plot” с разделением по кластерам. Результаты для признаков трёх датасетов представлены на **Error! Reference source not found.** и **Error! Reference source not found.**, **Error! Reference source not found.** и **Error! Reference source not found.**, .Рисунок 24 и Рисунок 25 соответственно.

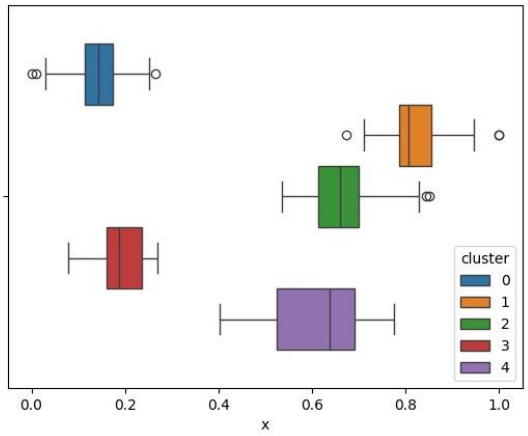


Рисунок 20 - диаграмма “box-plot” для признака x для lab2_blobs.csv

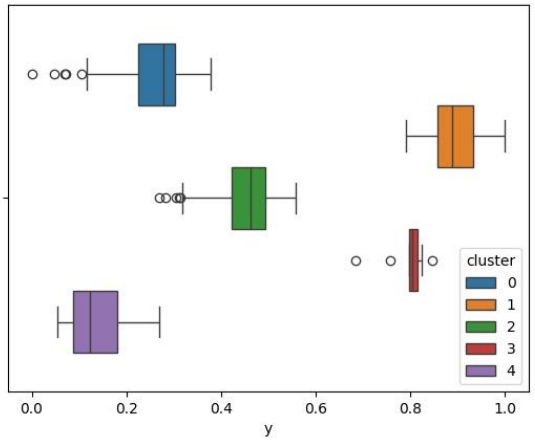


Рисунок 21 – диаграмма “box-plot” для признака y для lab2_blobs.csv

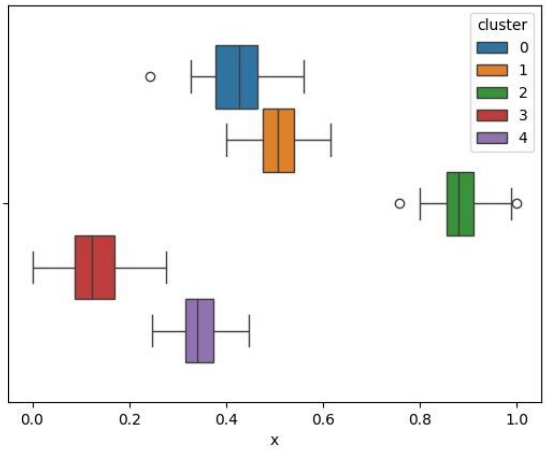


Рисунок 22 - диаграмма “box-plot” для признака x для lab2_checker.csv

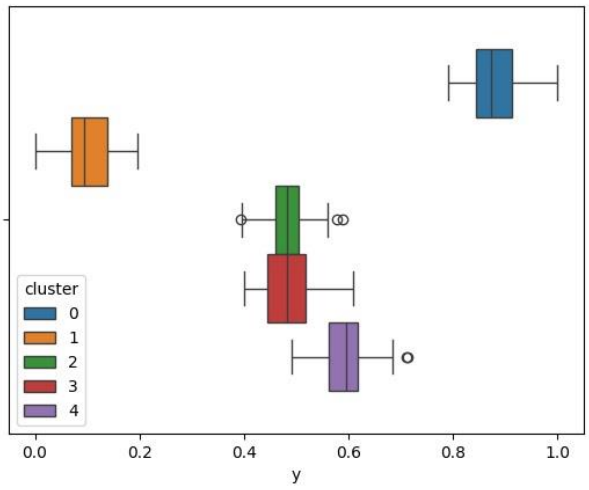


Рисунок 23 - диаграмма “box-plot” для признака y для lab2_checker.csv

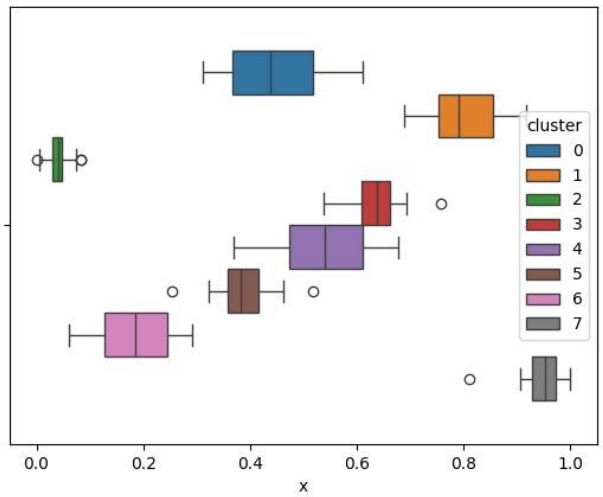


Рисунок 24 - диаграмма “box-plot” для признака x для lab2_noisymoons.csv

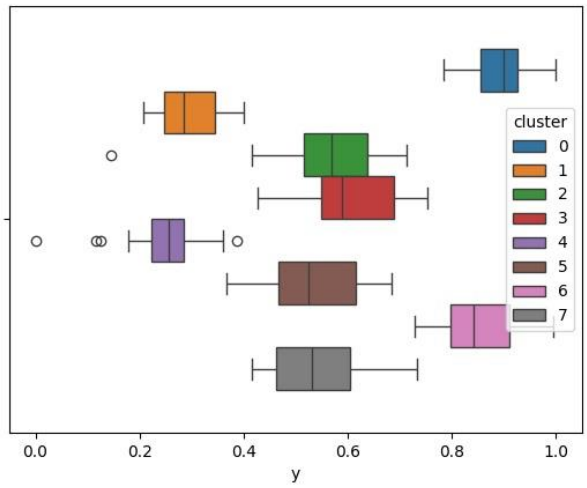


Рисунок 25 - диаграмма “box-plot” для признака y для lab2_noisymoons.csv

С помощью кластеризации алгоримом k-mean удалось разделить данные на кластеры.

Первые два датасета (lab2_blobs.csv и lab2_checker.csv), на мой взгляд, были разделены успешно. Возможны неточности для некоторых неоднозначных точек. Например, зелёный и фиолетовый кластеры в первом датасете и фиолетовый и красный кластеры во втором датасете имеют точки, которые визуально можно отнести и к одному, и к другому кластеру.

Третий датасет (lab2_noisymoons.csv) не удалось разделить на два кластера сложной формы. Вместо этого алгоритм предложил разделение таким образом, что полумесяцы содержат в себе несколько кластеров. Для эксперимента третий датасет был разделён на два кластера. Результат представленный на Рисунок 26, показывает, что кластеризовать датасет на два полумесяца методом k-mean не удалось.

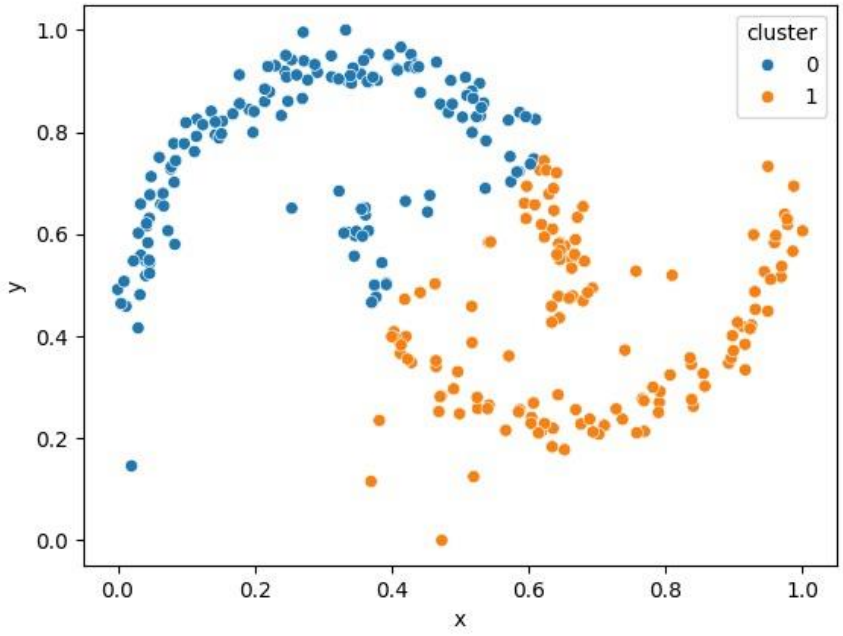


Рисунок 26 - диаграмма рассеяния результатов кластеризации на 2 кластера для lab2_noisymoons.csv методом k-means

Для каждого кластера было рассчитано количество точек, среднее, СКО, минимум и максимум. Результаты для трёх наборов данных представлены на Таблица 1, Таблица 2 и Таблица 3.

Таблица 1

lab2_blobs		
	x	y
claster = 0		
count	106	106
mean	0.660357	0.457495
std	0.063635	0.053037
min	0.535211	0.304931
max	0.849536	0.556882
claster = 1		
count	10	10
mean	0.188169	0.79409
std	0.062151	0.044957
min	0.079254	0.68434
max	0.268848	0.846848
claster = 2		
count	24	24
mean	0.623272	0.147838
std	0.115842	0.070087
min	0.402754	0.053518
max	0.84335	0.282783
claster = 3		
count	60	60
mean	0.821655	0.894281
std	0.063163	0.050281
min	0.672546	0.791121
max	1	1
claster = 4		
count	60	60
mean	0.140421	0.251639
std	0.055189	0.084429
min	0	0
max	0.265374	0.377383

По Таблица 1 видно, что claster=0 в таблице соответствует зелёному кластеру на графике, claster=1 в таблице соответствует красному кластеру, claster=2 в таблице соответствует фиолетовому кластеру, claster=3 в таблице соответствует оранжевому кластеру, claster=4 в таблице соответствует синему кластеру. Самым многочисленным кластером получился зелёный (106 точек), самым малочисленным – красный (10 точек), синий и оранжевый кластеры содержат 60 точек, фиолетовый содержит 24 точки. Самое большое СКО получилось у фиолетового (примерно 0.1 по x) и синего (примерно 0.8 по y), у остальных СКО составляет примерно 0.5-0.6.

Таблица 2

lab2_checker		
	x	y
claster = 0		
count	53	53
mean	0.122688	0.485998
std	0.061366	0.049018
min	0	0.400259
max	0.275951	0.608984
claster = 1		
count	46	46
mean	0.886668	0.481167
std	0.046826	0.047662
min	0.757464	0.392983
max	1	0.588039
claster = 2		
count	46	46
mean	0.504848	0.10393
std	0.045788	0.046222
min	0.399567	0
max	0.615417	0.195799
claster = 3		
count	50	50
mean	0.343477	0.594685
std	0.042936	0.053326
min	0.2469	0.491675
max	0.446452	0.713658
claster = 4		
count	55	55
mean	0.421779	0.877348
std	0.064685	0.043846
min	0.242856	0.790301
max	0.560222	1

По Таблица 2 видно, что claster=0 в таблице соответствует красному кластеру на графике, claster=1 в таблице соответствует зелёному кластеру, claster=2 в таблице соответствует оранжевому кластеру, claster=3 в таблице соответствует фиолетовому кластеру, claster=4 в таблице соответствует синему кластеру. Точки распределились по кластерам относительно равномерно, все кластеры имеют в себе 46-55 точек. Также СКО составляет у всех кластеров примерно 0.5-0.6.

Таблица 3

lab2_noisymoons		
claster = 0		
	x	y
count	28	28
mean	0.787061	0.283067
std	0.070722	0.058811
min	0.689229	0.207958
max	0.916304	0.400887
claster = 1		
count	42	42
mean	0.183748	0.849589
std	0.068148	0.066587
min	0.060668	0.727732
max	0.291649	0.995563
claster = 2		
count	47	47
mean	0.632788	0.604528
std	0.042132	0.093652
min	0.536798	0.427127
max	0.756821	0.752016
claster = 3		
count	31	31
mean	0.385959	0.541593
std	0.050397	0.095201
min	0.254105	0.382711
max	0.516882	0.684271
claster = 4		
count	28	28
mean	0.041227	0.561974
std	0.021596	0.113153
min	0	0.145795
max	0.08341	0.712705
claster = 5		
count	34	34
mean	0.530452	0.256461
std	0.079594	0.079487
min	0.369882	0
max	0.652416	0.387338
claster = 6		
count	24	24
mean	0.943219	0.530555
std	0.039938	0.097314
min	0.809548	0.384285
max	1	0.732646
claster = 7		
count	46	46
mean	0.444644	0.891583
std	0.084546	0.048124
min	0.311819	0.782882
max	0.610171	1

По Таблица 3 видно, что claster=0 в таблице соответствует оранжевому кластеру на графике, claster=1 в таблице соответствует розовому кластеру, claster=2 в таблице соответствует красному кластеру, claster=3 в таблице соответствует коричневому кластеру, claster=4 в таблице соответствует зелёному кластеру, claster=5 в таблице соответствует фиолетовому кластеру, claster=6 в таблице соответствует серому кластеру, claster=7 в таблице соответствует минему кластеру. Самым малочисленным кластером получился серый (24 точки), оранжевый (28 точки), коричневый (28 точек), Остальные кластеры имеют 31-46 точек. СКО у всех кластеров составляет примерно 0.4-0.1.

3. DBSCAN

С помощью функции clusterize (Листинг 3.1) была проведена кластеризация методом DBSCAN с выбранным eps и . min_samples.

Листинг 3.1

```
def clusterize(df, eps, min_samples):
    dbscan = DBSCAN(eps=eps, min_samples=min_samples)
    dbscan_clust = dbscan.fit_predict(df)
    norm_pd = pd.DataFrame(df, columns=['x', 'y'])
    norm_pd['cluster'] = dbscan_clust
    return norm_pd, dbscan_clust
```

Для данных lab2_blobs.csv были выбраны параметры eps=0.11, min_samples=5. Для данных lab2_checker.csv были выбраны параметры eps=0.1, min_samples=30. Для данных lab2_noisymoons.csv были выбраны параметры eps=0.1, min_samples=10.

Для lab2_blobs.csv и lab2_checker.csv были выбраны такие параметры, при которых все данные попадают в кластеры, так как на мой взгляд данные могут быть разделены без обозначения выбросов. При меньшем значении eps или большем значении min_samples на графиках появлялись точки, не входящие ни в один из кластеров (достаточно удалённые от общего скопления точек). При ещё большем значении min_samples кластеры, в которых было недостаточное количество точек, начинали полностью обозначаться как выбросы. При большем значении eps или меньшем значении min_samples кластеры начинали объединяться между собой.

Для данных lab2_noisymoons.csv были выбраны параметры параметры, которые оставляют удалённые от основного количества точек как выбросы, так как, на мой взгляд, эти данные действительно содержат выбросы. При меньшем

значении eps или большем значении min_samples в выбросы попадали точки, которые, на мой взгляд, должны входить в один из кластеров, например, правая часть нижнего полумесяца полностью уходила в выбросы. При большем значении eps или меньшем значении min_samples два кластера начинали объединяться между собой. При уменьшении обоих параметров происходила кластеризация на большое количество кластеров вместо двух.

Результаты кластеризации трёх датасетов методом DBSCAN представлены на Рисунок 27, Рисунок 28 и Рисунок 29.

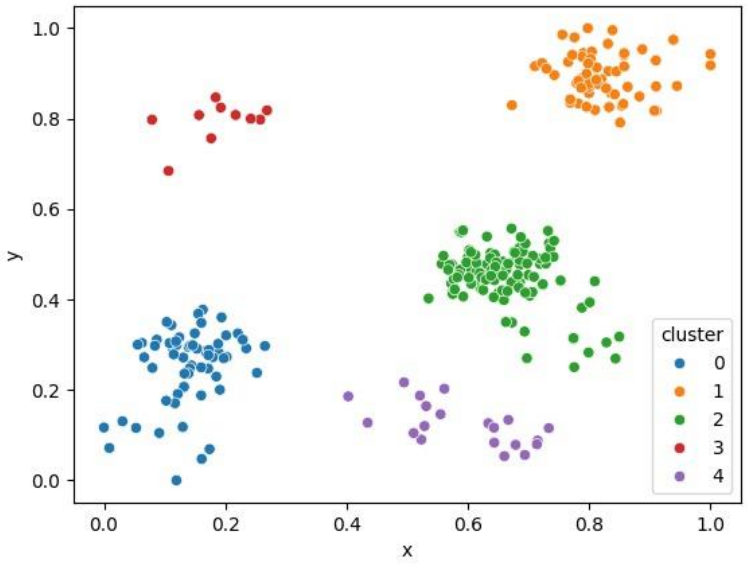


Рисунок 27 - диаграмма рассеяния результатов кластеризации для lab2_blobs.csv методом DBSCAN при eps=0.11, min_samples=5

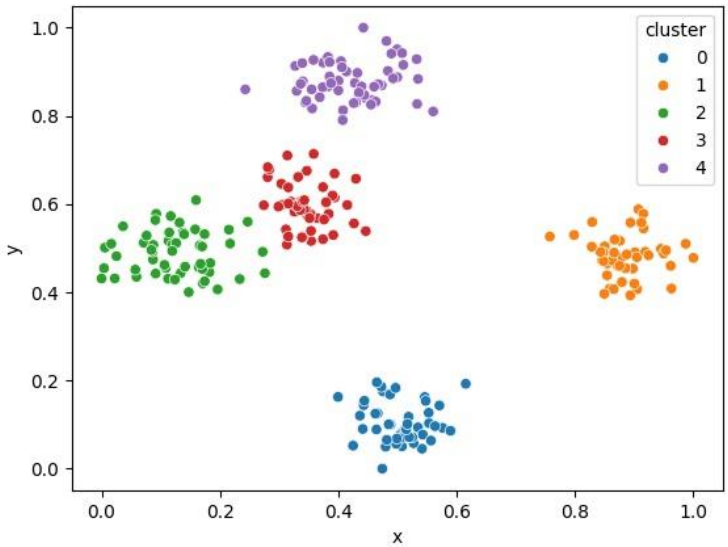


Рисунок 28 - диаграмма рассеяния результатов кластеризации для lab2_checker.csv методом DBSCAN при eps=0.1, min_samples=30

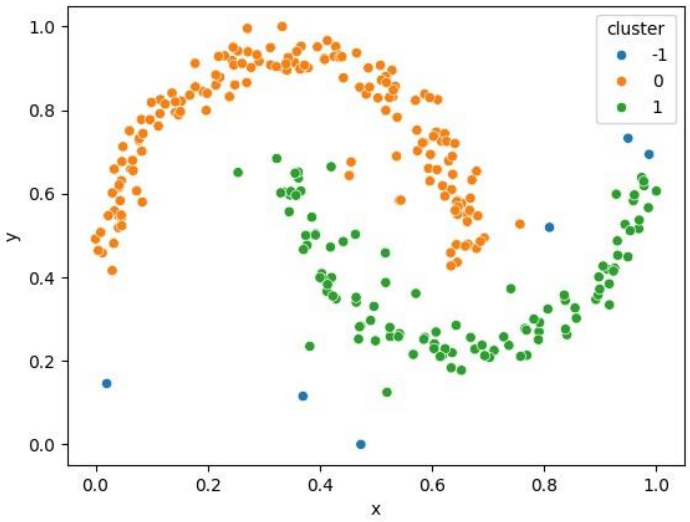


Рисунок 29 - диаграмма рассеяния результатов кластеризации для lab2_noisymoons.csv методом DBSCAN при eps=0.1, min_samples=10

Кластеризация методом DBSCAN прошла успешно. Удалось разделить на кластеры все три датасета. Была возможность регулировать получаемые кластеры, чтобы выбрать оптимальный вариант. Также была возможность варьировать количество точек, не входящих ни в один из кластеров, путём увеличения изменения параметров.

4. Иерархическая кластеризация

Была проведена иерархическая кластеризация при всех возможных параметрах linkage (ward и average). Для всех трёх наборов данных приведены дендрограммы на Рисунок 30 и Рисунок 31, Рисунок 32 и Рисунок 33, Рисунок 34 и Рисунок 35 соответственно.

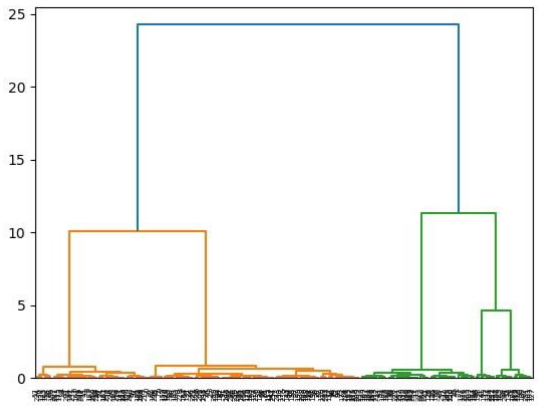


Рисунок 30 – дендрограмма иерархической кластеризации при linkage='ward' для lab2_blobs.csv

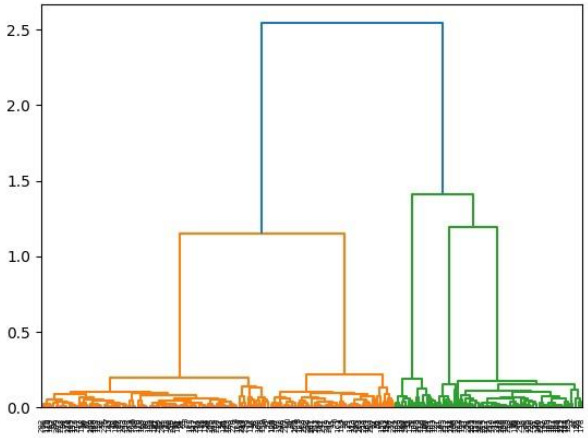


Рисунок 31 – дендрограмма иерархической кластеризации при linkage='avarage' для lab2_blobs.csv

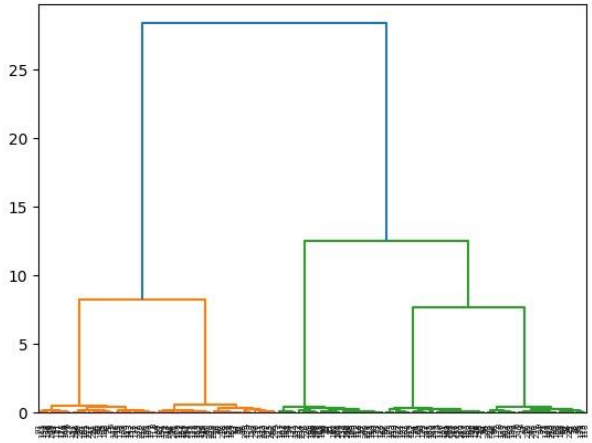


Рисунок 32 – дендрограмма иерархической кластеризации при linkage='ward' для lab2_checker.csv

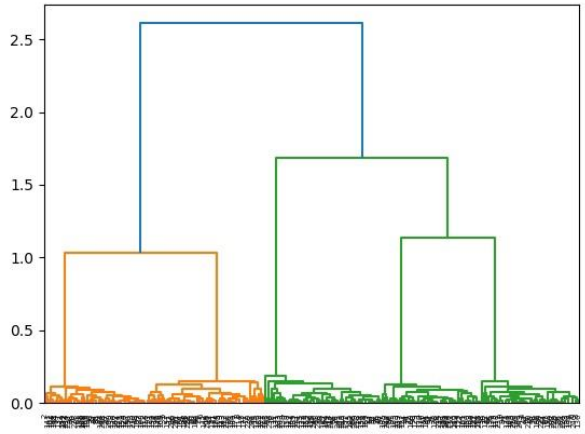


Рисунок 33 – дендрограмма иерархической кластеризации при linkage='avarage' для lab2_checker.csv

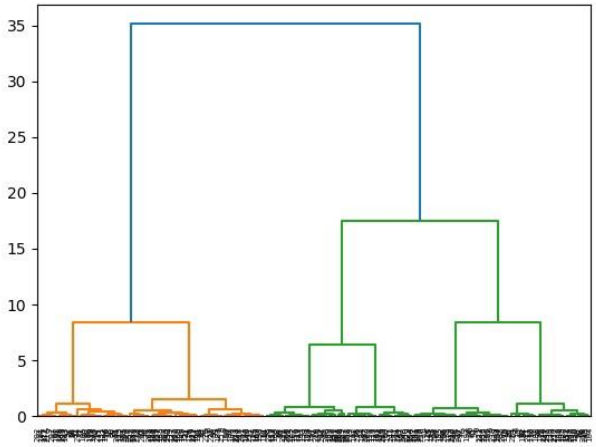


Рисунок 34 – дендрограмма иерархической кластеризации при linkage='ward' для lab2_noisymoons.csv

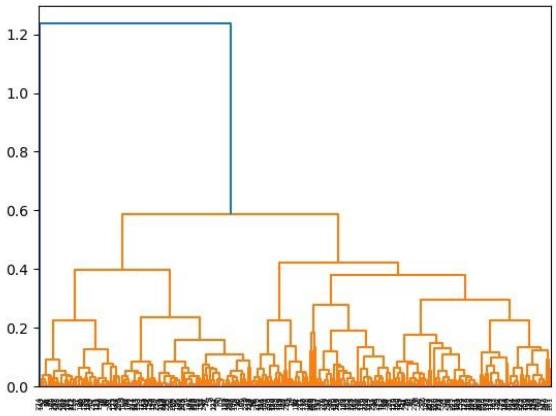


Рисунок 35 – дендрограмма иерархической кластеризации при linkage='avarage' для lab2_noisymoons.csv

Для данных lab2_blobs.csv и lab2_checker.csv было выбрано количество кластеров, равное 5. Для данных lab2_noisymoons.csv было выбрано количество кластеров, равное 2.

Диаграмма рассеяния для трёх датасетов представлены на Рисунок 36 и Рисунок 37, Рисунок 38 и Рисунок 39, Рисунок 40 и Рисунок 41.

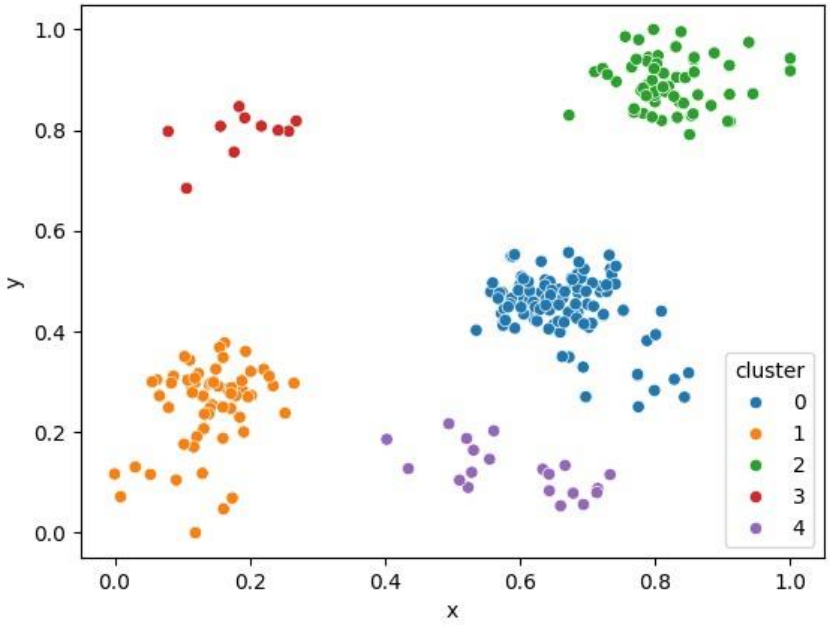


Рисунок 36 - диаграмма рассеяния иерархической кластеризации при linkage='ward' и количестве кластеров 5 для lab2_blobs.csv

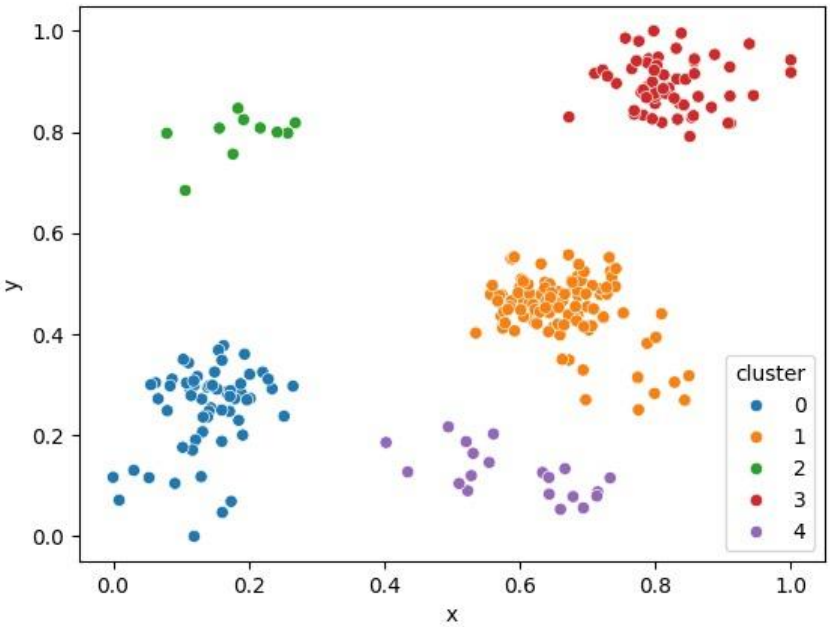


Рисунок 37 - диаграмма рассеяния иерархической кластеризации при linkage='avarage' и количестве кластеров 5 для lab2_blobs.csv

Как видно на Рисунок 36 и Рисунок 37, для lab2_blobs.csv при различных linkage получились одинаковые результаты. С помощью иерархической кластеризации удалось корректно разделить этот датасет на кластеры.

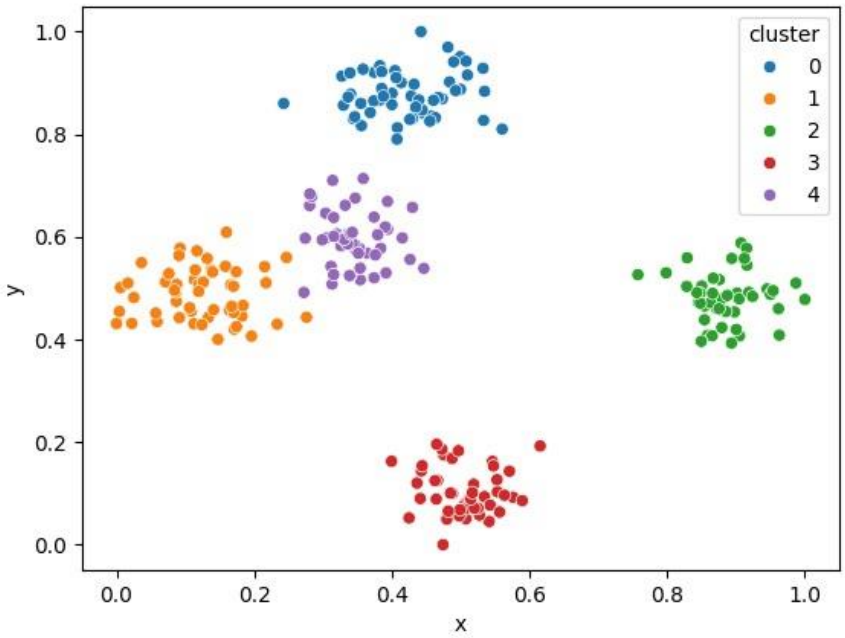


Рисунок 38 - - диаграмма рассеяния иерархической кластеризации при linkage='ward' и количестве кластеров 5 для lab2_checker.csv

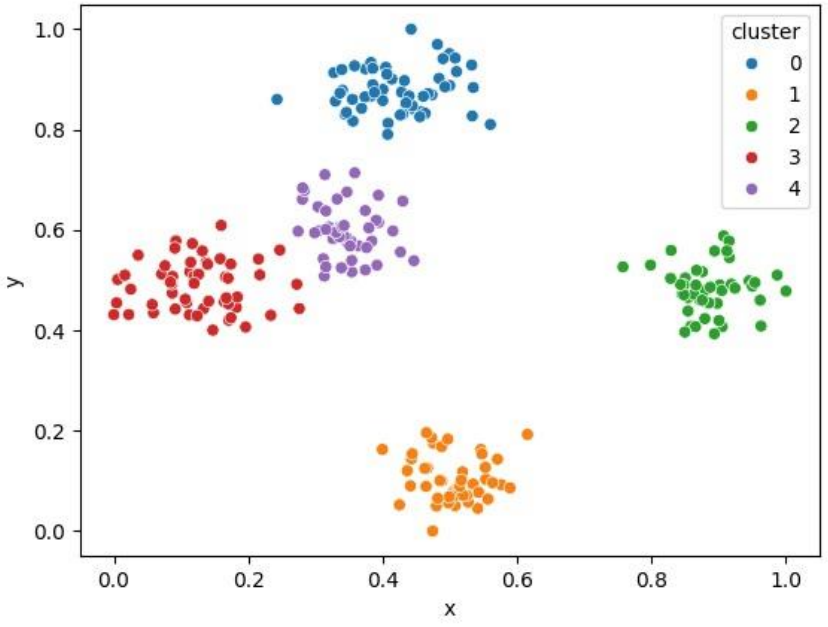


Рисунок 39 - диаграмма рассеяния иерархической кластеризации при linkage='avarage' и количестве кластеров 5 для lab2_checker.csv

Как видно на Рисунок 38 и Рисунок 39, для lab2_checker.csv при различных linkage получились схожие результаты. Отличия имеются только в неоднозначных точках, которые визуально могут относиться к обоим кластерам. С помощью иерархической кластеризации удалось корректно разделить этот датасет на кластеры.

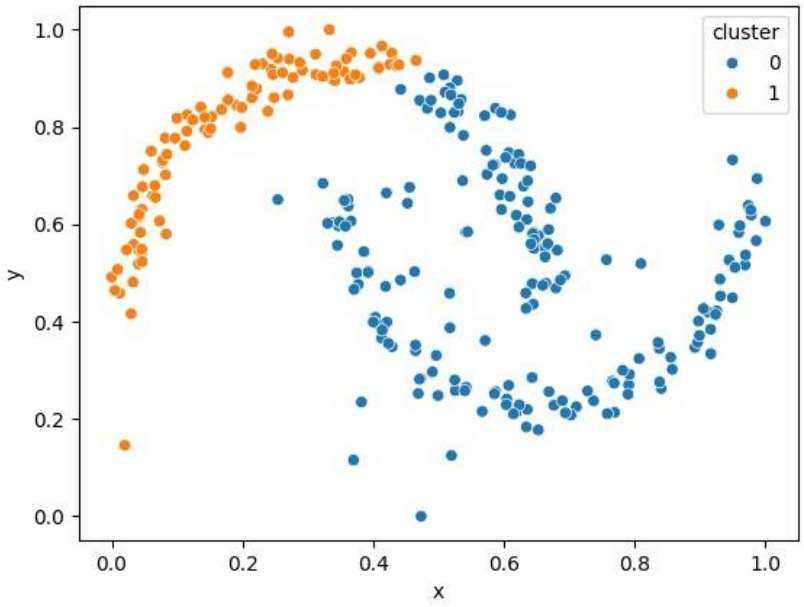


Рисунок 40 - диаграмма рассеяния иерархической кластеризации при linkage='ward' и количестве кластеров 2 для lab2_noisymoons.csv

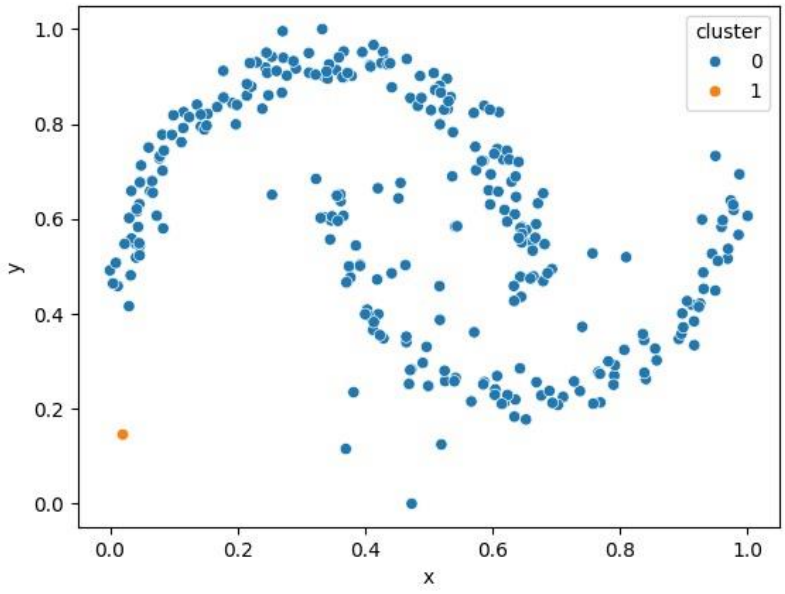


Рисунок 41 - диаграмма рассеяния иерархической кластеризации при linkage='avarage' и количестве кластеров 2 и lab2_noisymoons.csv

Как видно на Рисунок 40 и Рисунок 41, для lab2_noisymoons.csv при linkage='ward' данные разделились на два кластера в соотношении примерно 1:2, а при linkage='avarage' в один кластер вошла только одна точка, а во второй кластер вошли все остальные точки. С помощью иерархической кластеризации не удалось разделить этот датасет на кластеры в форме полумесяца.

Таким образом, для lab2_blobs.csv и lab2_checker.csv удалось получить корректное разделение на кластеры всеми тремя методами кластеризации. Для lab2_noisymoons.csv удалось разделить данные на два полумесяца только методом dbscan.

5. Изучение набора данных с большим количеством признаков

Для набора данных отмеченного lab2_winequality_red.csv была проведена предобработка (загрузка в формате датафрейма, проверка корректности данных, нормализация) с помощью методов, описанных в параграфе 1.

Было проведено исследование оптимального количества кластеров методом локтя и методом силуэтов.

С помощью функции show_elbow_method (Листинг 2.1) было проведено исследование оптимального количества кластеров методом локтя. На Рисунок 42 представлен график, помогающий определить оптимальное количество кластеров методом локтя.

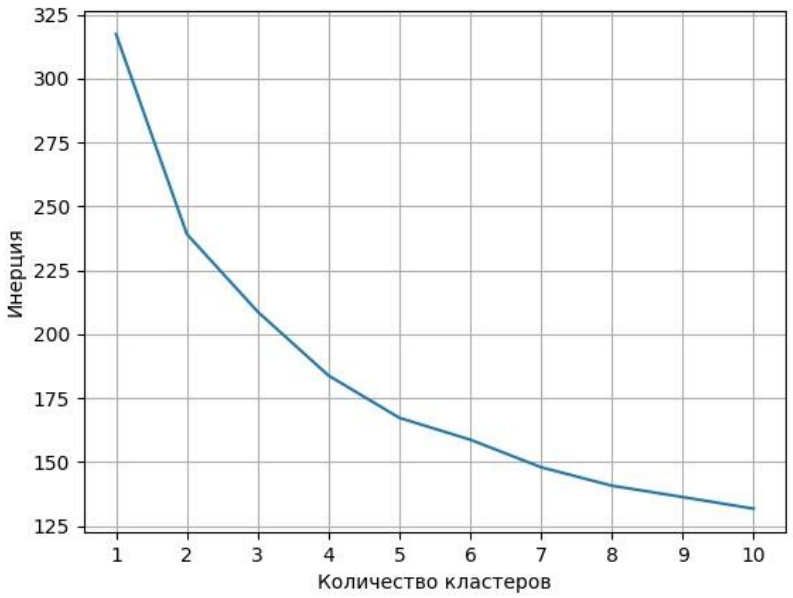


Рисунок 42 - график, помогающий определить оптимальное количество кластеров методом локтя

По Рисунок 42 видно, что по методу локтя оптимальное количество кластеров равно 5-8.

С помощью функции show_elbow_method (Листинг 2.2Листинг 2.1) было проведено исследование оптимального количества кластеров методом силуэта.

На Рисунок 43 представлены графики, помогающие определить оптимальное количество кластеров методом силуэта для трёх наборов данных соответственно.

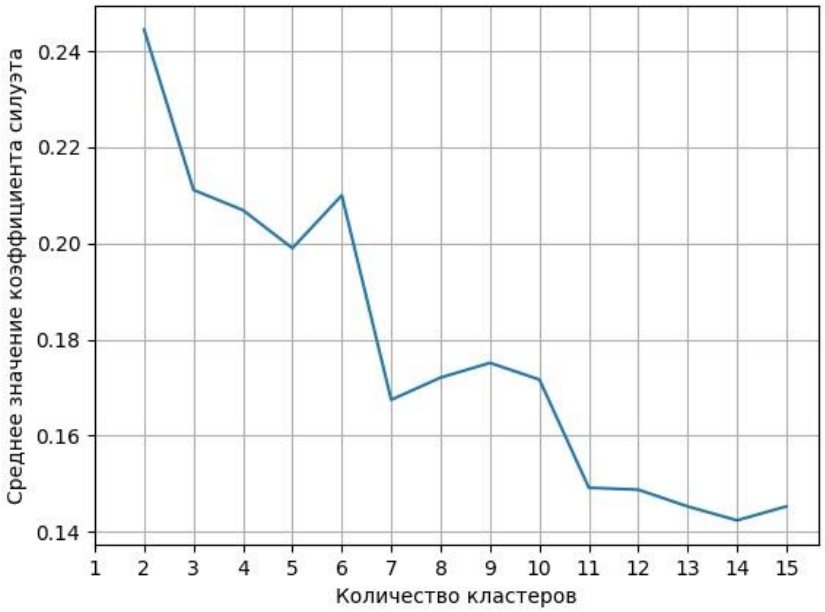


Рисунок 43 - график, помогающий определить оптимальное количество кластеров методом силуэта

По Рисунок 43 видно, что методом силуэта получается, что оптимально будет не делить датасет на кластеры. Однако, график имеет локальный максимум при 6-ти кластерах.

Также были проведены исследования для метода иерархичесткой кластеризации, и были получены схожие результаты.

Таким образом, для этих двух методов стоит рассматривать количество кластеров 5-8.

Были проведены эксперименты, чтобы определить оптимальный метод разбиения данных на кластеры.

В качестве первого метода был выбран DBSCAN. Этим методом не удалось разделить данные на оптимальные кластеры, так как получалось, что практически все данные уходили в выброс (Рисунок 44), либо, что абсолютное большинство данных принадлежат одному кластеру (Рисунок 45).

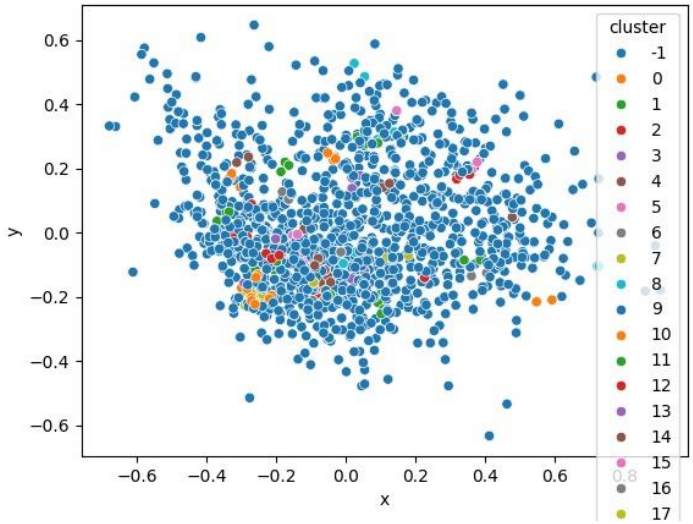


Рисунок 44 - диаграмма рассеяния с понижением размерности методом PCA неудачного использования dbscan 1 при eps=0.1 и min_samples=4

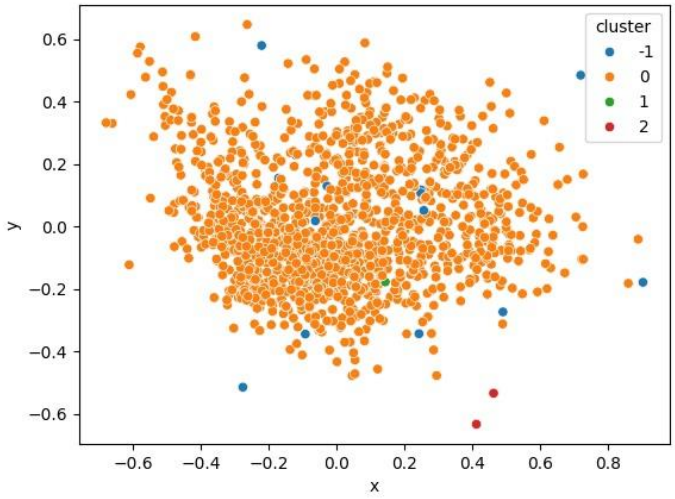


Рисунок 45- диаграмма рассеяния с понижением размерности методом PCA неудачного использования dbscan 2 eps=0.2 и min_samples=4

Далее были выбраны методы k-means и иерархическая кластеризация.

Эти методы давали похожие результаты, однако результат метода k-means был нестабилен. На Рисунок 46 и Рисунок 47 приведены диаграммы рассеяннния с понижением размерности методом TSNE для двух последовательно запущенных экспериментов методом k-means. Такой результат получается из-за итоговых различных центроид.

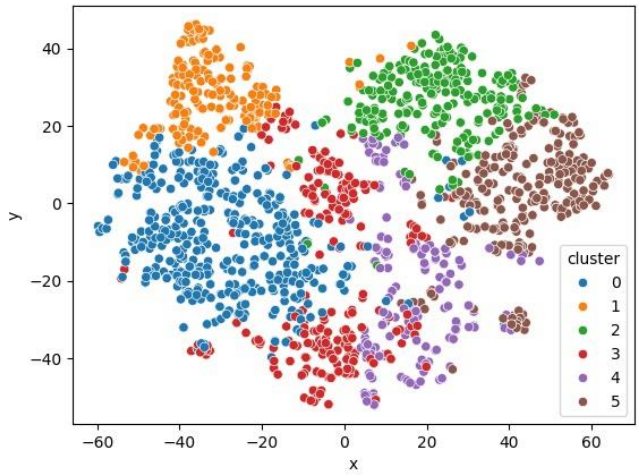


Рисунок 46 - диаграмма рассеяния с понижением размерности методом TSNE для k-means 1

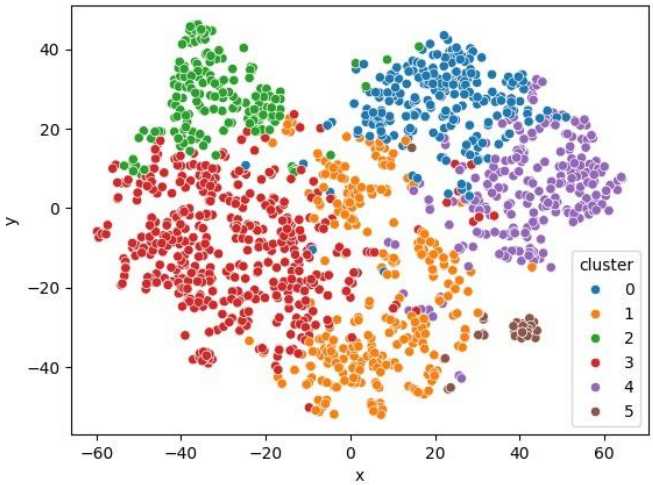


Рисунок 47 - диаграмма рассеяния с понижением размерности методом TSNE для k-means 2

Методом иерархичесткой кластеризации были получены стабильные результаты. Таким образом, для кластеризации был выбран именно метод иерархической кластеризации с количеством кластеров 6.

Были проведены эксперименты с количеством кластеров 5-8 и выбрано количество кластеров 6, так как при значении 5 разделение данных было заметно только на признаке citric acid и alcohol, остальные параметры были распределены практически равномерно для всех кластеров. Для значений больше 6-ти практически все кластеры были плохоразлечимы, но образовывались 1-2 кластера, которые собирали в себе все точки, имеющие по одному из признаков отличающиеся значения.

Например, при 8-им кластерах, имелось 7 схожих между собой кластеров и 1, имеющий нестандартные значения для residual sugar (Рисунок 48).

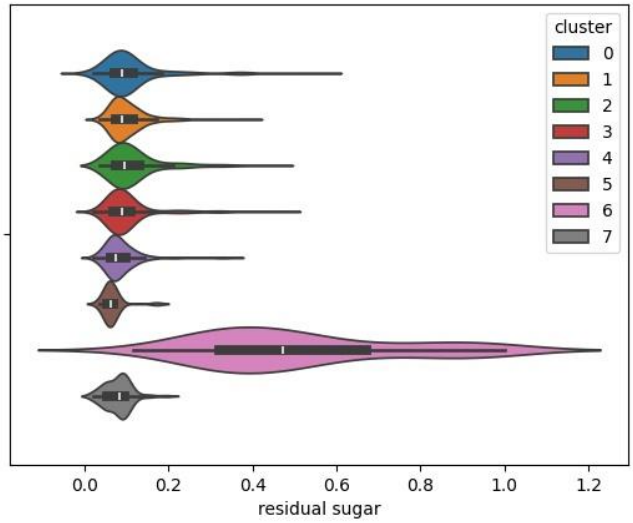


Рисунок 48 - распределение признака residual sugar при иерархической кластеризации на 8 кластеров

При значении 6 в кластеризации были задействовано большинство признаков.

На Рисунок 49 и Рисунок 50 представлены диаграммы рассеяния с понижением размерности для кластеризации иерархическим методом с количеством кластеров 6. Кластеры визуально различимы, но в разделении, согласно этим графикам, есть погрешности. Например, часть зелёного кластера зашла на зону коричневого, синиц кластер зашёл на кластеры вокруг.

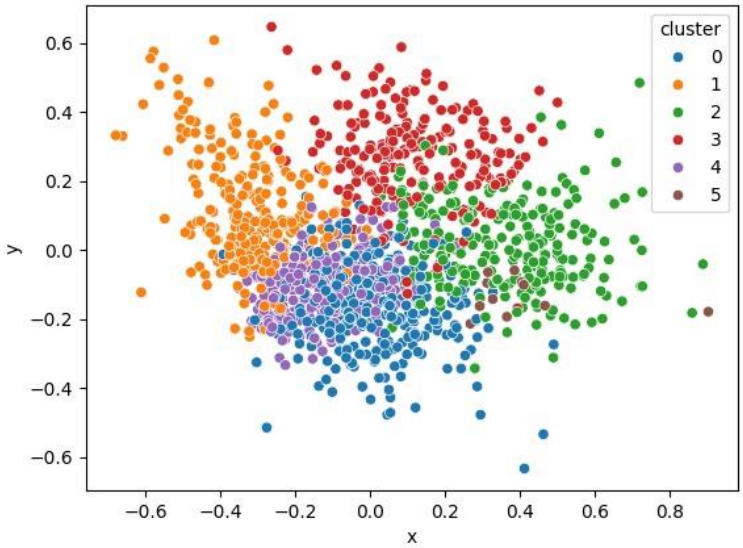


Рисунок 49 - диаграмма рассеяния с понижением размерности методом PCA для иерархической кластеризации

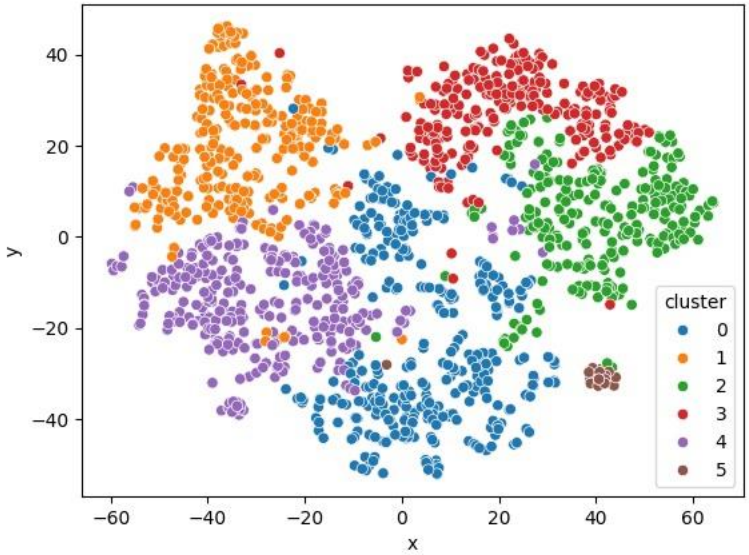


Рисунок 50 - диаграмма рассеяния с понижением размерности методом TSNE для иерархической кластеризации

Для более подробного анализа зависимости кластеров от признаков, была построена violin-диаграмма с помощью метода violinplot библиотеки seaborn. Результаты представлены на Рисунок 51, Рисунок 52, Рисунок 53, Рисунок 54, Рисунок 55, Рисунок 56, Рисунок 57, Рисунок 58, Рисунок 59, Рисунок 60 и Рисунок 61. Также была построена попарная диаграмма рассаяния, оценка плотности (Рисунок 62).

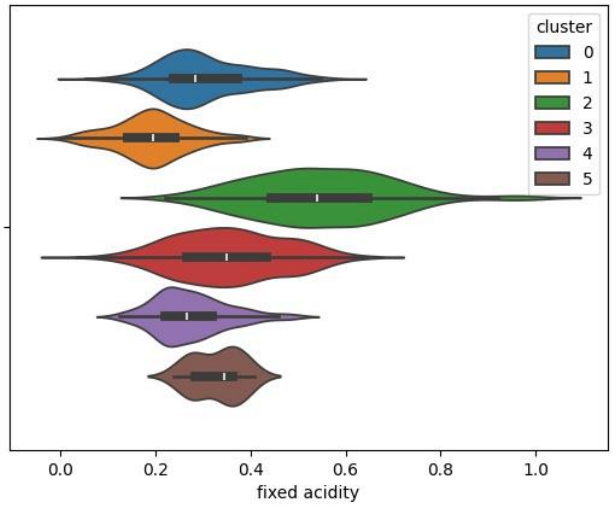


Рисунок 51 - violin-диаграмма признака fixed acidity

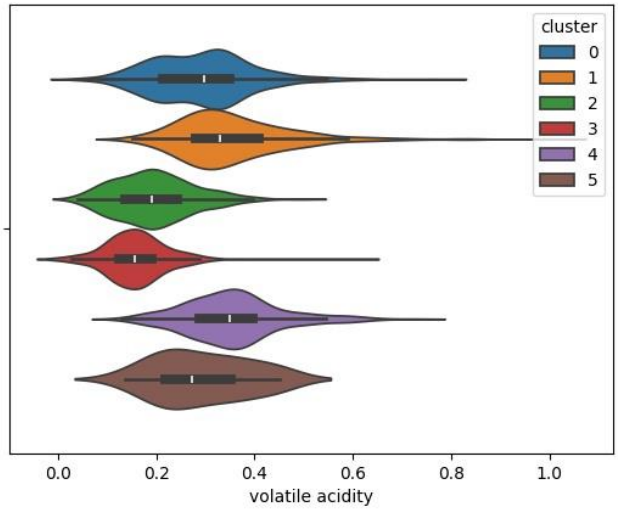


Рисунок 52 - violin-диаграмма признака volatile acidity

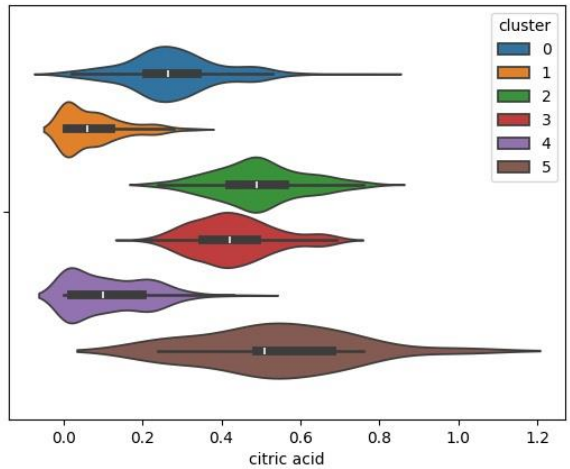


Рисунок 53 - violin-диаграмма признака citric acid

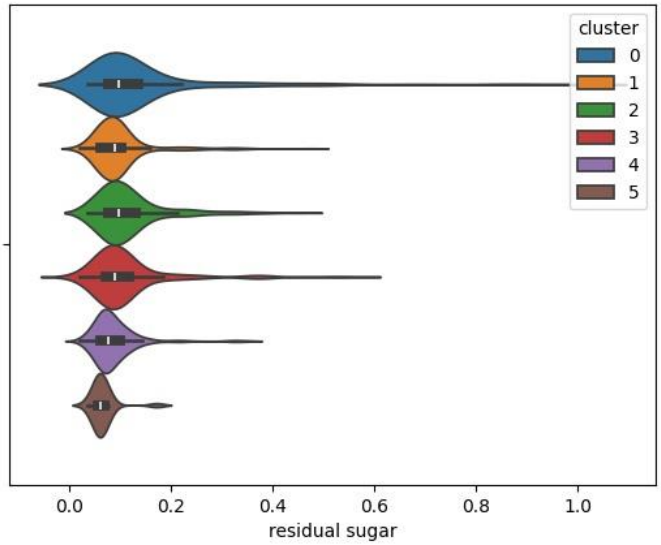


Рисунок 54 - violin-диаграмма признака residual sugar

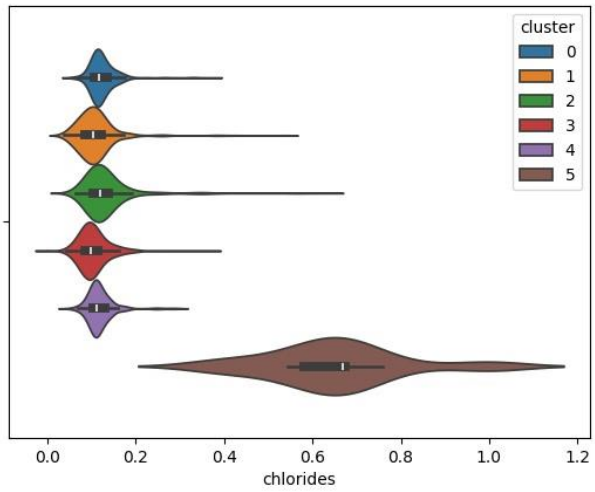


Рисунок 55 - violin-диаграмма признака chlorides

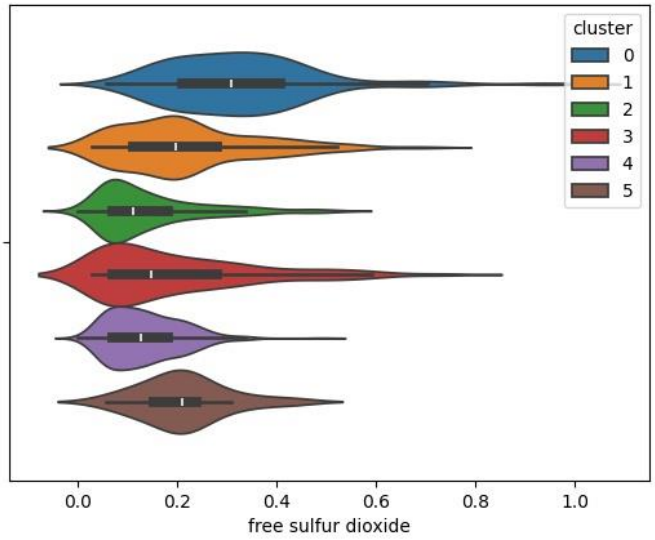


Рисунок 56 - violin-диаграмма признака free sulfur dioxide

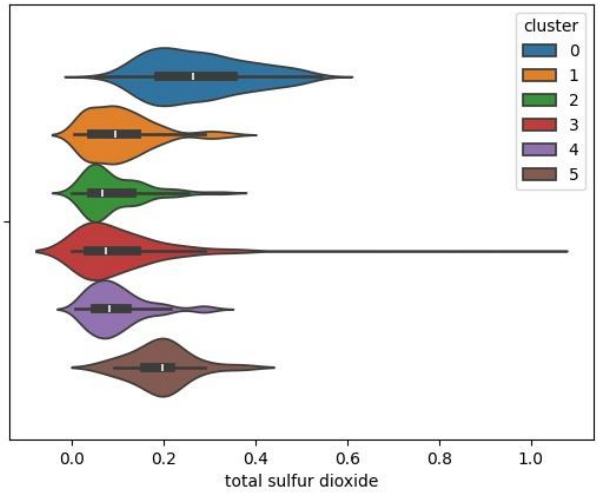


Рисунок 57 - violin-диаграмма признака total sulfur dioxide

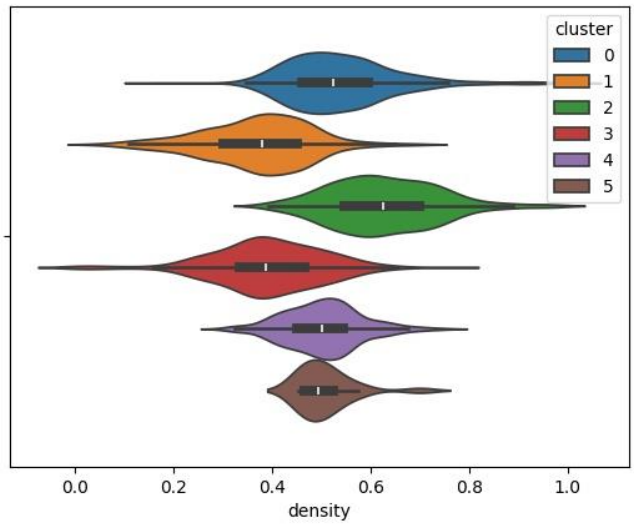


Рисунок 58 - violin-диаграмма признака density

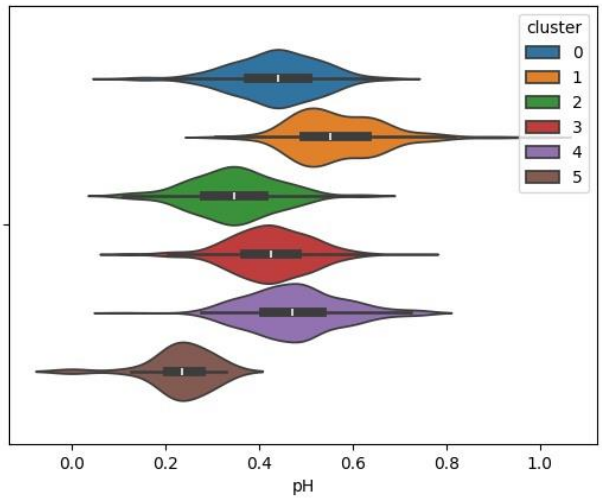


Рисунок 59 - violin-диаграмма признака pH

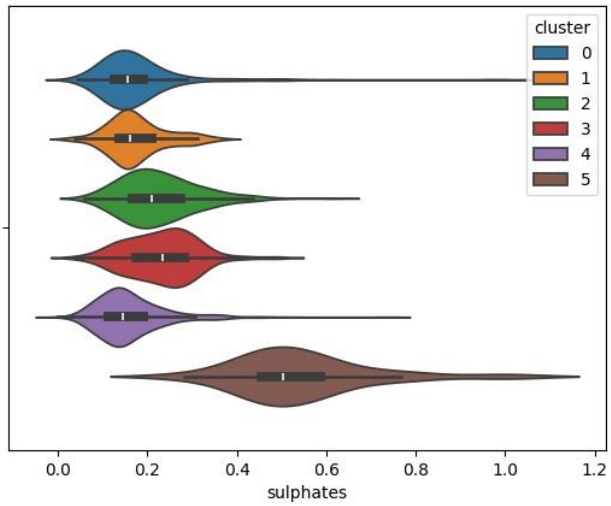


Рисунок 60 - violin-диаграмма признака sulphates

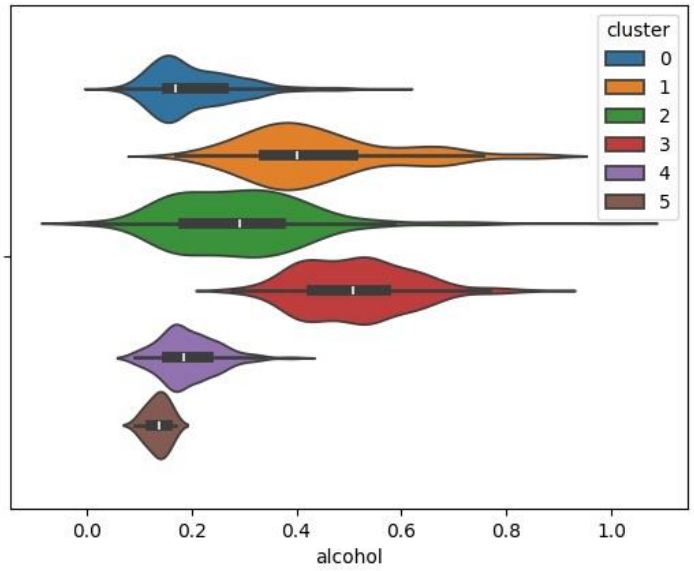


Рисунок 61 - violin-диаграмма признака alcohol

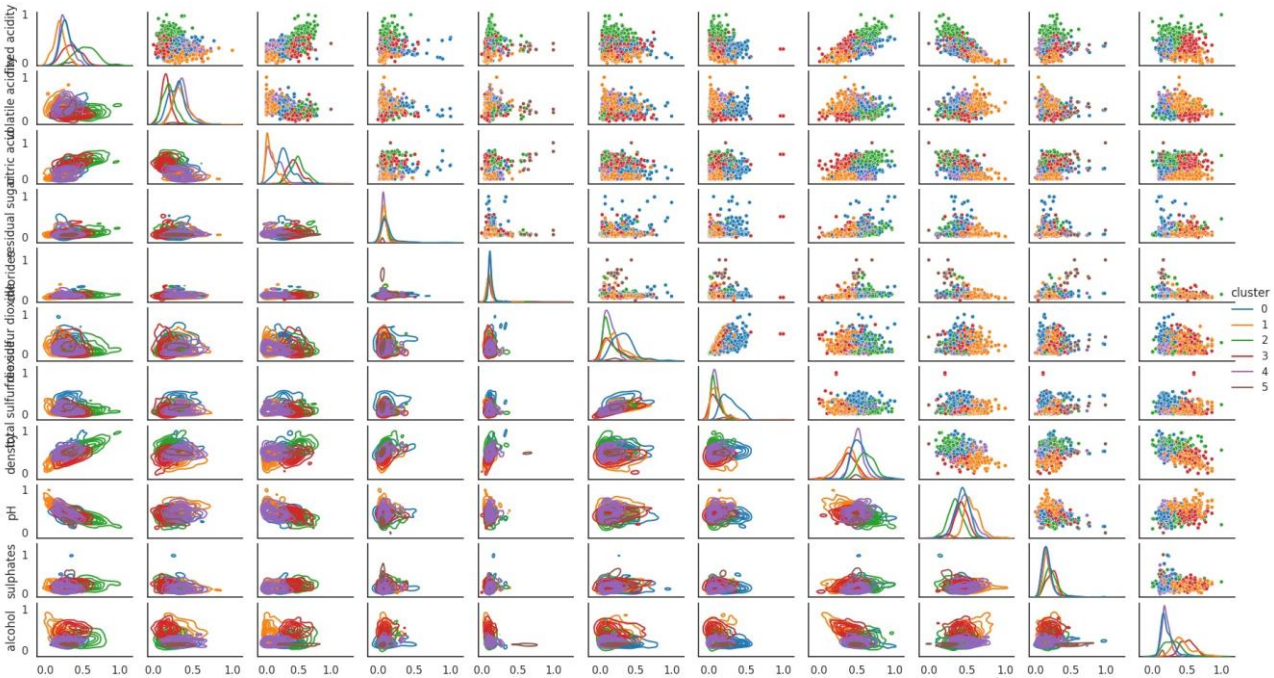


Рисунок 62 - попарная диаграмма характеристик и диаграмма плотности признаков.

В разделении активно участвовали все значения acidity, density, alcohol.

Практически не участвовали значения residual sugar, dioxide.

Таким образом, большинство признаков поучаствовали в кластеризации.

Заключение.

В ходе лабораторной работы были изучены методы кластеризации наборов данных, такое как K-Means, DBSCAN и Иерархическая кластеризация.

Было проведено исследование оптимального количества кластеров методами локтя и силуэта.

Была проведена кластеризация трёх наборов данных тремя разными методами с подбором оптимального количество кластеров или оптимального значения параметров.

Были построены диаграммы Вороного для оценки границ кластеров.

Также была проведена кластеризация для набора данных с большим количеством признаков. Был выбран наилучший метод кластеризации, определено количество кластеров для этого метода и построены графики (графики рассеяния с понижением плотности, violin-диаграммы, опарные диаграммы рассеяния, графики плотности признаков) для оценки результата кластеризации