Санкт-Петербургский политехнический университет Петра Великого

Институт машиностроения, материалов и транспорта

[**Высшая школа автоматизации и робототехники**](https://www.spbstu.ru/structure/vysshaya_shkola_avtomatizatsii_i_robototekhniki/)

**ЛАБОРАТОРНАЯ РАБОТА №4**

**«Методы кластеризация»**

по дисциплине

«Математические методы интеллектуальных технологий»

Выполнил

студент гр.3331506/60401 <*подпись*> Комаров А.Е.

Руководитель

<*подпись*> Орлова С.Р.

«\_\_\_» \_\_\_\_\_\_\_\_\_\_ 2021 г.

Санкт-Петербург

2021

***Цель работы:*** освоение методов решения задачи кластеризации

***Задачи работы:***

* освоить основные алгоритмы кластеризации;
* изучить различные способы измерения качества кластеризации;
* научиться применять полученные знания об алгоритмах и навыки по формированию признаков для получения максимально качественных моделей.

# Ход работы и результаты

# Часть 1. Анализ и подготовка данных

1. Описание базы

В таблице 1 представлено описание базы *House Prices - Advanced Regression Techniques*.

Таблица 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Признак | Описание признака | Тип | Крайние значения | |
| Id | Номер записи | int64 | 1 | 1460 |
| SalePrice | Цена продажи объекта в долларах | int64 | 34900 | 755000 |
| MSSubClass | Тип жилья | int64 | 20 | 190 |
| MSZoning | Класс продажи | object |  |  |
| LotFrontage | Длина в футах улицы, соединенной с объетом | float64 | 21 | 313 |
| LotArea | Размер в квадратных футах | int64 | 1300 | 215245 |
| Street | Тип подъездной работы | object |  |  |
| Alley | Тип подъездной аллеи к объекту | object |  |  |
| LotShape | Общая форма объета | object |  |  |
| LandContour | Плокскостность объета | object |  |  |
| Utilities | Имеющиеся коммунальные услуги (газ, вода, электричество) | object |  |  |
| LotConfig | Конфигурация объета | object |  |  |
| LandSlope | Склон объета | object |  |  |
| Neighborhood | Район в пределах города Эймс | object |  |  |
| Condition1 | Близость к главной дорге или железной дороге | object |  |  |
| Condition2 | Близость к главной дорге или железной дороге (при наличии второй) | object |  |  |
| BldgType | Тип жилья | object |  |  |
| HouseStyle | Стиль жилья | object |  |  |
| OverallQual | Общее качество материалов и отделки | int64 | 1 | 10 |
| OverallCond | Общая оценка состояния | int64 | 1 | 9 |
| YearBuilt | Дата постройки | int64 | 1872 | 2010 |
| YearRemodAdd | Дата реконструкции | int64 | 1950 | 2010 |
| RoofStyle | Тип крыши | object |  |  |
| RoofMatl | Материал крыши | object |  |  |
| Exterior1st | Внешнее покрытие дома | object |  |  |
| Exterior2nd | Наружное покрытие дома (если используется более одного материала) | object |  |  |
| MasVnrType | Тип облицовки кладки | object |  |  |
| MasVnrArea | Площадь облицовки кладки в квадратных футах | float64 | 0 | 1600 |
| ExterQual | Качество наружных материалов | object |  |  |
| ExterCond | Текущее состояние материала снаружи | object |  |  |
| Foundation | Тип фундамента | object |  |  |
| BsmtQual | Высота подвала | object |  |  |
| BsmtCond | Общее состояние подвала | object |  |  |
| BsmtExposure | Подвальные стены на уровне сада или выхода | object |  |  |
| BsmtFinType1 | Качество готовой площади подвала | object |  |  |
| BsmtFinSF1 | Готовые квадратные футы типа 1 | int64 | 0 | 5644 |
| BsmtFinType2 | Качество второй законченной области подвала (если есть) | object |  |  |
| BsmtFinSF2 | Готовые квадратные футы типа 2 | int64 | 0 | 1474 |
| BsmtUnfSF | Незавершенные квадратные футы подвала | int64 | 0 | 2336 |
| TotalBsmtSF | Общая площадь подвала в квадратных футах | int64 | 0 | 6110 |
| Heating | Тип отопления | object |  |  |
| HeatingQC | Качество и состояние отопления | object |  |  |
| CentralAir | Наличие центральное кондиционирование | object |  |  |
| Electrical | Описание электрической системы | object |  |  |
| 1stFlrSF | Квадратные футы на первом этаже | int64 | 334 | 4692 |
| 2ndFlrSF | Квадратные футы второго этажа | int64 | 0 | 2065 |
| LowQualFinSF | Квадратные футы с отделкой низкого качества (все этажи) | int64 | 0 | 572 |
| GrLivArea | Квадратные футы жилой площади над уровнем земли | int64 | 334 | 5642 |
| BsmtFullBath | Полные ванные комнаты в подвале | int64 | 0 | 3 |
| BsmtHalfBath | Полуподвальные ванные комнаты | int64 | 0 | 2 |
| FullBath | Полные ванные комнаты над цокольным этажом | int64 | 0 | 3 |
| HalfBath | Полуванны над цокольным этажом | int64 | 0 | 2 |
| Bedroom | Количество спален над цокольным этажом | int64 | 0 | 8 |
| Kitchen | Количество кухонь | int64 | 0 | 3 |
| KitchenQual | Качество кухни | object |  |  |
| TotRmsAbvGrd | Общее количество комнат над цокольным этажом (без ванных комнат) | int64 | 2 | 14 |
| Functional | Рейтинг функциональности дома | object |  |  |
| Fireplaces | Количество каминов | int64 | 0 | 3 |
| FireplaceQu | Качество камина | object |  |  |
| GarageType | Расположение гаража | object |  |  |
| GarageYrBlt | Год постройки гаража | float64 | 1900 | 2010 |
| GarageFinish | Внутренняя отделка гаража | object |  |  |
| GarageCars | Размер гаража по вместимости машин | int64 | 0 | 4 |
| GarageArea | Размер гаража в квадратных футах | int64 | 0 | 1418 |
| GarageQual | Качество гаража | object |  |  |
| GarageCond | Состояние гаража | object |  |  |
| PavedDrive | Подъездная дорога с твердым покрытием | object |  |  |
| WoodDeckSF | Площадь деревянного настила в квадратных футах | int64 | 0 | 857 |
| OpenPorchSF | Площадь открытой веранды в квадратных футах | int64 | 0 | 547 |
| EnclosedPorch | Площадь веранды в квадратных футах | int64 | 0 | 552 |
| 3SsnPorch | Площадь веранды на три сезона в квадратных футах | int64 | 0 | 508 |
| ScreenPorch | Площадь крыльца в квадратных футах | int64 | 0 | 480 |
| PoolArea | Площадь бассейна в квадратных футах | int64 | 0 | 738 |
| PoolQC | Качество бассейна | object |  |  |
| Fence | Качество забора | object |  |  |
| MiscFeature | Прочие функции, не включенные в другие категории (лифт, сарай, теннисный корт и др.) | object |  |  |
| MiscVal | Ценность дополнительных функций | int64 | 0 | 15500 |
| MoSold | Месяц продажи | int64 | 1 | 12 |
| YrSold | Год продажи | int64 | 2006 | 2010 |
| SaleType | Тип продажи | object |  |  |
| SaleCondition | Состояние продажи | object |  |  |

1. Устранение “технических” ошибок

Первым шагом были выявлены признаки, не содержащие в себе достаточное количество информации для последующего анализа (малое количество ненулевых строк), и удалены из дальнейшего рассмотрения.

Была получена таблица с количеством не-*NA/null* записей в каждом признаке, после чего обнаружено, что 6 признаков могут быть удалены из дальнейшего рассмотрения. Ниже представлены фрагмент кода и таблица 2 – фрагмент таблицы *count.xlsx*, отсортированной по количеству не-*NA/null* записей.

DataFrame.count().to\_excel('count.xlsx')  
  
DataFrame = DataFrame.drop(['PoolQC', 'Alley', 'FireplaceQu', 'Fence', 'MiscFeature', 'LotFrontage'], axis=1)

Таблица 2

|  |  |
| --- | --- |
| Признак | count() |
| PoolQC | 7 |
| MiscFeature | 54 |
| Alley | 91 |
| Fence | 281 |
| FireplaceQu | 770 |
| LotFrontage | 1201 |
| GarageType | 1379 |
| GarageYrBlt | 1379 |
| GarageFinish | 1379 |
| … | |

Далее были удалены все строки, содержащие *NA/null* записи.

DataFrame = DataFrame.dropna()

1. Устранение “смысловых” ошибок

Для устранения выбросов в числовых признаках использовался следующий фрагмент кода.

DataFrame = DataFrame[(DataFrame[["GrLivArea"]] < 4000).all(axis=1)]  
DataFrame = DataFrame[(DataFrame[["GarageCars"]] < 4).all(axis=1)]  
DataFrame = DataFrame[(DataFrame[["OverallQual"]] > 2).all(axis=1)]  
DataFrame = DataFrame[(DataFrame[["1stFlrSF"]] < 2000).all(axis=1)]  
DataFrame = DataFrame[(DataFrame[["GrLivArea"]] < 2500).all(axis=1)]  
DataFrame = DataFrame[(DataFrame[["TotalBsmtSF"]] < 2000).all(axis=1)]  
DataFrame = DataFrame[(DataFrame[["TotalBsmtSF"]] > 450).all(axis=1)]  
DataFrame = DataFrame[(DataFrame[["GarageArea"]] < 1000).all(axis=1)]

После устранения выбросов было произведено удаление всех строк, содержащих *NA/null* записи, и обновление номеров строк, что отражает представленный ниже фрагмент кода.

DataFrame = DataFrame.dropna()  
DataFrame.reset\_index(drop=True)

Далее был произведен перевод категориальных признаков в числовые.

for feature in DataFrame.columns:  
 if DataFrame[feature].dtype == object:  
 DataFrame[feature] = pandas.Categorical(DataFrame[feature])  
 sex\_map\_train = dict(zip(DataFrame[feature].cat.codes, DataFrame[feature]))  
 DataFrame[feature] = DataFrame[feature].cat.codes

1. Определение целевой переменной и итогового набора данных

В качестве целевой переменной выбрана цена продажи объекта в долларах. Также для кластеризации был добавлен новый признак, отражающий стоимость квадратного фута жилой площади *PricePerFoot*.

Для определения итогового набора данных были построены матрицы корреляции признаков между собой. Вследствие большого количества признаков и следующей из этого сложности анализа корреляций признаков с целевой переменной на основании графиков, были найдены корреляции всех признаков с ценой продажи объекта с помощью фрагмента кода, представленного ниже.

corr = DataFrame[['SalePrice'] + DataFrame.columns.to\_list()].corr().iloc[0]  
corr.sort\_values().to\_excel('correlation.xlsx')

На основании полученных данных были выбраны для дальнейшего анализа признаки с большими значениями корреляции. В таблице 3 представлены корреляции выбранных признаков со стоимостью жилья.

Таблица 3

|  |  |
| --- | --- |
| Признак | Корреляция |
| YearBuilt | 0,577645775 |
| PricePerFoot | 0,587639934 |
| GrLivArea | 0,69984561 |
| OverallQual | 0,786033352 |
| SalePrice | 1 |

На рисунке 1 представлена диаграмма рассеяния датасета, где по осям x и y расположены два признаки, полученные после применения метода главных компонент к выбранным для анализа пяти признакам.

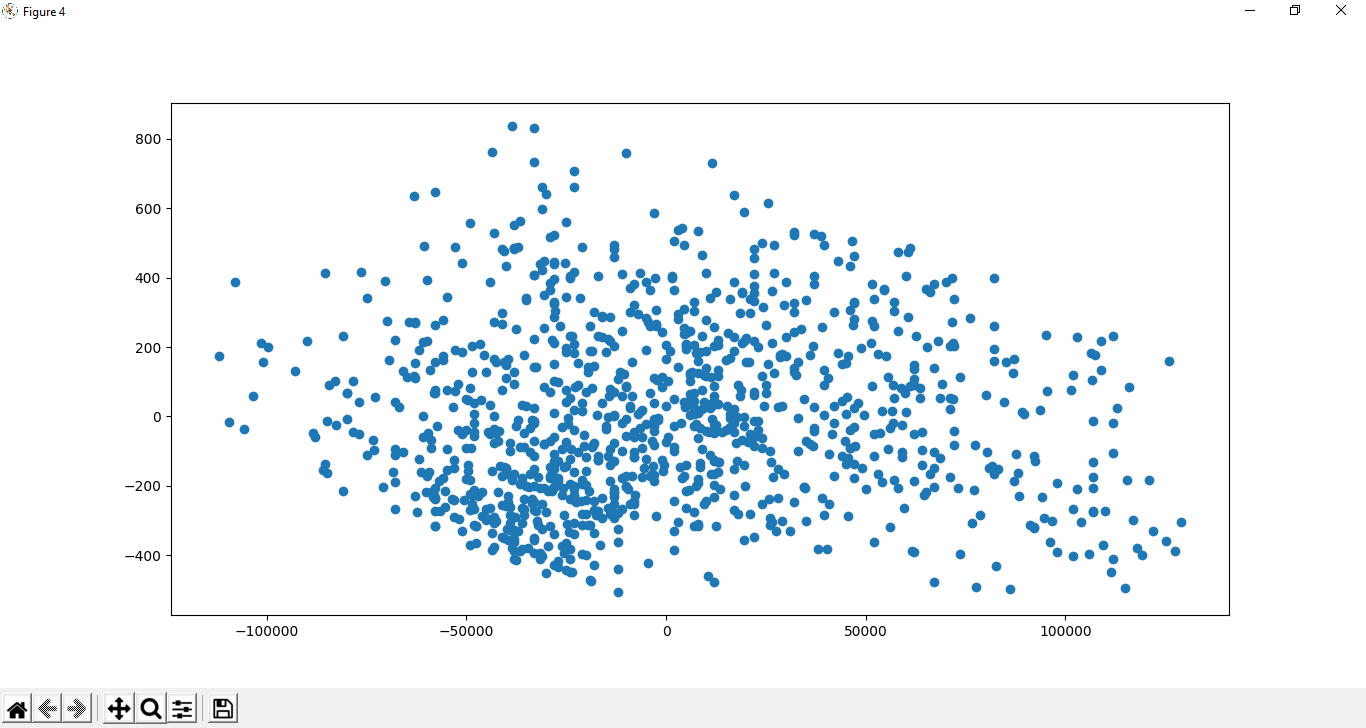


Рисунок 1

# Часть 2. Кластеризация

1. Используемые метрики

В качестве одной из метрик используется *Silhouette Coefficient* - силуэт. Силуэт одного объекта вычисляется по формуле ниже.

где – среднее расстояние между рассматриваемым объектом и всеми объектами того же класса,

- среднее расстояние между рассматриваемым объектом и всеми объектами ближайшего класса.

Силуэт группы объектов определяется как среднее значение силуэтов для каждого объекта в отдельности.

Значение силуэта меняется в диапазоне от -1 до 1, где -1 соответствует неправильной кластеризации, 1 – соответствует плотной кластеризации с хорошо разделенными кластерами, что говорит о качественной работе. Значение около 0 означает наличие перекрывающихся кластеров.

Второй используемой метрикой является *Davies-Bouldin Index* – .

где – среднее расстояние между всеми объектами класса и центроидом этого класса,

– расстояние между центроидами кластеров и .

Таким образом, для каждого кластера находиться наиболее схожий, вычисляются расстояния между выбранными парами кластеров и находиться среднее всех этих значений.

Наименьшее возможное значение равно нулю. Близкие к нулю значения соответствуют лучшей кластеризации.

Так как датасет не предназначен для кластеризации и выделение кластеров практически не представляется возможным, было принято решение о выделении трех групп стоимости (сбалансированных по количеству элементов), удаление ряда объектов на границах групп и повторное применение методов кластеризации к описанному измененному датасету. Так как после описанных действий получаются размеченные данные, для анализа качества кластеризации используется также метрика, учитывающая разметку, а именно скорректированный индекс Рэнда. На формуле ниже представлен простой индекс Рэнда.

где – число пар объектов, находящихся в одном кластере в размеченных и кластеризованных данных,

- число пар объектов, находящихся в разных кластерах в размеченных и кластеризованных данных,

– число всевозможных пар объектов.

Так как индекс Рэнда не обеспечивает близость к нулю для случайного разбиения, применяется скорректированный индекс Рэнда, которому соответствует формула ниже.

где – ожидаемый индекс Рэнда,

– максимальный индекс Рэнда.

Формула для вычисления скорректированного индекса Рэнда для простейшего варианта с двумя кластерами также может быть записана в виде, представленном ниже.

где – число пар объектов, находящихся в одном кластере в размеченных данных и в разных кластерах в кластеризованных данных,

– число пар объектов, находящихся в разных кластерах в размеченных данных и в одном кластере в кластеризованных данных.

Значение скорректированного индекса Рэнда изменяется в диапазоне от -1 до 1, где 1 соответствует правильной кластеризации, а 0 – случайному разбиению на кластеры.

1. Решение задачи кластеризации

Под задачей кластеризации понимаем выделение различных групп жилья на основании выбранных признаков.

Результат кластеризации методом *K-Means* на оригинальных данных представлен на рисунке 2.

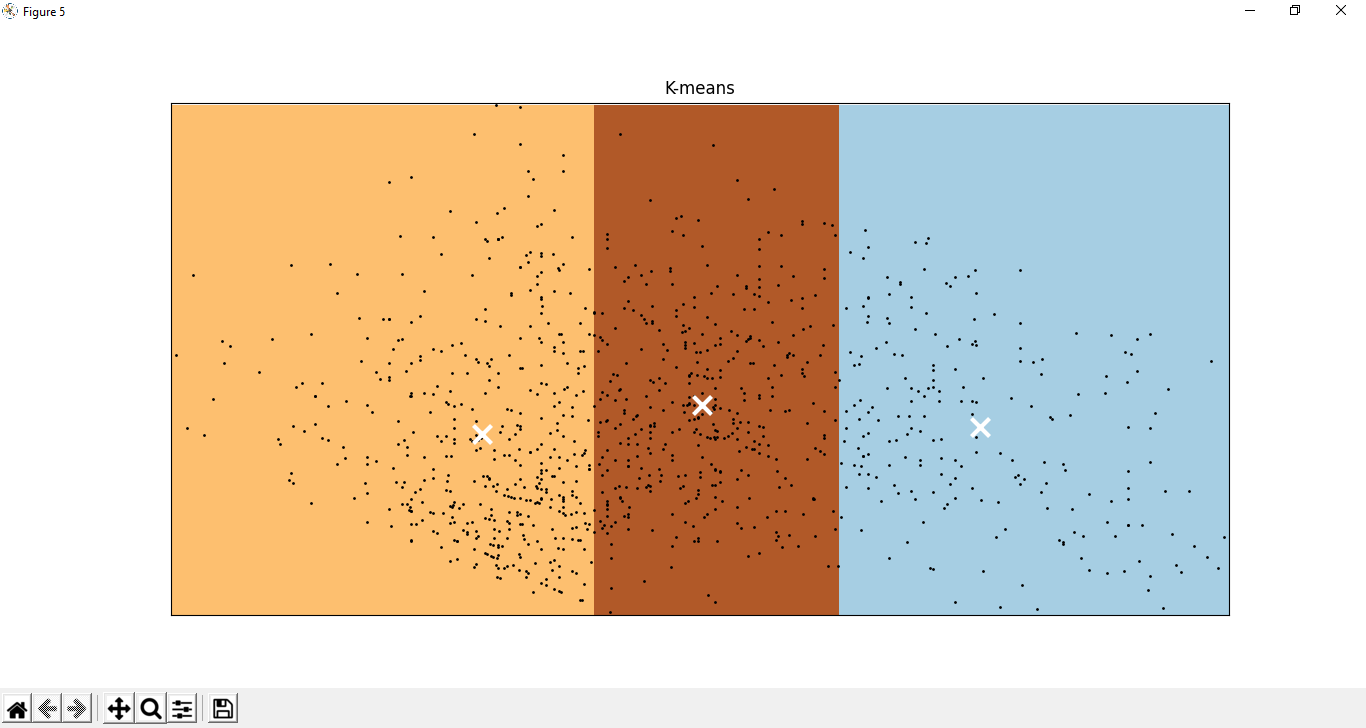


Рисунок 2

Значения метрик для кластеризации методом *K-Means*:

silhouette\_score 0.5595855732301378

davies\_bouldin\_score 0.5337344473594228

По результатам работы методы *K-Means* были выделены 3 кластера. Значения метрик далеки от идеальных значений (*silhouette\_score* = 1, *davies\_bouldin\_score* = 0), что может быть объяснено большой скученностью объектов на диаграмме рассеяния, что приводит к тому, что размеры кластеров близки к значениям расстояния между центроидами класстеров.

Кластеризация методом смеси Гауссовых распределений *Gaussian Mixture* работает как вероятностное улучшение над алгоритмом *K-Means*. На первоначальном этапе методом *K-Means* происходит разметка кластеров. *K-Means* ожидает, что кластеры имеют формы гиперсфер, но очень часто кластеры в лучшем случае на плоскости описываются эллипсами или в худшем случае более сложными формами. *Gaussian Mixture* после первоначальной разметки кластеров размечает точки вероятностными принадлежностями ко всем кластерам и в дальнейшем на основе этих данных происходит переопределение кластеров на основе не только *K-Means*, но и гладкой гауссовской модели.

Результат кластеризации методом смеси Гауссовых распределений *Gaussian Mixture* на оригинальных данных представлен на рисунке 3.

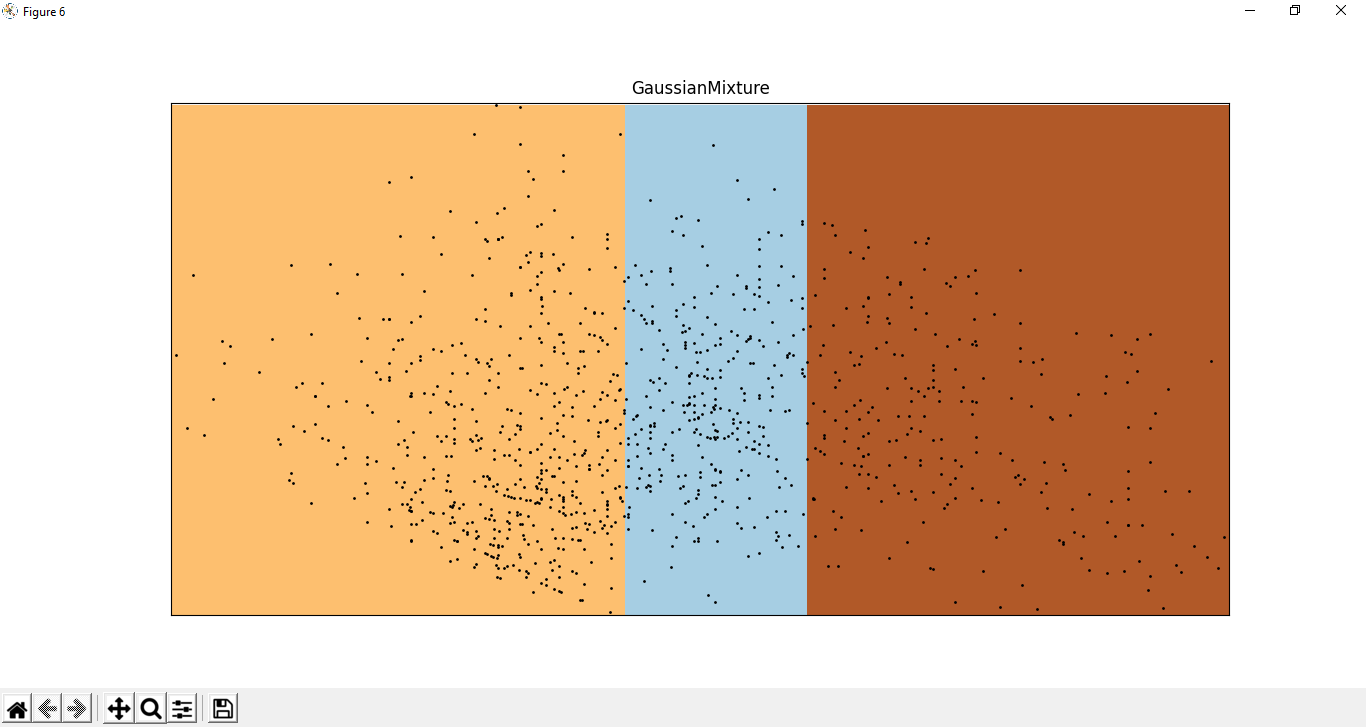


Рисунок 3

Значения метрик для кластеризации методом *Gaussian Mixture*:

silhouette\_score 0.5355486394787062

davies\_bouldin\_score 0.5231766034686424

Значения метрик также как и у метода *K-Means* далеки от идеальных значений.

При кластеризации методом *DBSCAN* первый шагом берется случайная точка на плоскости и присваивается ей метка первого кластера. Далее выбирается ближайший сосед к этой точке и ей тоже присваивается метка первого кластера. Процесс останавливается, когда расстояние от всех точек, уже обработанного кластера, до ближайшей нерассмотренной ранее точки превышает заданное. После этого берется ранее не рассмотренная точка и ей присваивается метка нового кластера. Процесс повторяется пока не останется таких точек, что расстояния от них до ближайших точек кластеров превышает заданное.

Метод *DBSCAN* хорошо работает на плотных облаках. Также данный метод хорошо борется с выбросами. Особенностью метода является возможность не задавать количество кластеров.

Результат кластеризации методом *DBSCAN* на оригинальных данных представлен на рисунке 4.

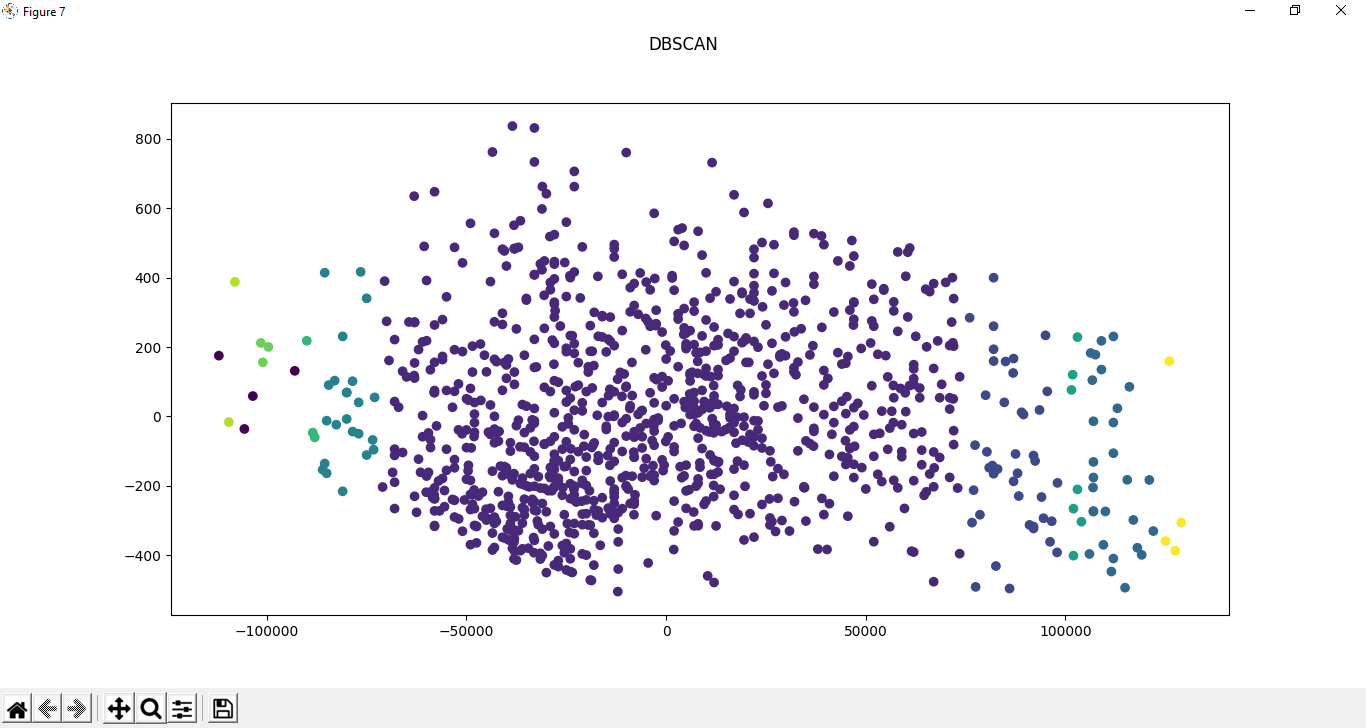


Рисунок 4

Значения метрик для кластеризации методом *DBSCAN*:

silhouette\_score 0.1785502539801736

davies\_bouldin\_score 0.8627225965441054

Значения метрик близки к идеальным значениям, однако как можно заметить из диаграммы рассеяния, представленной на рисунке 4, в результате работы метода *DBSCAN* был выделен один большой кластер, содержащий значительное большинство всех объектов датасета, и несколько кластеров на границе основного облака всех объектов, что не может быть названо качественной кластеризацией.

Выбранный датасет содержит одно облако объектов на диаграмме рассеяния и методами оптимизации метрик было установлено, что наиболее выгодно объединять все точки в один кластер, или выбирать только одну точки и все остальное считать выбросами. Соответственно всегда получается один кластер.

Алгоритм спектральной кластеризации основан на случайном «блуждании» по графу и спектральному разложению матрицы смежности. Поскольку нужно учитывать веса ребер в графах, то в матрице смежности записываются не просто наличие связей между узлами, а их расстояния. Исходя из этого, логично предположить, что выборка должна быть нормализованной. В противном случае при расчете степеней узлов, значения будут сильно отличаться и алгоритм не сможет корректно определить количество связных компонент на графе.

Результат работы методом спектральной кластеризации на нормализованных данных представлен на рисунке 5.

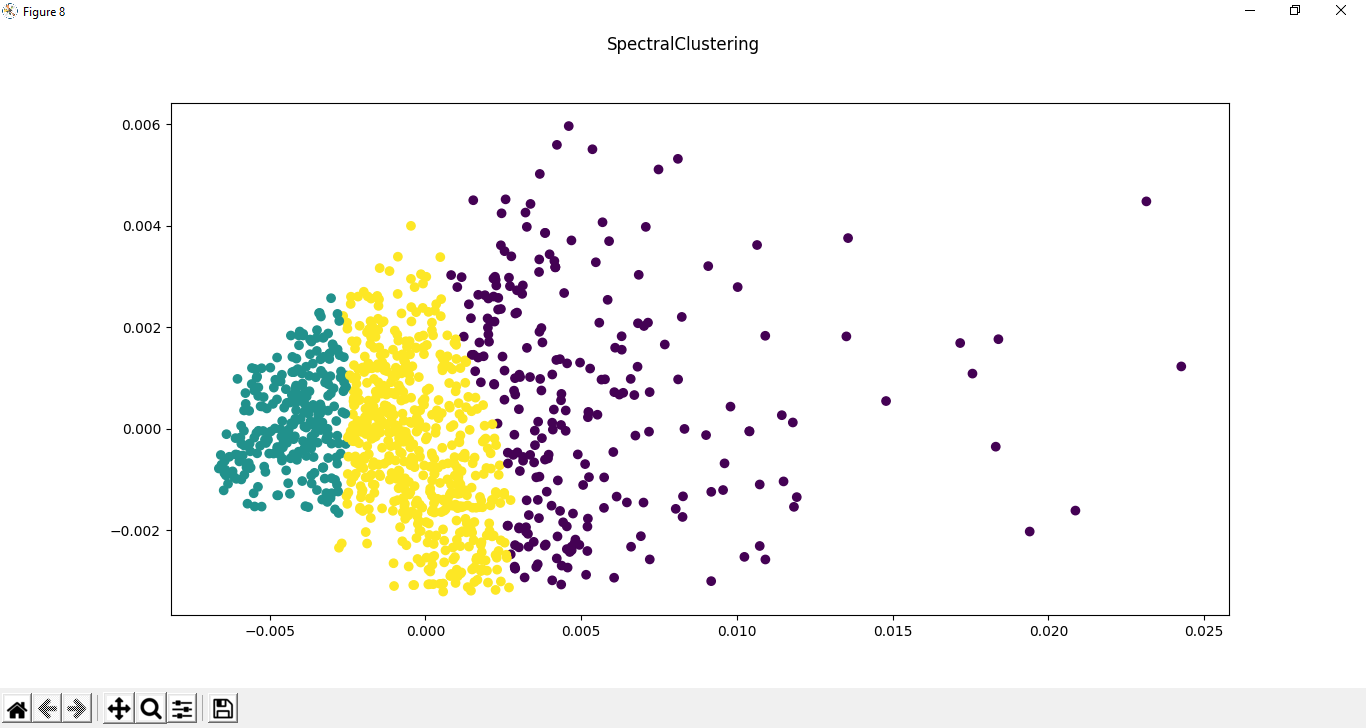


Рисунок 5

Значения метрик для метода спектральной кластеризации:

silhouette\_score 0.353604584639317

davies\_bouldin\_score 0.9045063643763945

Значения обоих метрик хуже значений для кластеризации методами *K-Means* и *Gaussian Mixture*.

В таблице 4 представлены время обучения и время предсказания (где возможно) для всех примененных ранее методов.

Таблица 4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *K-Means* | *Gaussian Mixture* | *DBSCAN* | Спектральная кластеризация |
| Обучение | 0.538 | 0.140 | 0.013 | 0.349 |
| Предсказание | 0.005 | 0.018 | - | - |

Диаграмма рассеяния датасета с размеченными данными и искусственно добавленными границами между кластерами представлена на рисунке 6. Результаты работы всех описанных ранее методов представлены на рисунках 7-10.

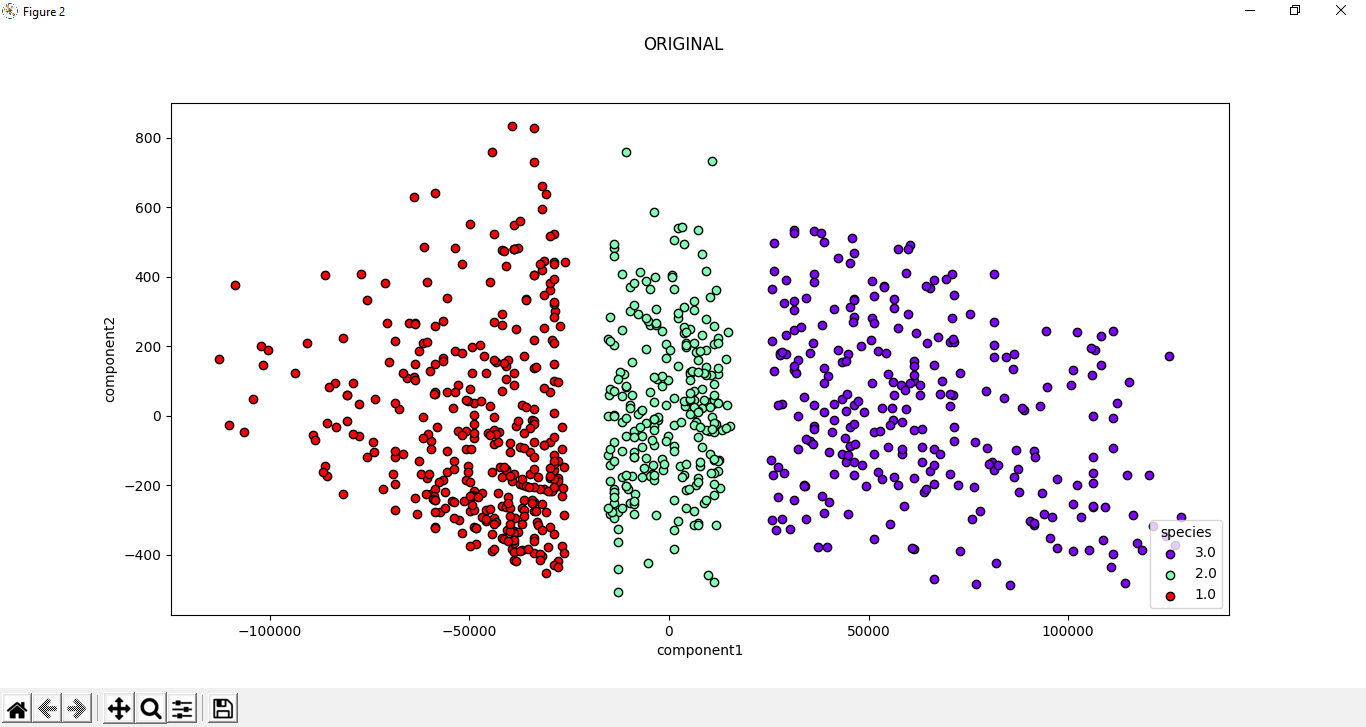


Рисунок 6

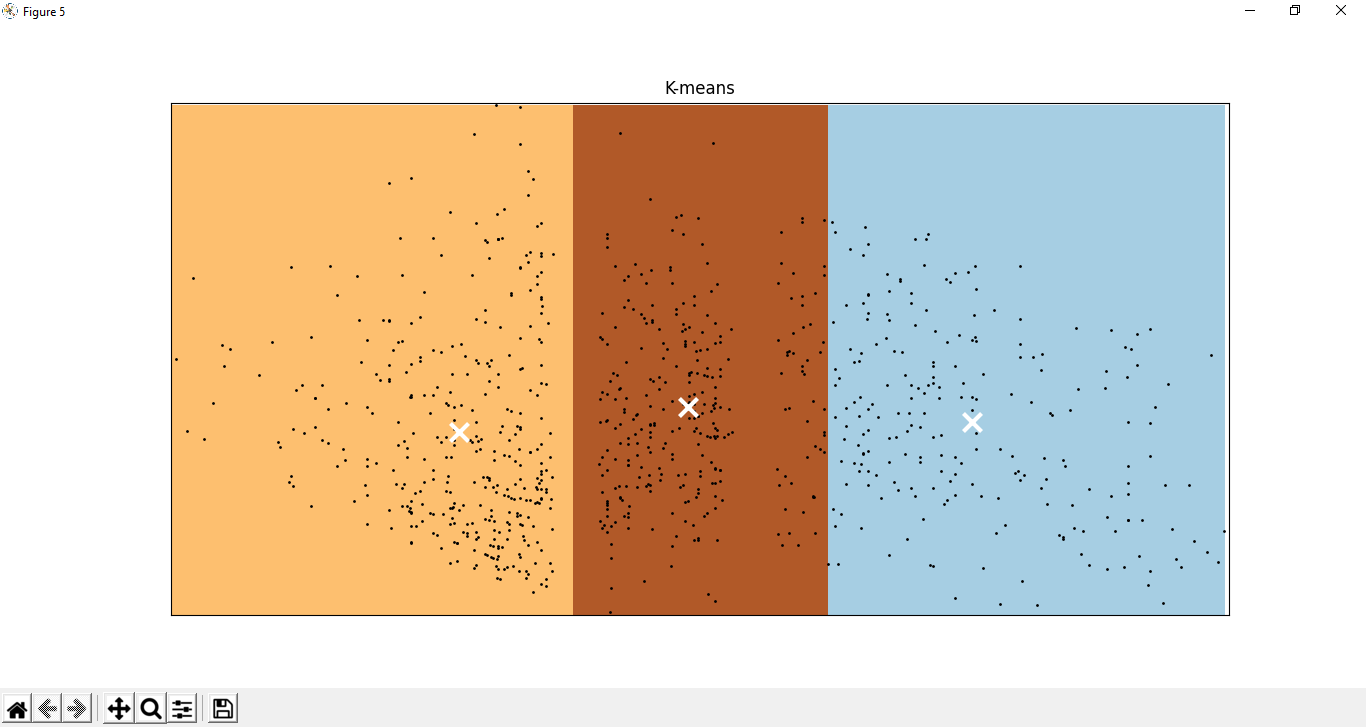


Рисунок 7

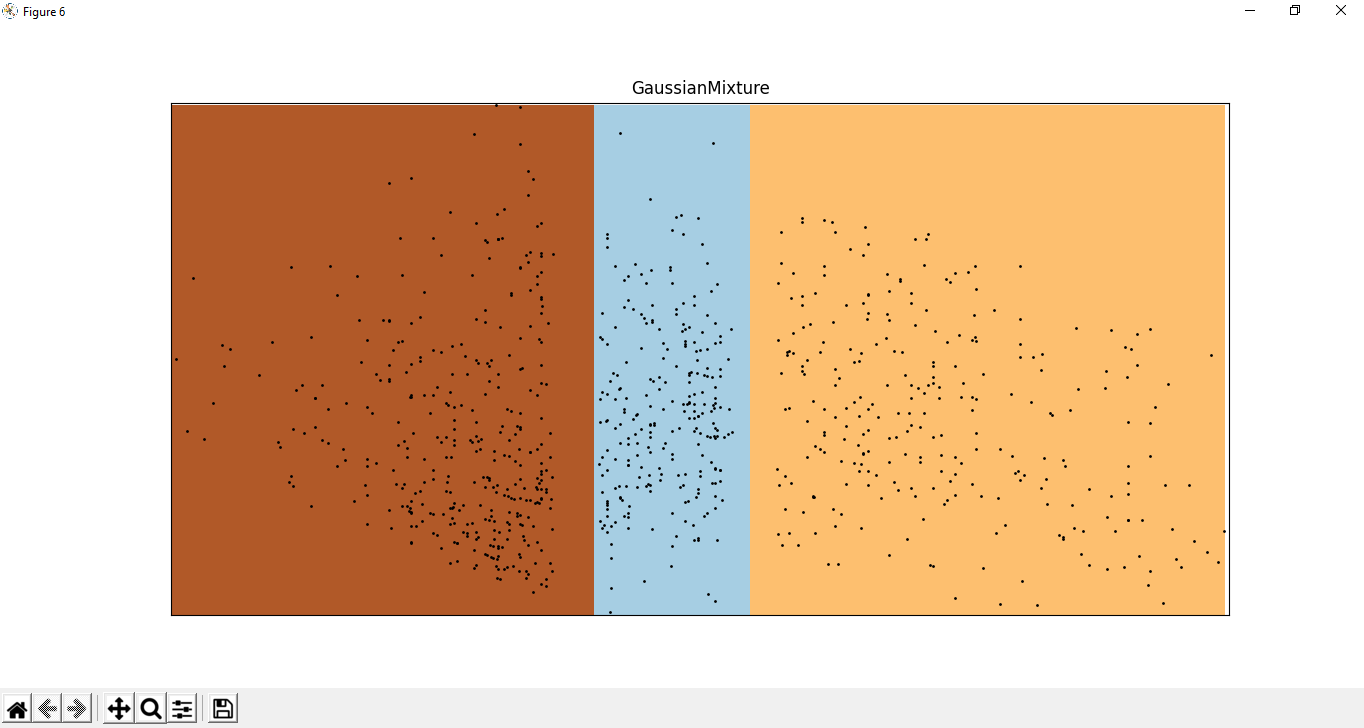


Рисунок 8

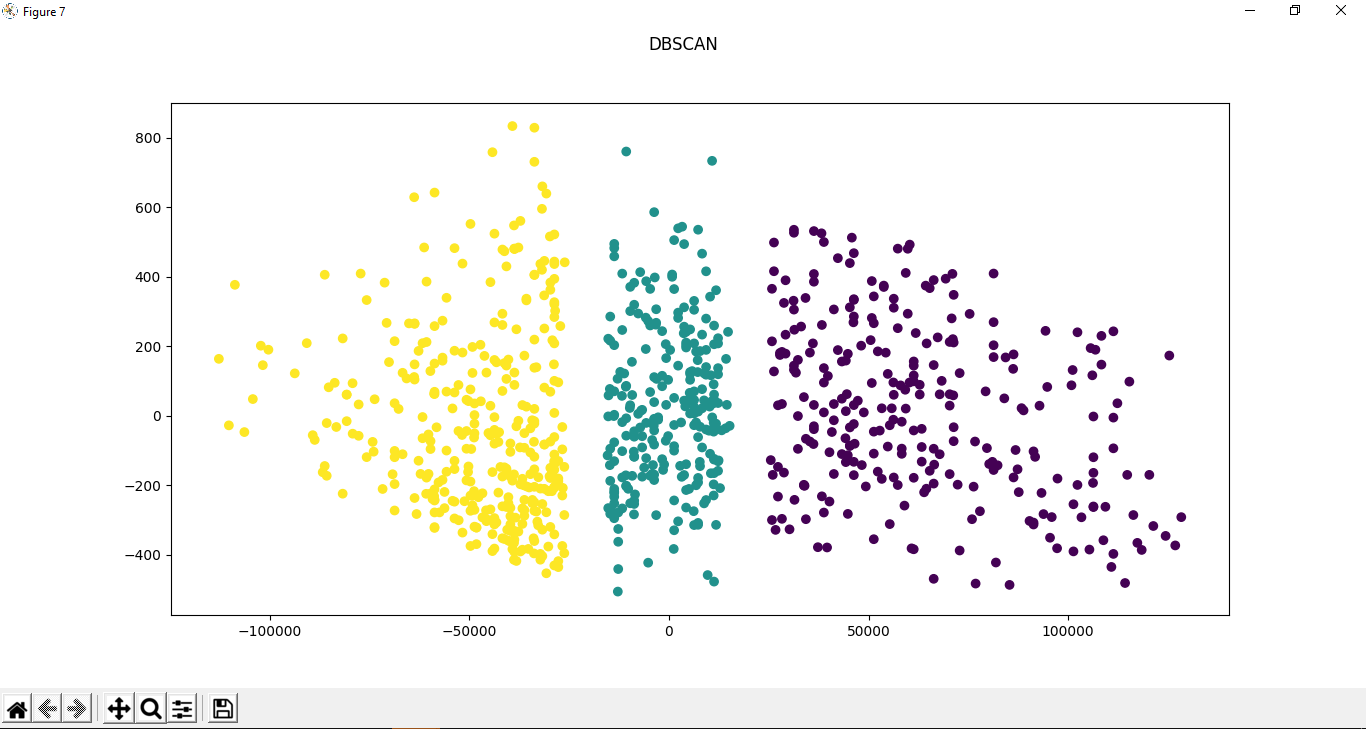


Рисунок 9

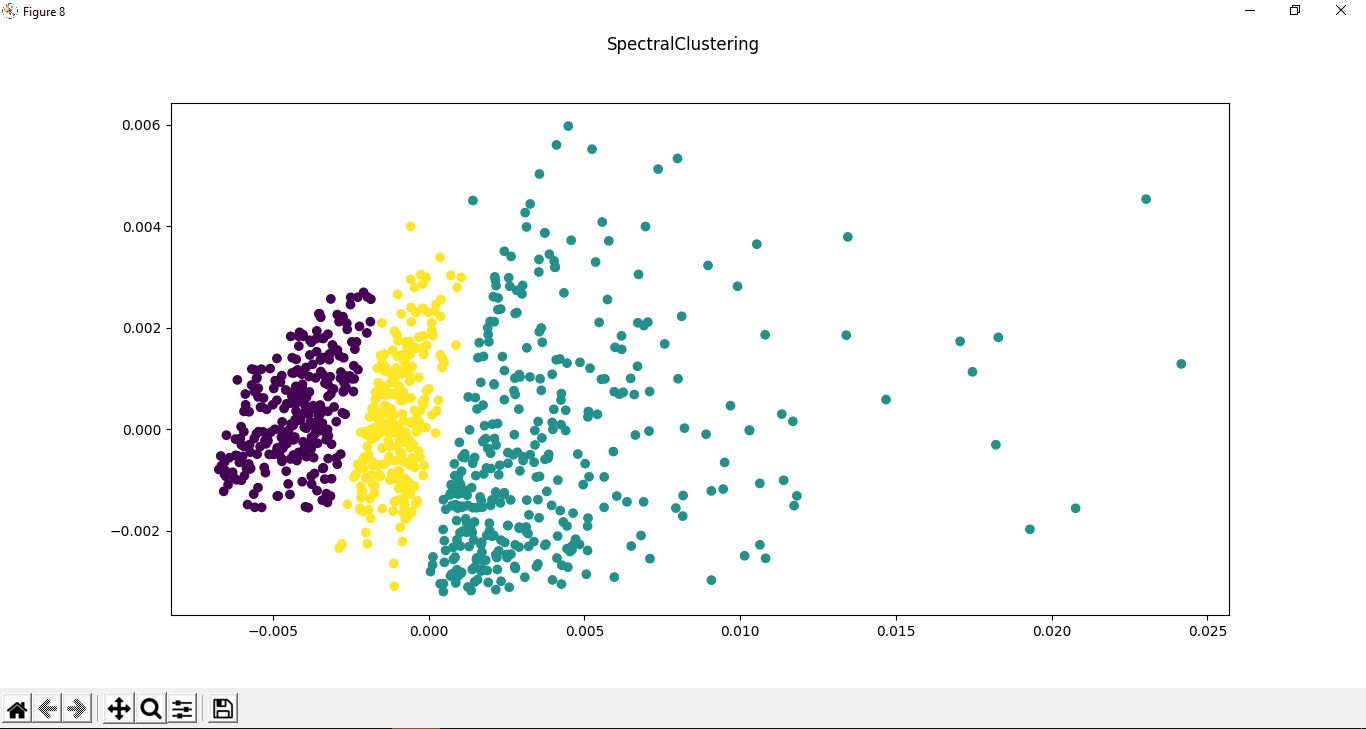


Рисунок 10

Значения трех метрик для всех примененных методов представлены ниже.

------- K-means --------

silhouette\_score 0.6055969142520004

davies\_bouldin\_score 0.4837671127228833

adjusted\_rand\_score 0.8565139232355925

------- GaussianMixture --------

silhouette\_score 0.5937128915124407

davies\_bouldin\_score 0.4712299632446013

adjusted\_rand\_score 1.0

------- DBSCAN --------

silhouette\_score 0.5937128915124407

davies\_bouldin\_score 0.4712299632446013

adjusted\_rand\_score 1.0

------- SpectralClustering --------

silhouette\_score 0.3287892738743041

davies\_bouldin\_score 0.8846889266185444

adjusted\_rand\_score 0.9965533711799966

В таблице 5 представлены время обучения и время предсказания (где возможно) для всех примененных ранее методов на разменных данных.

Таблица 4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *K-Means* | *Gaussian Mixture* | *DBSCAN* | Спектральная кластеризация |
| Обучение | 0.191 | 0.064 | 0.009 | 0.147 |
| Предсказание | 0.004 | 0.011 | - | - |

# Заключение

В настоящей работе использовалась та же база, что и в лабораторной работе 3. Вся предварительная обработка данных (удаление выбросов, выбор высоко коррелирующих признаков) осталась без изменений по сравнению с лабораторной работой 3. Единственное отличие состоит в том, что были выбраны немного другие признаки, коррелирующие с ценой, а также добавлен новый признак, отражающий стоимость квадратного фута жилой площади

Первым этапом работы была проведена кластеризация всех объектов четырьмя различными методами. Результаты работы методов были оценены двумя метриками, по результатам оценки которых были сделаны выводы о том, что датасет плохо подходит под задачу кластеризации вследствие того, что на диаграмме рассеяния все объекты сгруппированы вместе, в одном облаке. Особенно это повлияло на кластеризацию методом *DBSCAN*, результатом работы которого стало выделение одного большого кластера и нескольких небольших кластеров на краях облака объектов на диаграмме рассеяния.

Вторым этапом работы стало выделение трех групп жилья по стоимости и удаление ряда граничных объектов. Вследствие описанных действий на диаграмме рассеяния даже визуально были заметны границы между кластерами. В результате, методами *DBSCAN* и *Gaussian Mixture* была достигнута идеальная кластеризации, что отражается значением скорректированного метода Рэнда равным 1.

При анализе времени предсказания можно сделать вывод о том, что метод *DBSCAN* наименее время затратный метод, в то время как *K-Means* – наиболее время затратный. Однако, время предсказания методом *K-Means* в несколько раз меньше времени предсказания методом *Gaussian Mixture.*