ГУАП

КАФЕДРА № 41

ОТЧЕТ   
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| cтарший преподаватель |  |  |  | В. В. Боженко |
| должность, уч. степень, звание |  | подпись, дата |  | инициалы, фамилия |

|  |
| --- |
| ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ №4 |
| Кластеризация 2024 по курсу: ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ |
|  |
|  |

РАБОТУ ВЫПОЛНИЛ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СТУДЕНТ гр. № | 4217 |  |  |  | В. А. Милованов |
|  |  |  | подпись, дата |  | инициалы, фамилия |

Санкт-Петербург 2024

**Цель работы**

Изучение алгоритмов и методов кластеризации на практике.

**Индивидуальное задание**

Вариант 4

Набор данных 4heart2.csv

Данные о болезнях сердца:

1. возраст: возраст пациента (лет)

2. анемия: снижение количества эритроцитов или гемоглобина (логическое значение)

3. высокое кровяное давление: если у пациента гипертония (логическое значение)

4. креатининфосфокиназа (КФК) уровень фермента КФК в крови (мкг/л)

5. диабет: если у пациента диабет (логическое значение)

6. фракция выброса: процент крови, покидающей сердце при каждом сокращении (в процентах)

7. тромбоциты: тромбоциты в крови (килотромбоциты/ мл)

8. пол: женщина или мужчина (бинарный)

9. креатинин сыворотки: уровень креатинина сыворотки в крови (мг/дл)

10. натрий сыворотки: уровень натрия сыворотки в крови (мэкв/л)

11. курение: если пациент курит или нет (логическое)

12. время: период наблюдения (дни)

13. событие смерти: если пациент умер в течение периода наблюдения (логическое значение)

**Ход работы**

Датафреймы были сгенерированы с помощью функции make\_data, которая создала 3 группы с заданным числом объектов и двумя признаками. Эти данные обеспечивают основу для последующей кластеризации. Созданы тренировочный и тестовый наборы данных с разными параметрами seed для воспроизводимости (см. рис. 1).

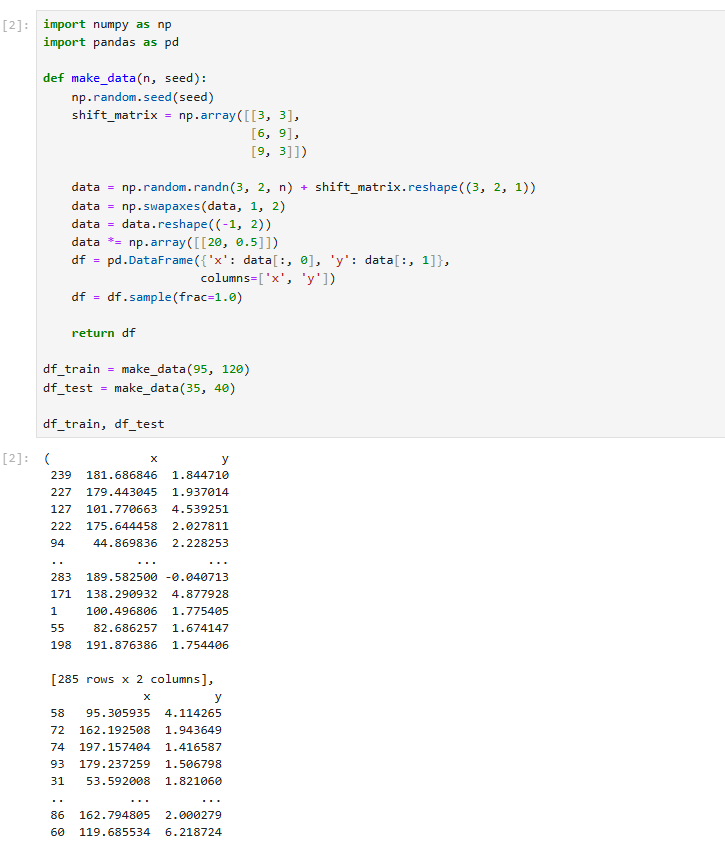
****

Рисунок 1 – Функция генерации данных

Данные были стандартизированы с использованием StandardScaler. Это привело к привидению признаков к одному масштабу, устранив влияние различий в их диапазонах. Стандартизация улучшает работу модели, особенно для чувствительных к масштабу алгоритмов, таких как K-means. Был импортирован соответствующий метод для обучения данных (см. рис. 2).

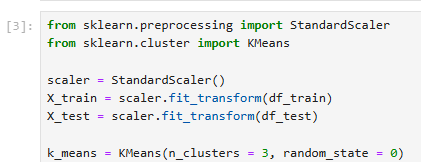


Рисунок 2 – Стандартизация

Используем метод, соединяющий в себе свойства fit и predict, для обучения модели (см. рис. 3).



Рисунок 3 – Обучение с помощью fit\_predict

Модель K-means успешно обучена на тренировочных данных. Были выделены 3 кластера, и каждому объекту присвоена соответствующая метка. Координаты центров кластеров были рассчитаны с помощью метода cluster\_centers\_. Это позволяет понять их положение в пространстве признаков. Было выведено визуальное отображение созданных кластеров, обозначенных разными цветами (см. рис. 4).

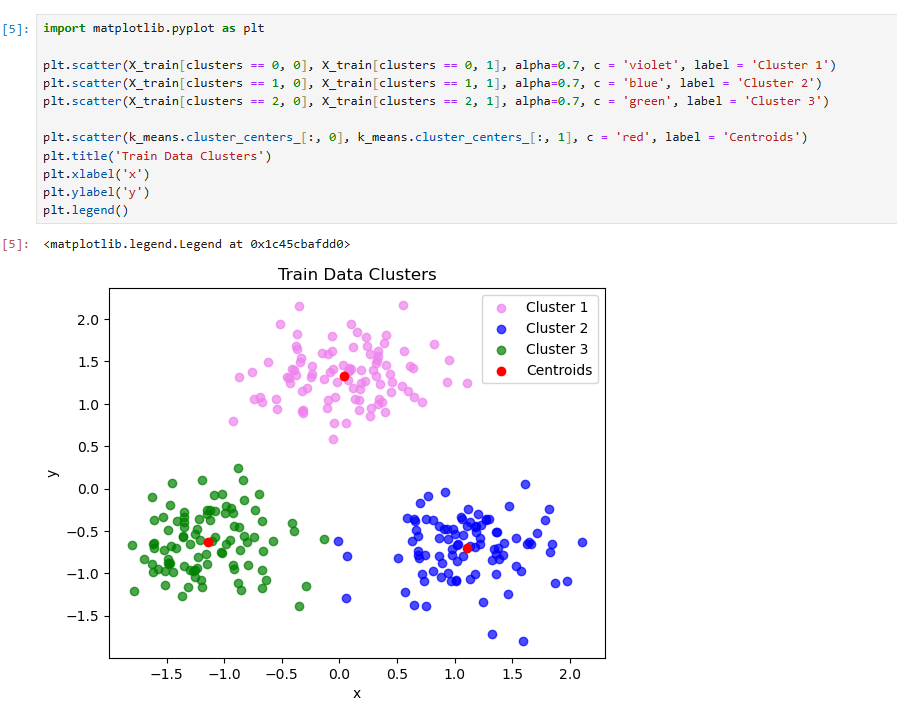
****

Рисунок 4 – Визуализация тренировочной выборки

Кластеры визуализированы с использованием метода scatter. Объекты из разных кластеров покрашены в разные цвета, что наглядно показывает границы между группами. Центры кластеров выделены, что помогает оценить их относительное положение.

Для тестовых данных успешно выполнено предсказание кластеров с использованием обученной модели K-means. Построен график, на котором объекты из разных кластеров окрашены по принадлежности. Центры кластеров также отображены, что подтверждает корректность прогнозирования на новых данных (см. рис. 5).

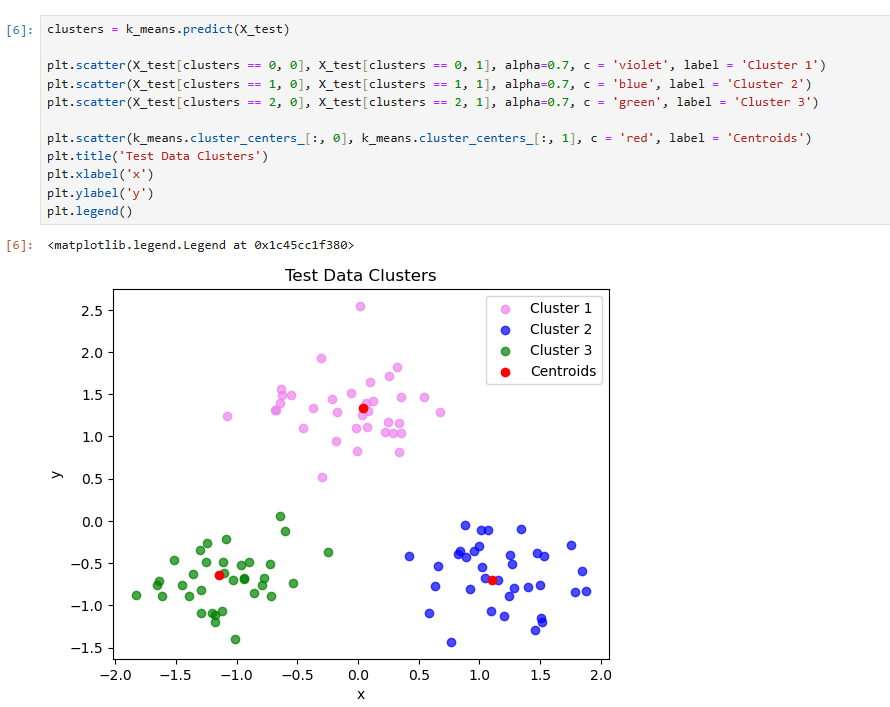


Рисунок 5 – Визуализация тестовой выборки

Был импортирован метод для подсчета значения коэффициента силуэта для данной модели (см. рис. 6).

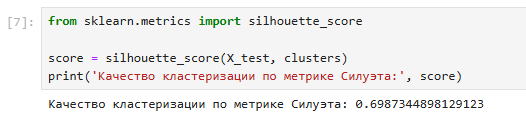


Рисунок 6 – коэффициент Силуэта

Коэффициент силуэта показал значение, близкое к 1, что указывает на хорошую плотность кластеров и чёткое разделение между ними. Это подтверждает высокое качество кластеризации при выбранном числе кластеров

Были проведены аналогичные расчеты для количества кластеров, равного 2 и 4 (см. рис. 7-8).

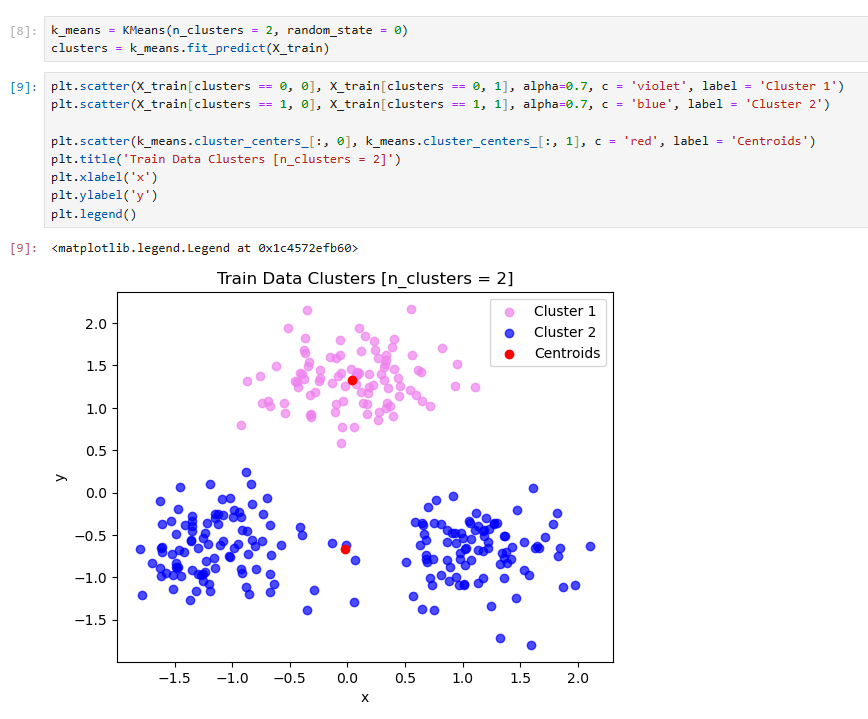


Рисунок 7 – Визуализация при n\_clusters = 2

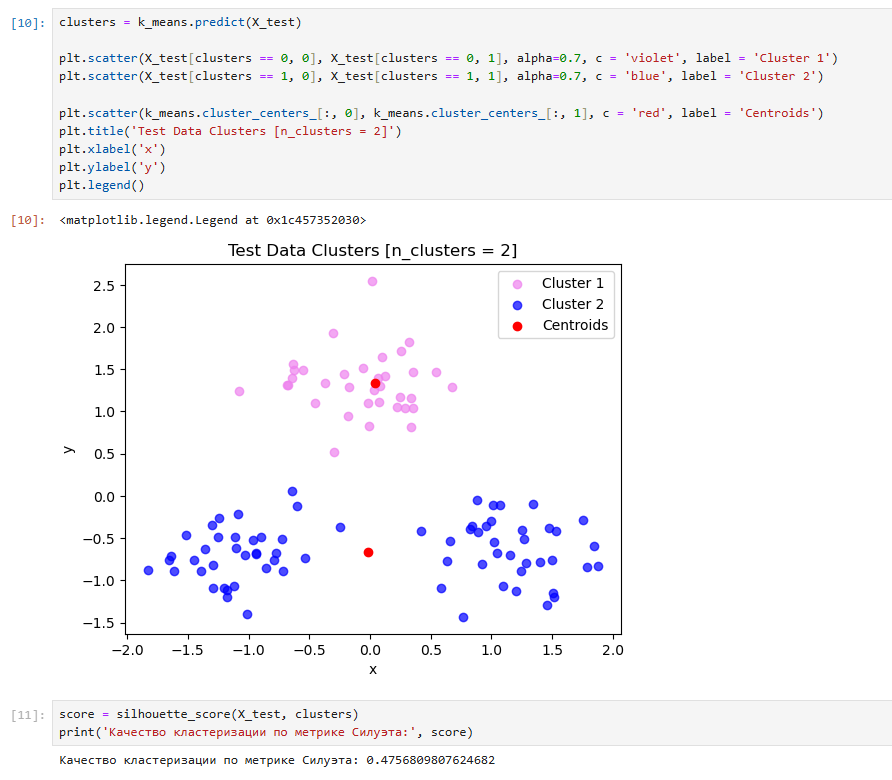


Рисунок 8 – Визуализация тестовой выборки при n\_clusters = 2

Можно заметить неоднородность кластеров по отношению друг к другу и падение показателей метрики. Ситуация аналогична на обоих выборках (см. рис. 9).

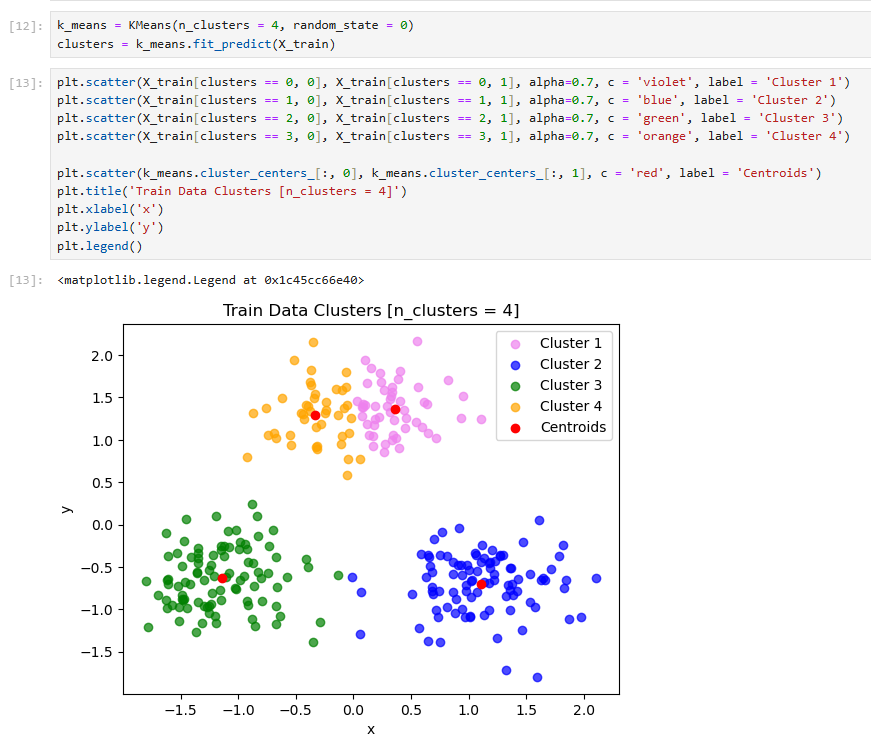


Рисунок 9 – Визуализация при n\_clusters = 4

Уже на нынешнем этапе снова заметна неоднородность кластеров. Номера 1 и 4 делят одну практически цельную группу (см. рис. 10).

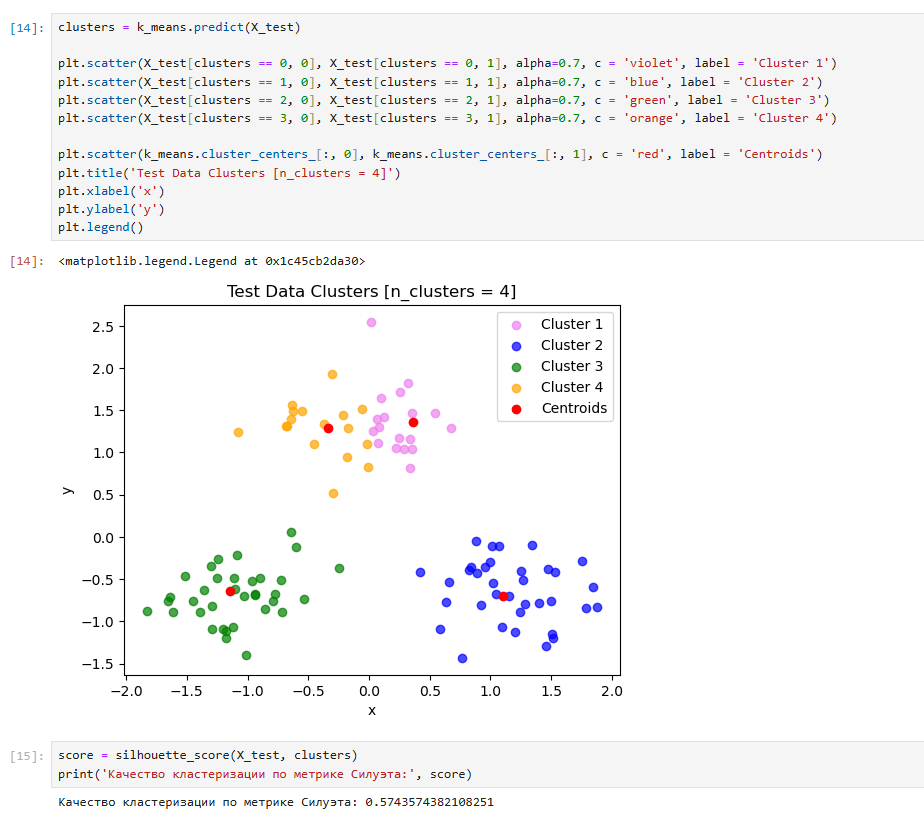


Рисунок 10 – Визуализация тестовой выборки при n\_clusters = 4

Параметр n\_clusters в алгоритме K-means указывает количество кластеров, на которые необходимо разделить данные. От выбора этого параметра зависит размер и структура модели. При n\_clusters=2 наблюдалось снижение коэффициента силуэта, что указывает на недостаточное разделение кластеров. При n\_clusters=4 кластеры были разделены слишком подробно, что привело к снижению коэффициента силуэта. Оптимальным оказалось число кластеров n\_clusters=3.

Наблюдения были проверены с использованием графика, построенному по "методу локтя". Метод локтя — это способ определения оптимального числа кластеров в алгоритмах кластеризации. Этот метод основан на анализе метрики инерции, которая измеряет степень рассеяния точек внутри кластеров (см. рис. 11).

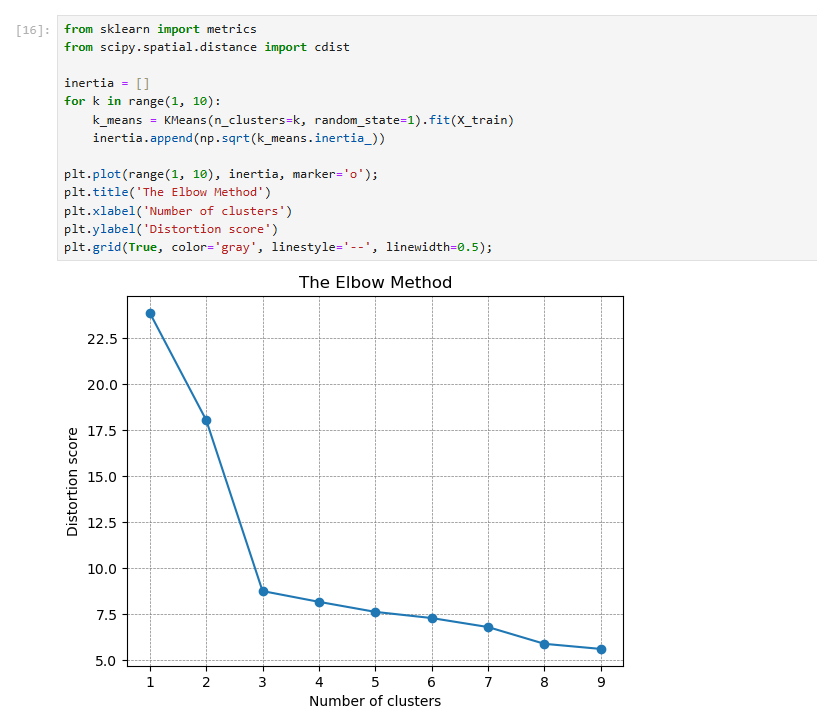


Рисунок 11 – Метод локтя

График метода локтя продемонстрировал "излом" при числе кластеров, равном 3. Это подтверждает, что три кластера являются оптимальным числом для данной задачи. При дальнейшем увеличении числа кластеров улучшение инерции становится незначительным.

Была импортирована база данных согласно варианту (см. рис. 12).

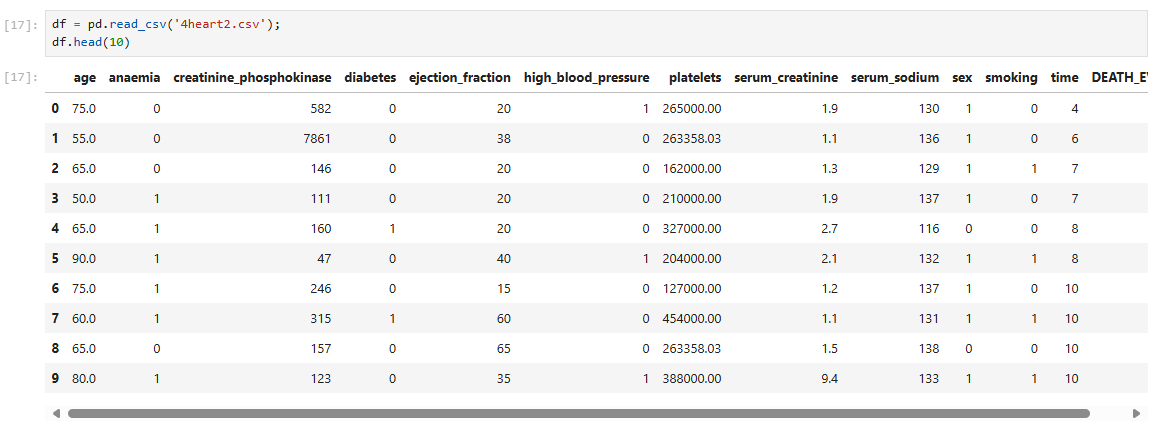


Рисунок 12 – База данных

Проводится первичная обработка данных. Проверяются столбцы, типы и правильность информации. Обработка данных необходима, чтобы избежать ошибок, связанных с некорректными значениями или несоответствием типов данных (см. рис. 13).

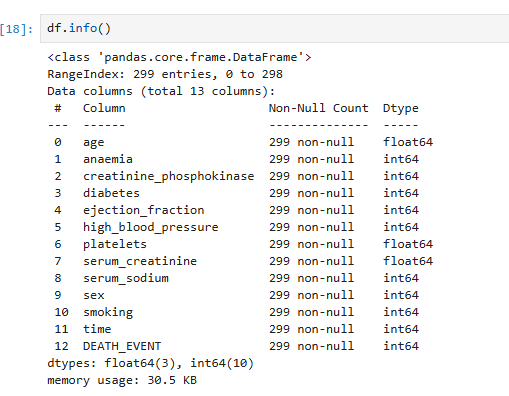


Рисунок 13 – Информация о БД

Так как возраст использовается в виде целых значений, было решено поменять его тип данных с float на int (см. рис. 14-15).

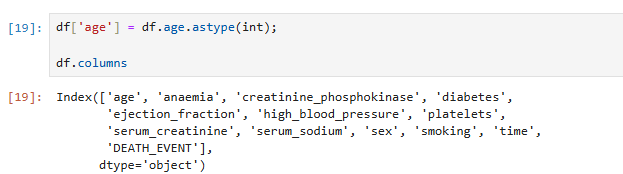


Рисунок 14 – Изменение типа данных

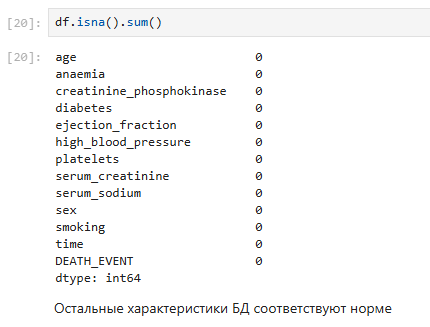


Рисунок 15 – Проверка данных

Остальные характеристики БД соответствуют норме.

С помощью pairplot из библиотеки Seaborn построена матрица диаграмм рассеяния, где данные каждого класса выбранной переменной выделены разными цветами (см. рис. 16).

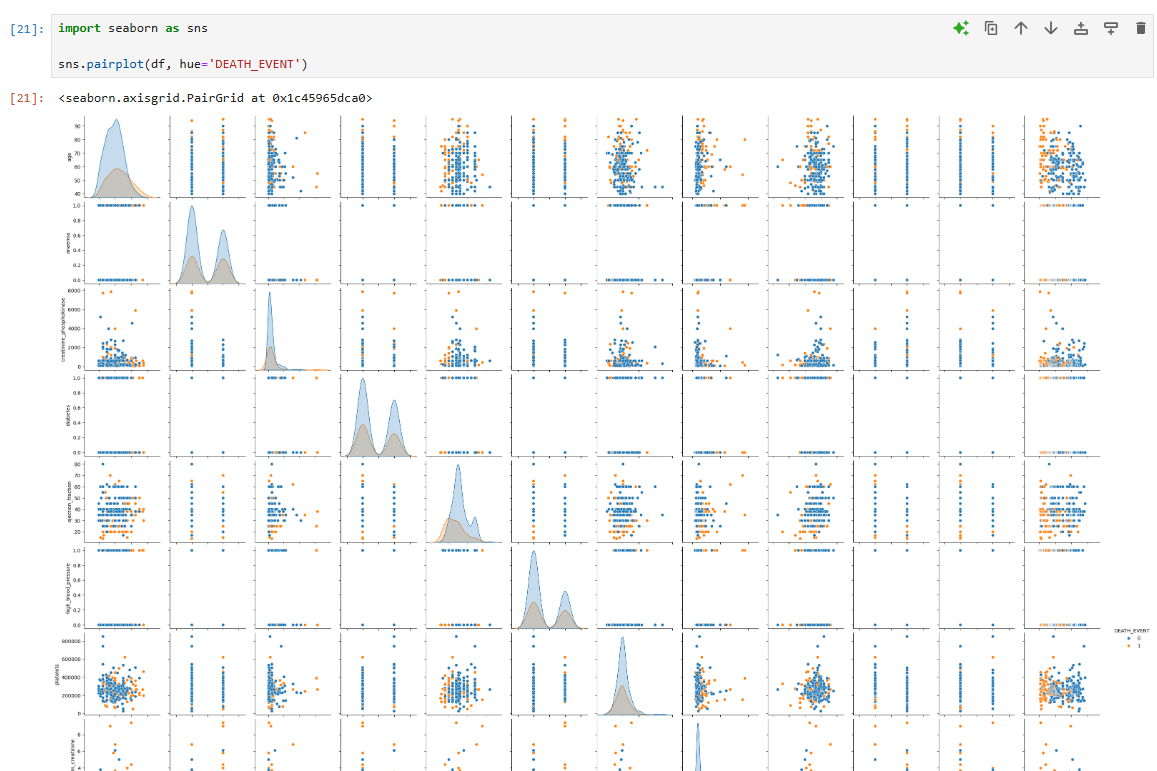


Рисунок 16 – Матрица рассеяния

Данные были масштабированы с использованием StandardScaler, который приводит каждый признак к распределению со средним значением 0 и стандартным отклонением 1. Стандартизация данных обязательна для методов, чувствительных к масштабам признаков. Если признаки находятся в разных единицах измерения (например, метры и километры), это может привести к неверным результатам кластеризации (см. рис. 17).

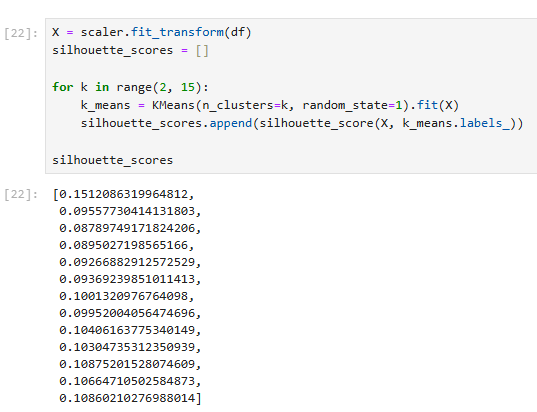


Рисунок 17 – Поиск оптимального n\_clusters

С помощью внутренней метрики подбиралось наиболее эффективное количество кластеров для загруженной БД. В результате самый весомый результат показал n\_clusters=2. Такой итог может быть обусловлен большим набором признаков, содержащих в себе бинарные значения (см. рис. 18).

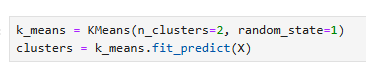


Рисунок 18 – Обучение добавленной базы данных

После обучения модели были выведены средние значения данных каждого кластера по каждому признаку. Анализ средних значений позволяет интерпретировать кластеры, выявляя ключевые признаки, которые их различают. Например, один кластер может представлять объекты с высокими значениями определенных признаков (см. рис. 19).

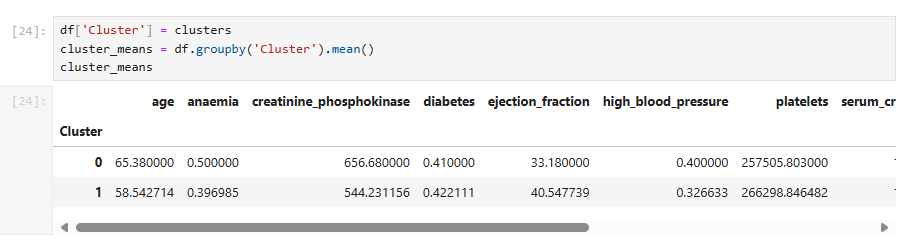


Рисунок 19 – Вывод средних значений

В результате большая часть признаков содержится во втором кластере.

Были вычислены метрики: Calinski-Harabasz Score: показывает соотношение межкластерной дисперсии к внутрикластерной. Davies-Bouldin Score: оценивает качество кластеров, где меньшие значения означают лучшее разделение (см. рис. 20).

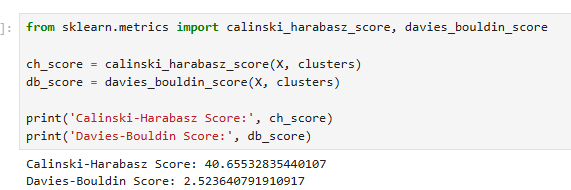


Рисунок 20 – Вывод метрик

Данные внутренние метрики показывают успешный результат обучения.

Выполнена агломеративная кластеризация с построением дендрограммы. Определено оптимальное количество кластеров на основе анализа дендрограммы. Иерархическая кластеризация объединяет объекты в группы, начиная с самых мелких (один объект — один кластер) и заканчивая одним кластером для всех объектов. Дендрограмма показывает процесс объединения кластеров (см. рис. 21-22).

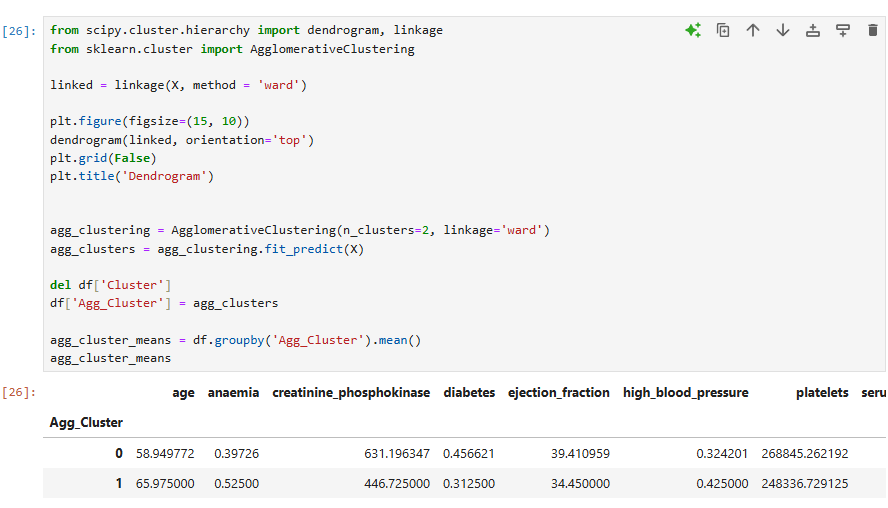


Рисунок 21 – Обучение с использованием агломеративной кластеризации

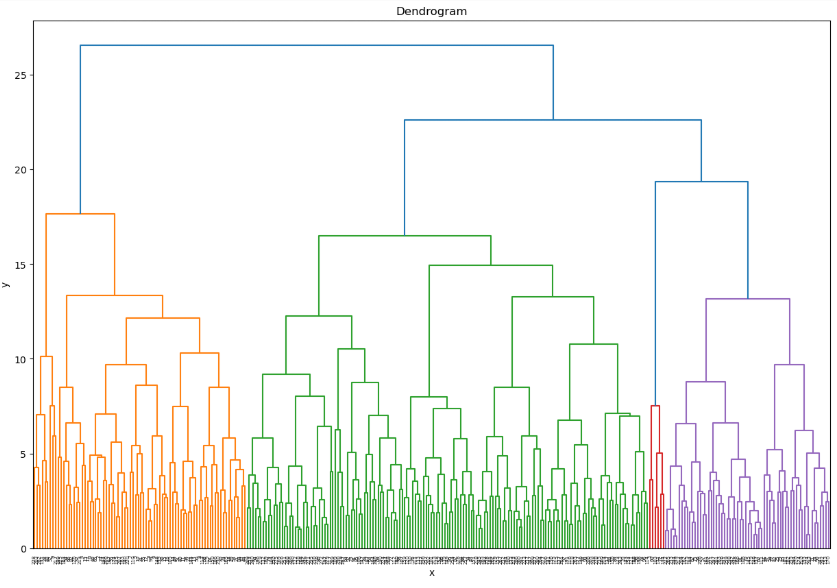


Рисунок 22 – Дендрограмма

Методы K-means и иерархической кластеризации показали сходные результаты, но каждый имеет свои преимущества: K-means лучше для больших данных и быстрее работает. Иерархическая кластеризация дает более наглядное представление о структуре данных через дендрограмму (см. рис. 23).

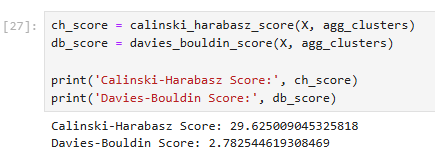


Рисунок 23 – Вывод метрик по второму методу обучения

Ссылка на файл:

<https://github.com/Vvvvv55531/LR4.git>

**Вывод**

В ходе работы была выполнена задача кластеризации данных, которая включала предварительную обработку, построение модели, визуализацию результатов и анализ качества. Кластеризация относится к методам обучения без учителя, позволяя разделить данные на группы на основе их сходства без предварительных меток. В данной работе был использован алгоритм K-means, который делит данные на k кластеров, минимизируя внутрикластерную дисперсию. Для определения оптимального числа кластеров были применены метод локтя и коэффициент силуэта. Метод локтя показал "излом" на графике при n\_clusters=3, а максимальное значение коэффициента силуэта также подтвердило, что три кластера обеспечивают лучшее качество группировки. Стандартизация данных с помощью StandardScaler позволила устранить влияние различных масштабов признаков, что улучшило результаты кластеризации.

Для анализа качества кластеризации использовались метрики инерции, коэффициент силуэта и Calinski-Harabasz Score. Инерция уменьшалась с увеличением числа кластеров, но после трех кластеров снижение становилось незначительным. Коэффициент силуэта достиг максимума при n\_clusters=3, указывая на высокую плотность кластеров и хорошее их разделение. Calinski-Harabasz Score также подтвердил высокое качество модели при трёх кластерах. На тренировочных данных кластеры визуализированы с использованием scatter-графиков, где объекты из разных кластеров выделены цветами, а центры кластеров обозначены маркерами. Для тестовых данных была выполнена кластеризация с использованием обученной модели, и результаты показали, что модель сохраняет стабильность при работе с новыми данными.

Дополнительно были протестированы значения n\_clusters=2 и n\_clusters=4. При двух кластерах разделение данных оказалось слишком грубым, объединяя разные группы в одну, а при четырех кластерах наблюдалось избыточное разделение, что снизило коэффициент силуэта. Это подтверждает, что выбор оптимального числа кластеров играет ключевую роль в обеспечении качества модели. Построение графика метода локтя и использование коэффициента силуэта доказали, что три кластера обеспечивают лучший баланс между плотностью групп и их разделением.

В результате работы было установлено, что алгоритм K-means является эффективным инструментом для кластеризации данных, особенно при правильной предварительной обработке и выборе параметров. Стандартизация данных и использование метрик качества позволили достичь устойчивых результатов. Для анализа данных, требующих детальной иерархической структуры, могут применяться альтернативные методы, такие как агломеративная кластеризация, но для больших объемов данных и задач с фиксированным числом кластеров K-means является оптимальным выбором.