ГУАП

КАФЕДРА № 41

ОТЧЕТ   
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

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

|  |
| --- |
| ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ № 4. |
| КЛАСТЕРИЗАЦИЯ. |
| по курсу: ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ. |
|  |
|  |

РАБОТУ ВЫПОЛНИЛ

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| СТУДЕНТ ГР. № | 4217 |  |  |  | | Ш.С. Махкамов |
|  |  |  | подпись, дата | |  | инициалы, фамилия |

**Цель работы**

Изучение алгоритмов и методов кластеризации на практике.

**Индивидуальное задание**

Вариант №2

Ссылка на GoogleColab:

<https://colab.research.google.com/drive/1_zdHeAMZy2lV9Jr3Ck9iwfZ9uGpfp_UZ?usp=sharing>

**Выполнение работы**

**Часть 1 - Кластеризация набора данных**

Используется функция для создания набора данных с параметрами: n - число объектов в каждой группе, `seed` - зерно, необходимое для воспроизводимости случайных значений.

В результате применения получается `DataFrame`, содержащий заданное число объектов в каждой группе (всего 3 группы) с двумя признаками: x и y.

С помощью этой функции создаются два набора данных: `train` и `test`. Параметр n = 99, `seed` для `train` равен 10, для `test` равен 52.



Рисунок 1 – Создание набора данных

Далее проводится стандартизация данных. Для проверки выводятся первые 5 элементов каждого стандартизованного массива (рис. 2).

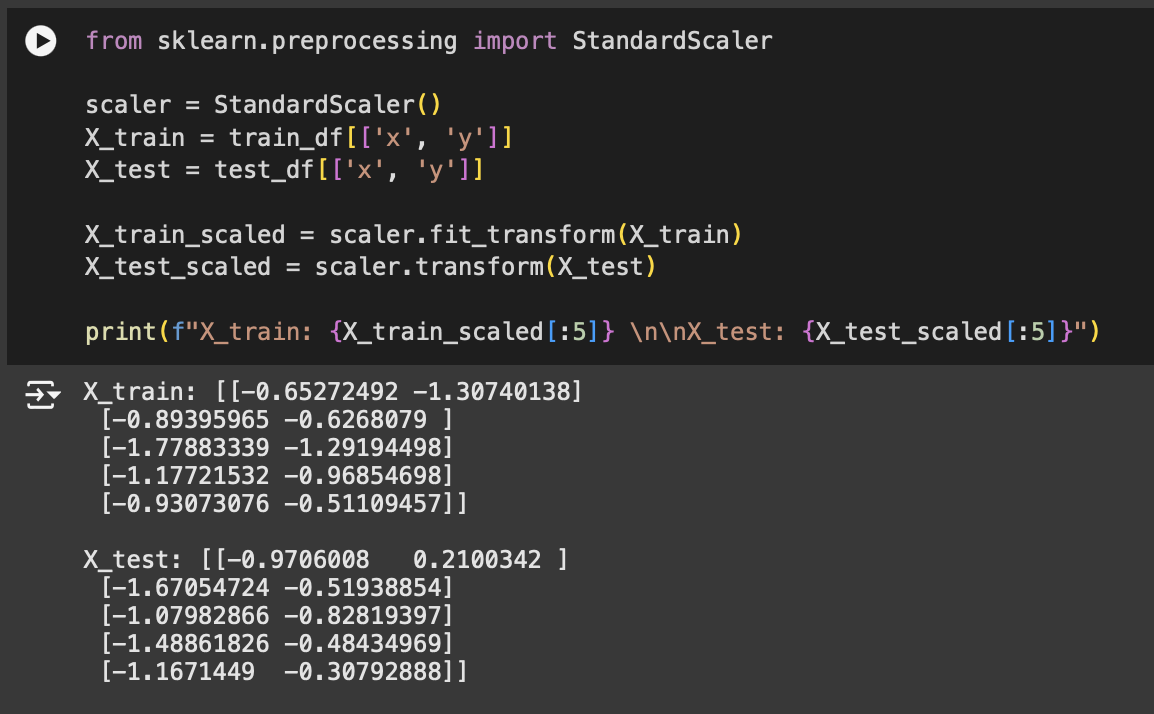


Рисунок 2 – Стандартизация

С помощью метода `fit\_predict` модель обучается для кластеризации методом `K-means` (К-средних) для трёх кластеров. Создаются метки кластеров (рис. 3).

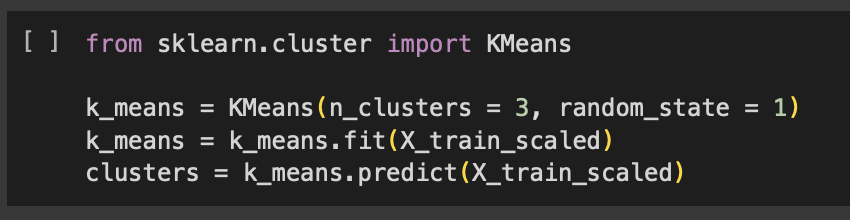


Рисунок 3 – Обучение модели

Проводится визуализация кластеров с помощью `scatter` (рис. 3). На графике помечаются центры кластеров. Объекты из разных кластеров помечаются разными цветами.



Рисунок 4 – Визуализация результатов кластеризации

На представленном графике отображаются результаты кластеризации методом K-средних. Визуализация разделяет данные на три различные группы, каждая из которых обозначена своим цветом.

По графику можно сделать вывод, что метод К-средних довольно эффективно и без ошибок определяет класс, к которому относится наблюдение.

Далее, с использованием тестовых данных `test` выполняется предсказание. Строятся графики для тестовых данных (рис. 5). На графике центры кластеров помечаются треугольниками. Объекты из разных кластеров красятся разными цветами.

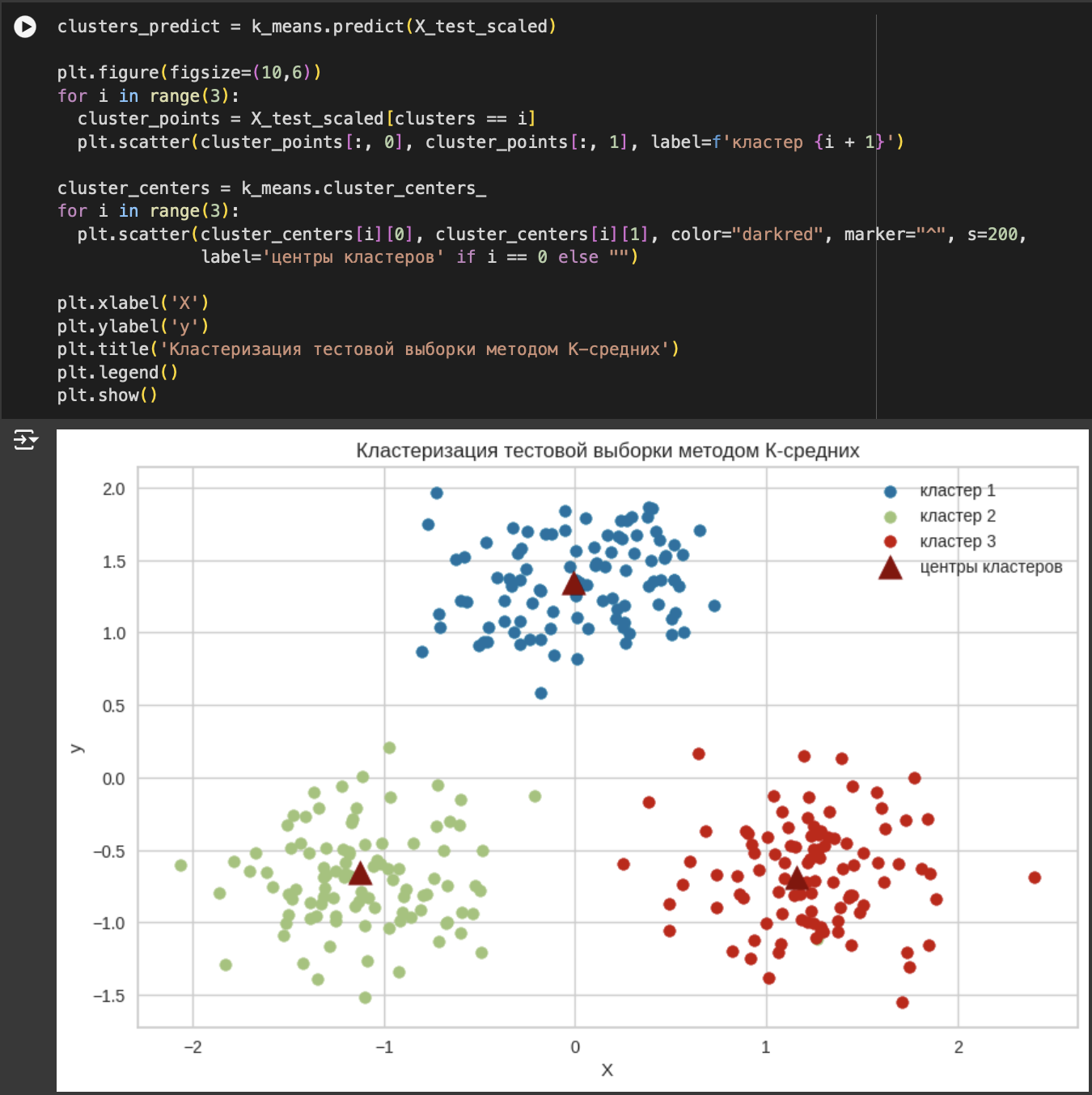


Рисунок 5 – Визуализации кластеризации тестовой выборки

Модель хорошо показывает себя на тестовой выборке. По графику видно, что определение классов для наблюдений прошло успешно.

Для достоверности подсчитывается коэффициент силуэта (рис. 6).

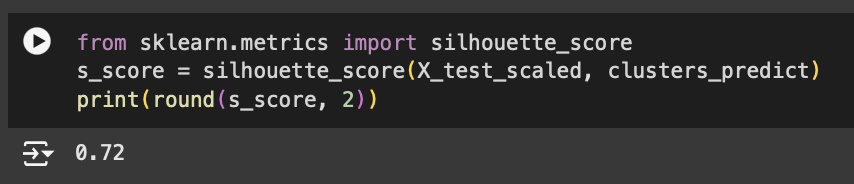


Рисунок 6 – Подсчет коэффициента силуэта

Коэффициент силуэта показывает, насколько хорошо была выполнена кластеризация, данный показатель находится в диапазоне от -1 до +1, при этом оценка около 0 указывает на перекрывающиеся кластеры.

Значение в 0,72 указывает на то, что обученная модель с высокой точностью определяет правильные классы.

Далее те же самые действия проделываются для n\_clusters = 2 (рис. 7) и n\_clusters = 4 (рис. 8). Сравниваются результаты и делаются выводы



Рисунок 7 – Модель К-средних для 2 кластеров

По графикам кластеризации на 2 группы можно сделать вывод, что в целом кластеризация выполняется верно - модель правильно определяет принадлежность к группам по К-средним, но по графику отчетливо видно, что кластеров в данных три.



Рисунок 8 – Модель К-средних для 4 кластеров

По графикам кластеризации на 4 группы можно сделать вывод, что 1 и 3 группу модель определила верно и в целом, 2 и 3 группу тоже разделила логически верно, но на данном графике, также как и на прошлом, видно, что кластеров в данных три.

Далее строится график по “методу локтяˮ с помощью визуалайзера.

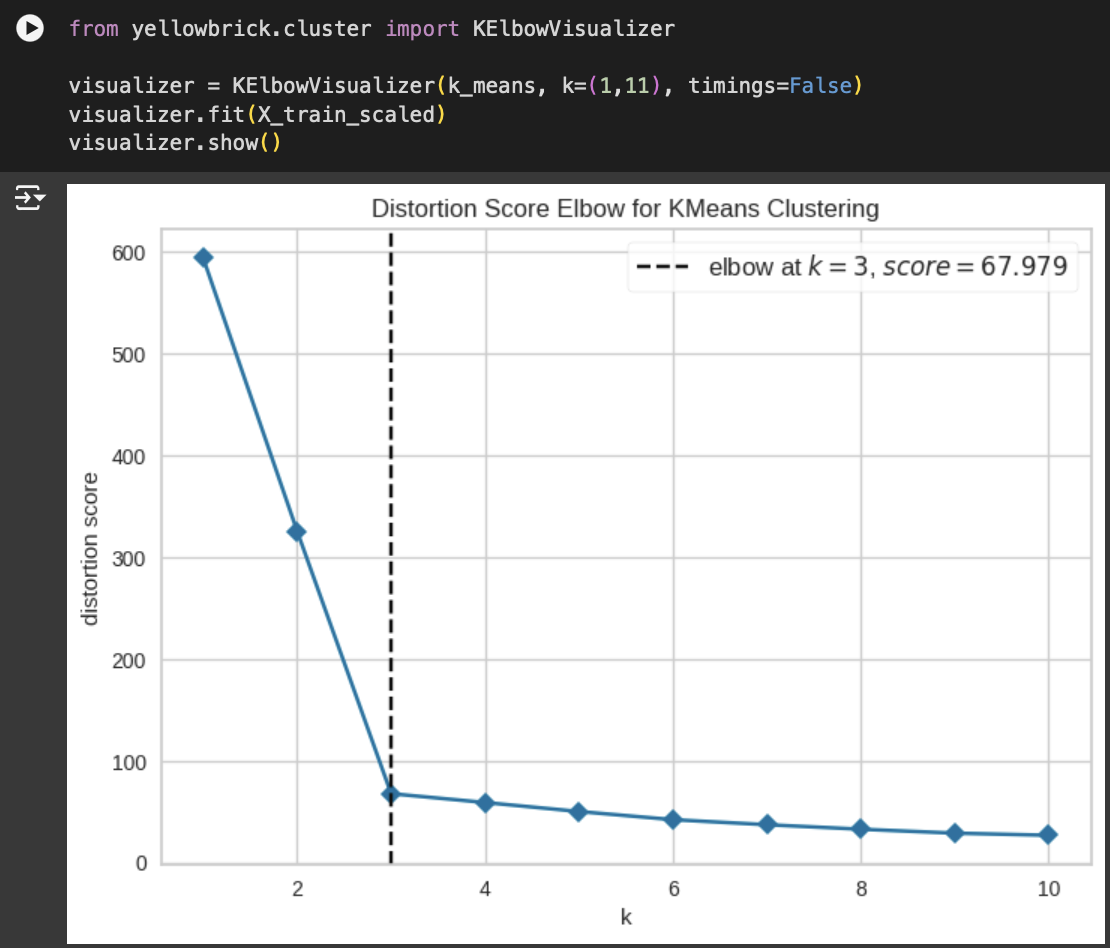


Рисунок 9 – Использование «метода локтя»

По графику четко видно и выделено, что количество кластеров равно 3, что подтверждается первой построенной диаграммой рассеяния.

**Часть 2 - Кластеризация по варианту**

Производится загрузка CSV файла и вывод первых 5 строк (рис. 10).

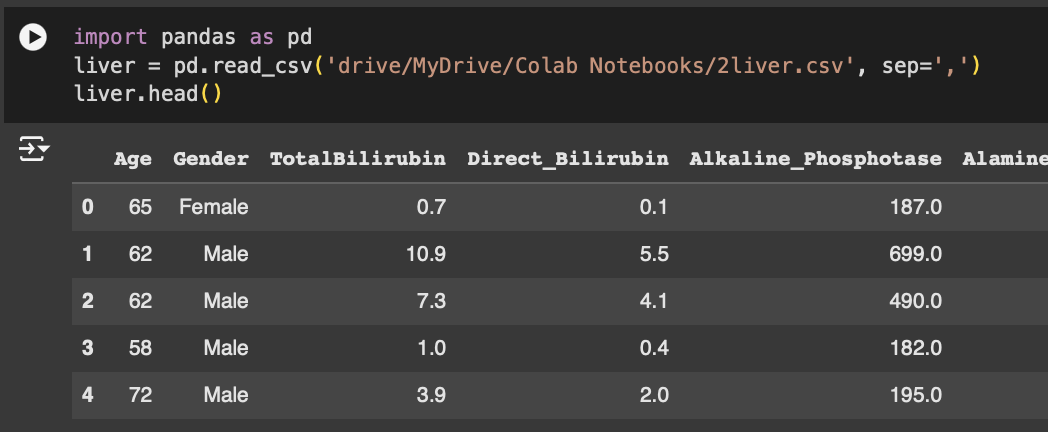


Рисунок 10 – Выгрузка файла с данными

Набор данных `2liver.csv` содержит информацию о заболеваемости печени:

1. Возраст пациента
2. Пол пациента
3. Общий билирубин
4. Прямой билирубин
5. Щелочная фосфатаза
6. Аламиноаминотрансфераза
7. Аспартатаминотрансфераза
8. Всего белков
9. Альбумин
10. Соотношение альбуминов и глобулинов
11. Целевое поле: два значения - пациент с заболеванием печени или без заболевания.

Проводится предварительная обработка данных. Для начала выводится общая информация о датасете (рис. 11).

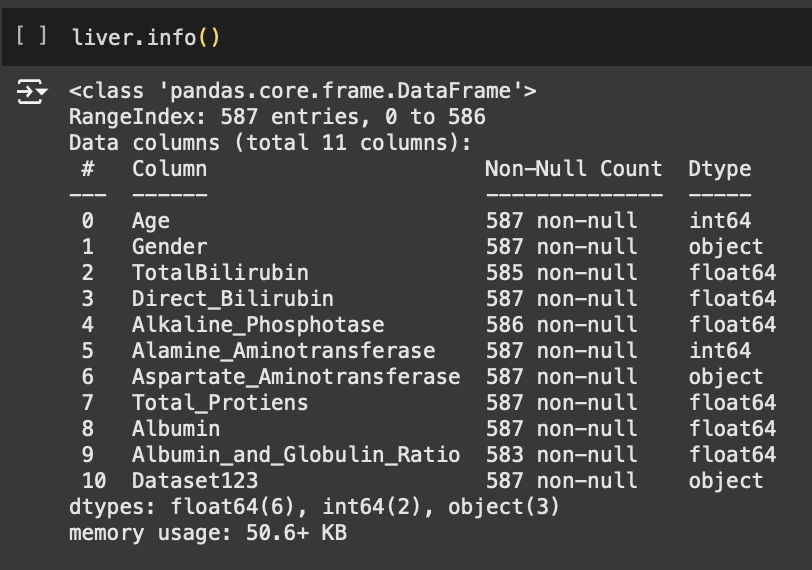


Рисунок 11 – Информация о датасете

Значения столбца `Aspartate\_Aminotransferase` приводятся к типу `float64` (рис. 12).

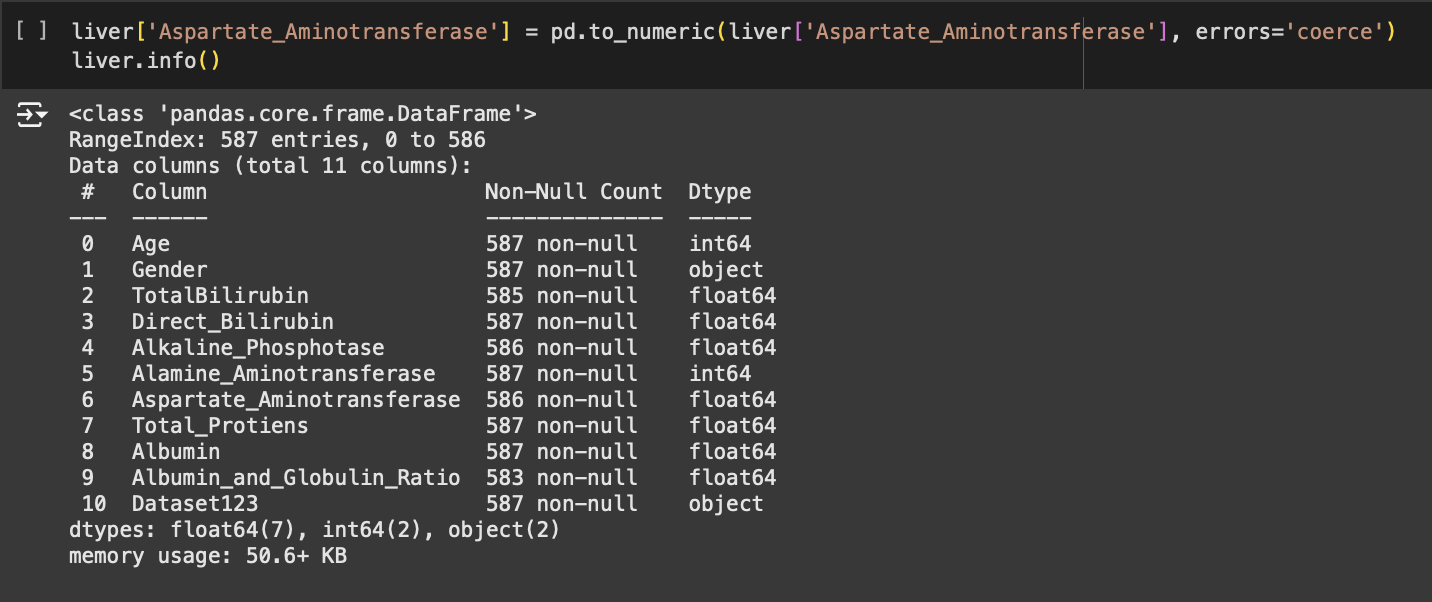


Рисунок 12 – Перевод строкового столбца в числовой

Датасет проверяется на наличие пропущенных значений (рис. 13).



Рисунок 13 – Поиск пропущенных значений

Чтобы модель правильно классифицировала наблюдения, вместо заполнения пустых полей они просто удаляются, так как их очень мало (рис. 14).

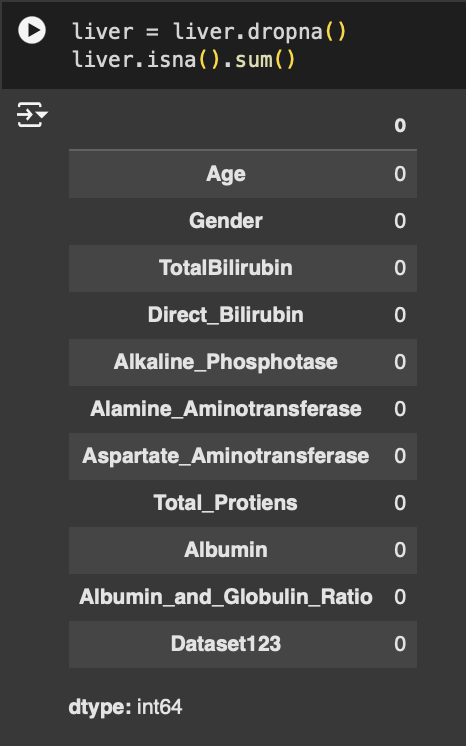


Рисунок 14 – Удаление пропущенных значений

Проверяется корректность заголовков столбцов (рис. 15).

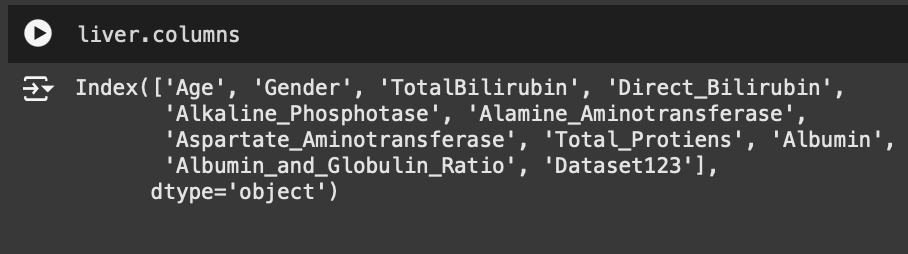


Рисунок 15 – Вывод названий столбцов

Столбец `Dataset123` переименовывается в `Disease` (рис. 16).

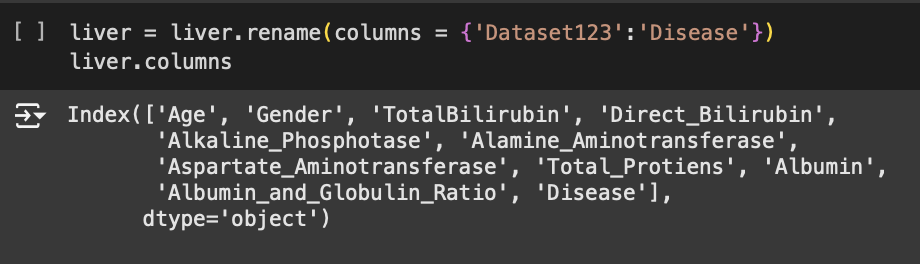


Рисунок 16 – Переименование заголовка столбца

Далее производится проверка на явные дубликаты (рис. 17).

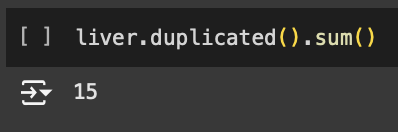


Рисунок 17 – Проверка на явные дубликаты

Дубликаты удаляются, после чего обновляется нумерация индексов (рис. 18).

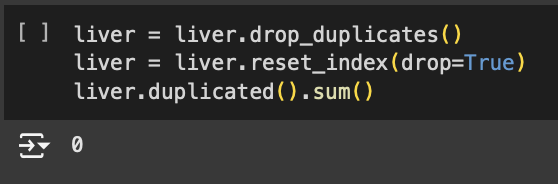


Рисунок 18 – Удаление явных дубликатов

Далее производится проверка на наличие неявных дубликатов (рис. 19).

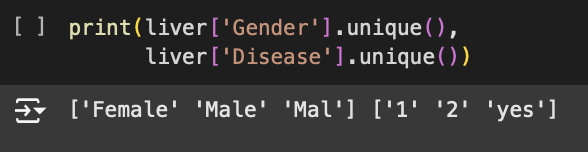


Рисунок 19 – Проверка на неявные дубликаты

Перед устранением неявных дубликатов необходимо определить какое из значений "1" и "2" означает, что человек болен, а какое - не болен. Для этого подсчитывается кол-во наблюдений с обоими признаками (рис. 20).

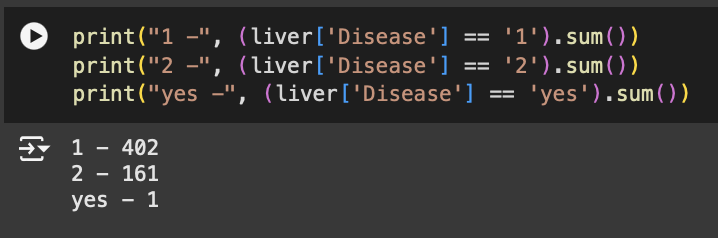


Рисунок 20 – Подсчет кол-ва наблюдений целевой переменной

Так как по статистике в наблюдениях реже встречается статус "болен" можно сделать вывод, что "1" означает, что человек не болен, а "2", что болен. Значение "yes" относится к значению "2".

Дубликаты заменяются на корректные значения, а значения «1» и «2» переименовываются в «yes» и «no» (рис. 21).

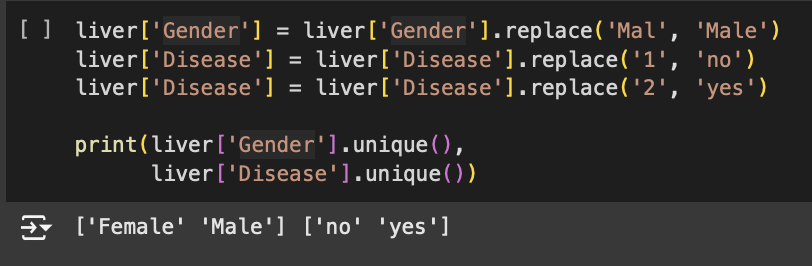


Рисунок 21 – Удаление неявных дубликатов

Для кластерного анализа данных необходимо привести строковые данные к числовым. Для этого используется `OrdinalEncoder` (рис. 22).



Рисунок 22 – Перевод категориальных признаков в числовые

В столбце `Gender` "0" означает "Female", а "1" - "Male". В столбце `Disease` "0" означает "no", а "1" - "yes".

Целевой переменной, которую необходимо предсказать, является столбец `Disease` (болен человек или нет). Целевая переменная не включается в модель.

Строится матрица диаграмм рассеяния (рис. 23). Значения целевой переменной выделяются разными цветами.

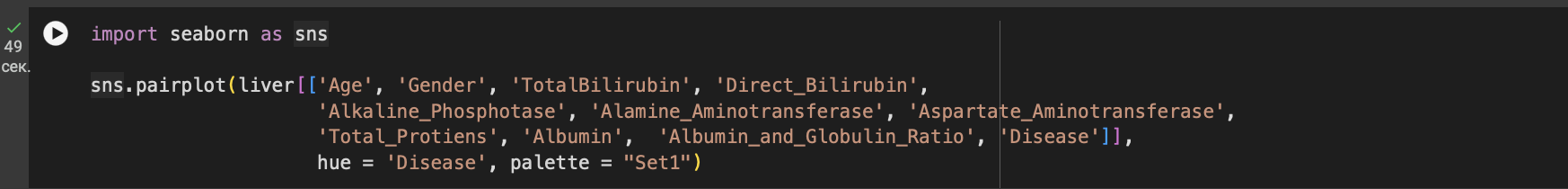




Рисунок 23 – Матрица диаграмм рассеяния

На представленной матрице зависимостей (матрица рассеяния) можно сделать следующие выводы:

* болезнь печени встречается у людей с низким показателем "TotalBilirubin", "Direct\_Bilirubin", "Alamine Aminotransferase" и "Aspartate Aminotransferase";
* болезнь печени встречается во всех возрастах примерно одинаково, поэтому данный признак в обучение модели не включается;
* наличие болезни печени не зависит от пола, поэтому данный признак тоже не включается в обучение модели;
* наблюдается сильная положительная корреляция между "TotalBilirubin" и "DirectBilirubin";
* наблюдается умеренная положительная корреляция между "Total\_Protiens" и "Albumin";
* наблюдается умеренная положительная корреляция между "Albumin" и "Albumin\_and\_Clobuline\_Ratio".

Строится тепловая карта корреляции числовых признаков для определения коллинеарных признаков (рис. 24).

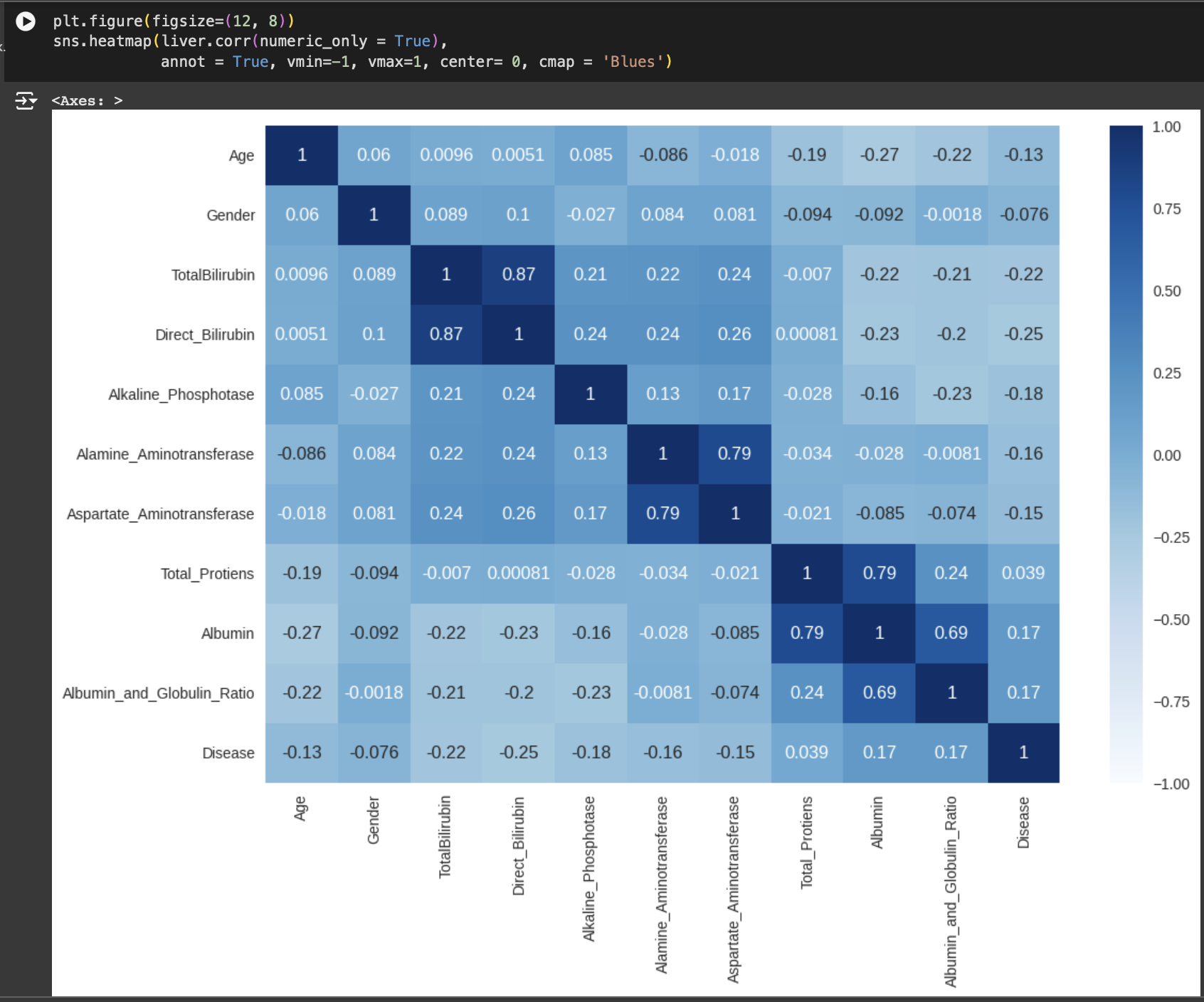


Рисунок 24 – Тепловая карта корреляции

По тепловой карте видно:

* 'TotalBilirubin' и 'Direct\_Bilirubin' являются по сути одним и тем же (показатели билирубина), поэтому между ними силная корреляция, что говорит о коллинеарности признаков, поэтому один из признаков убирается при обучении модели, в данном случае не учитывается 'Direct\_Bilirubin';
* между 'Alamine\_Aminotransferase' и 'Aspartate\_Aminotransferase' наблюдается высокая корреляция, но данные признаки отвечают за разные показатели, поэтому оба применяются при обучении;
* 'Total\_Proteins' и 'Albumin' имеют высокую корреляцию, что говорит об их коллинеарности;
* 'Albumin' и 'Albumin\_and\_Globulin\_Ratio' также имеют значительную корреляцию, но они представляют разные данные по-разному (если основываться медицинских формулах): 'Albumin' — абсолютное значение альбумина. 'Albumin\_and\_Globulin\_Ratio' — соотношение альбумина к глобулинам.

В итоге 'Total\_Proteins' исключается при обучении модели, так как этот показатель полностью зависит от других признаков, а

'Albumin', 'Albumin\_and\_Globulin\_Ratio' оставляются.

Далее данные делятся на тренировочные и тестовые (рис. 25).

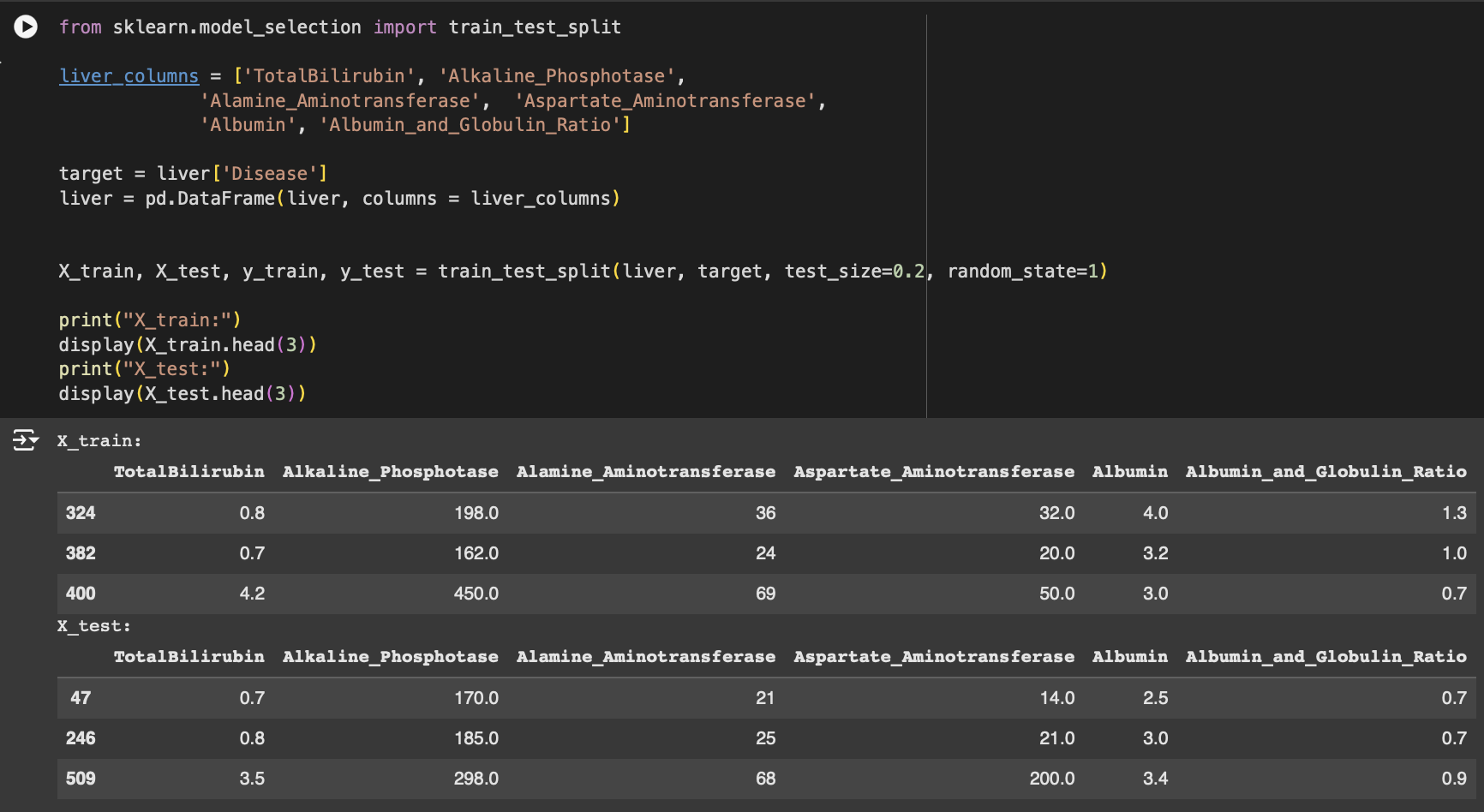


Рисунок 25 – Разделение на данных на тренировочную и тестовую выборки

Выполняется стандартизация числовых данных с помощью `StandardScaler` (рис. 26).

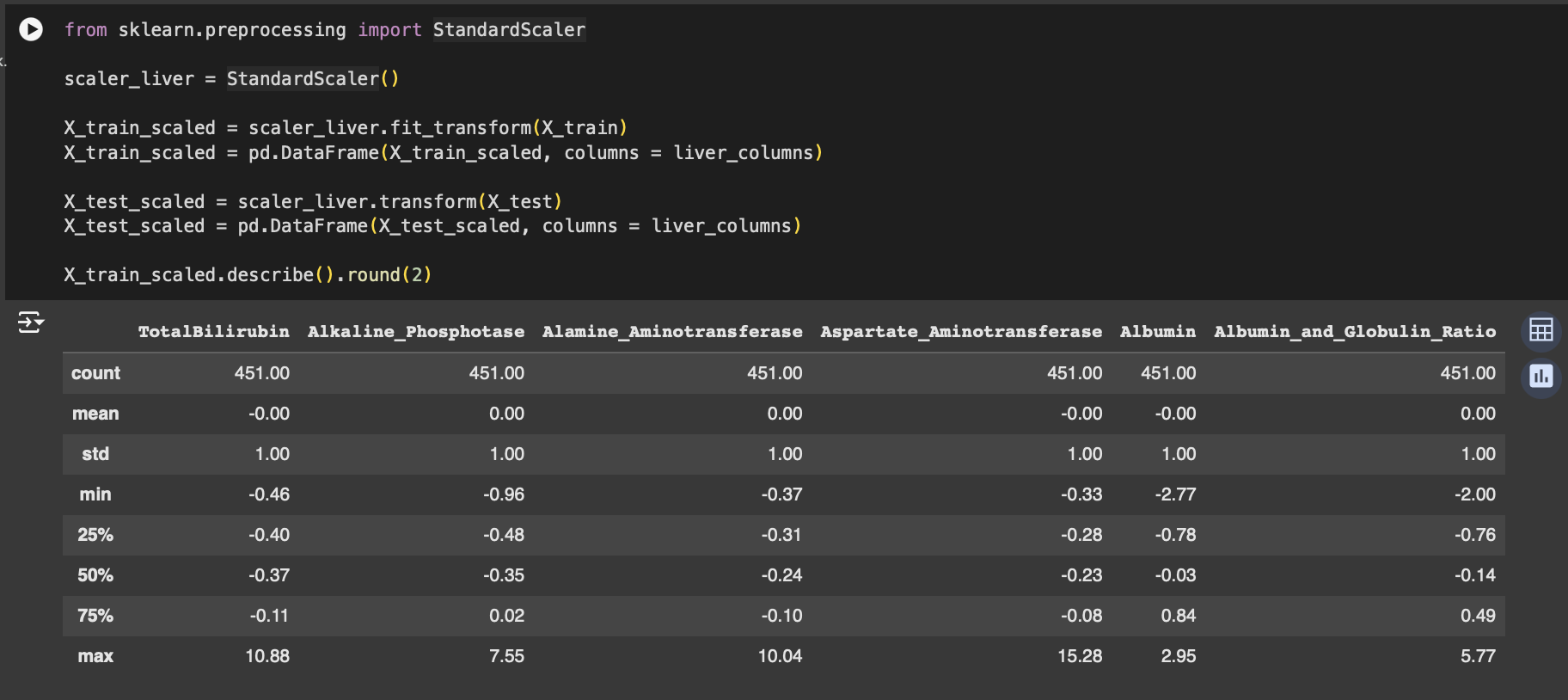


Рисунок 26 – Стандартизация данных

Стандартизация была проведена успешно, все средние равны нулю, а стандартных отклонения равны 1.

Для определения кол-ва кластеров используется "метод локтя" (рис. 27).

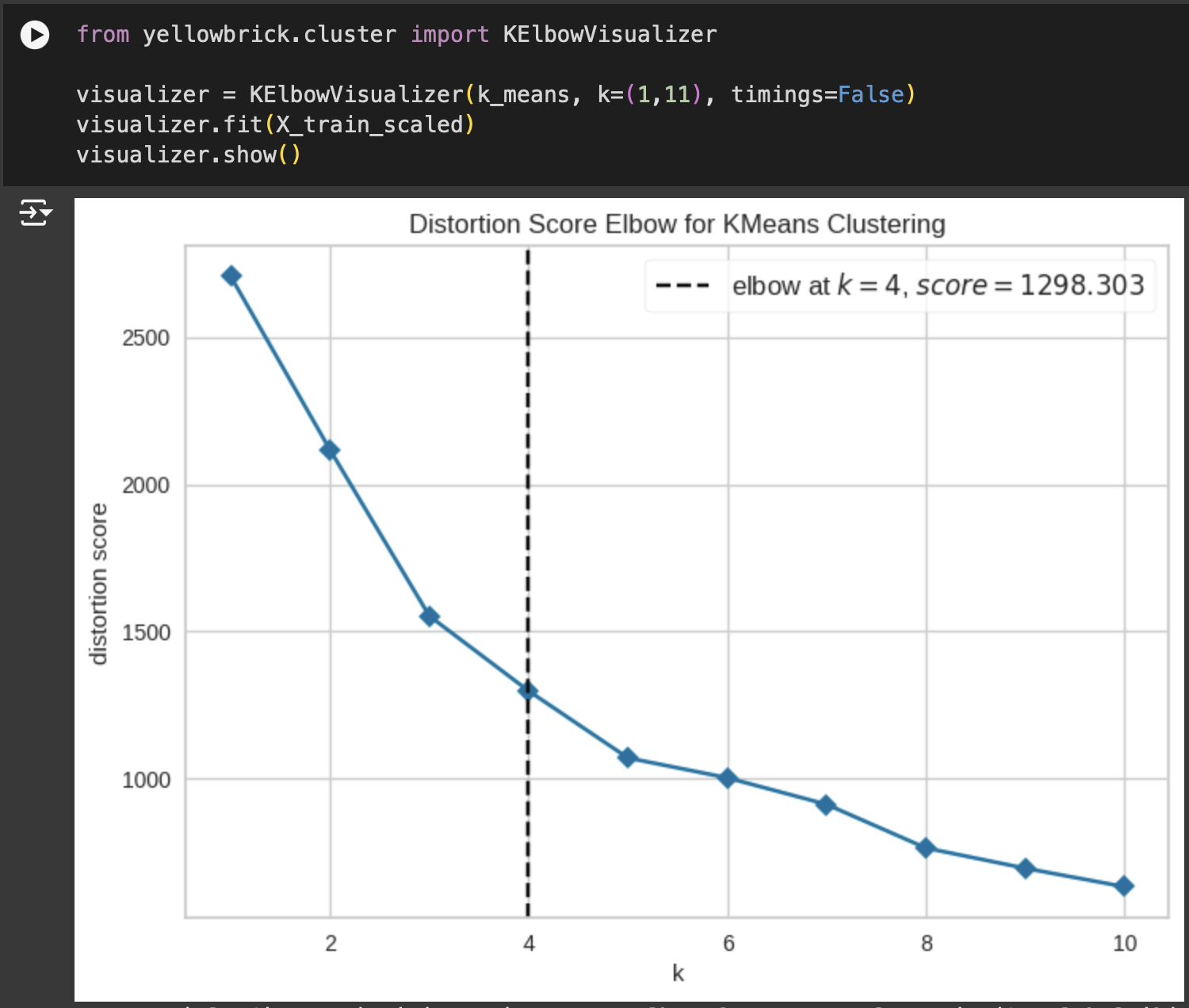


Рисунок 27 – Определение кол-ва кластеров

По "методу локтя" не получилось однозначно определить число кластеров, но видно, что distortion score (внутрикластерная дисперсия) быстро снижается при увеличении числа кластеров от 1 до 4, а дальнейшее увеличение не приводит к значительному улучшению, поэтому число кластеров равно 4.

Выполняется кластеризация объектов методом k-means. Подбирается количество кластеров равное 4 (рис. 28).

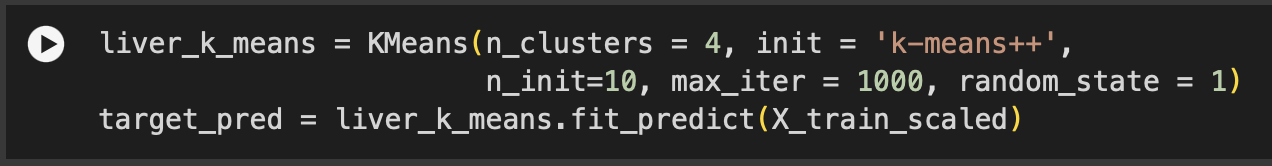


Рисунок 28 – Обучение модели и предсказание меток кластеров

Определяются объекты, относящиеся к одному кластеру, и делаются выводы о каждом кластере (рис. 29).



Рисунок 29 – Визуализация объектов и предсказанных меток кластеров

По графику, построенному по признакам 'Albumin\_and\_Globulin\_Ratio' и 'TotalBilirubin', можно сделать следующие вывод, что модель K-средних довольно хорошо разделила наблюдения на 4 класса. Хоть и нельзя судить о всей кластеризации основываясь на графике только двух признаков из 6, все равно можно сказать, что разделение на классы хорошо прослеживается.

На графике также видны точки, которые значительно удалены от остальных данных. Эти выбросы могут указывать на уникальные наблюдения, которые выходят за рамки нормального распределения для большинства классов.

Далее, для объективной оценки работы модели подсчитываются средние показатели по каждому признаку каждого кластера. Для этого сначала создается новый датафрейм на основе тренировочных данных, затем к нему присоединяется целевая переменная (рис. 30).



Рисунок 30 – Создание нового датафрейма

Далее производится группировка по целевому признаку и подсчет средних значений каждого признака. Для удобства датафрейм транспонируется

(рис. 31).

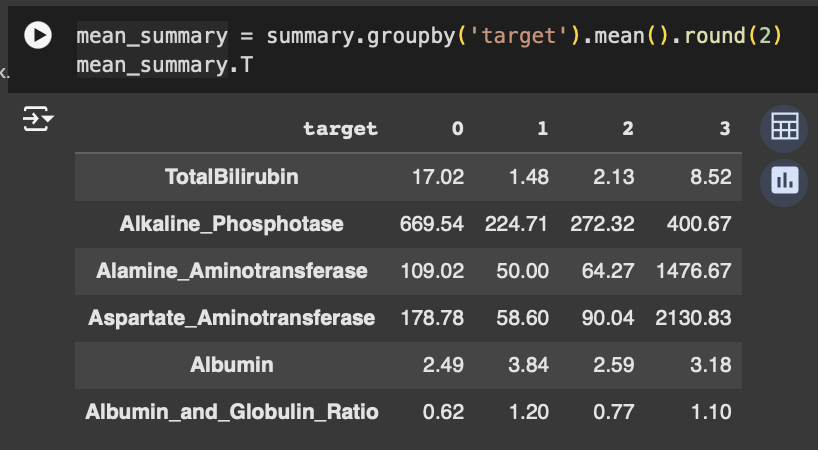


Рисунок 31 – Средние значения каждого признака для каждого кластера

На основе представленных средних значений для разных признаков можно сделать следующие выводы:

* класс 0 имеет высокие значения TotalBilirubin, Alkaline\_Phosphotase, ALT и AST, что скорее всего говорит о том, что в данную группу попали люди с заболеванием печени;
* у 1 и 2 классов в среднем очень похожие значения признаков, что говорит о том, что возможно данные классы могли быть объединены в один;
* в классе 3 наблюдаются очень высокие уровни Alamine\_Aminotransferase и Aspartate\_Aminotransferase, что также вероятно говорит о наличии болезни.

Далее считается коэффициент силуэта (рис. 32).

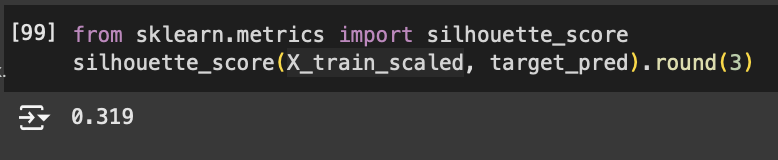


Рисунок 32 – Подсчет коэффициента силуэта

Значение 0.319 указывает на то, что кластеризация выполняется на низком уровне. Значения выше 0 демонстрируют, что объекты внутри кластеров находятся ближе друг к другу, чем к объектам в других кластерах.

Далее подсчитывается метрика ARI (рис. 33).

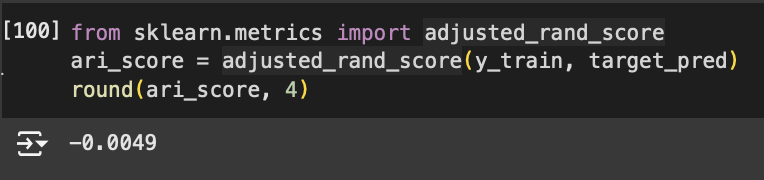


Рисунок 33 – Подсчет ARI

ARI принимает значение -0,0049, что указывает на то, что кластеризация была проведена не лучшим образом. Это может означать, что объекты неправильно сгруппированы, или группы сильно пересекаются.

Далее подсчитаем коэффициент однородности (рис. 34).

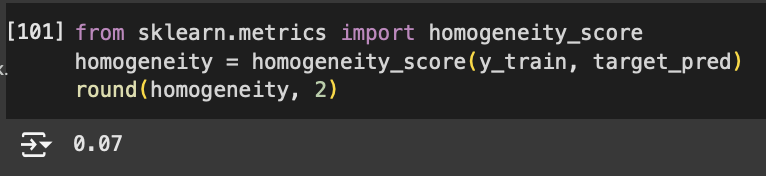


Рисунок 34 – Подсчет коэффициента однородности

Значение 0.07 указывает на то, что объекты в кластерах скорее всего принадлежат не к одному классу, они не однородны.

Далее выполняется кластеризация иерархическим агломеративным методом. Для определения кол-ва кластеров строится дендрограмма (рис. 35).

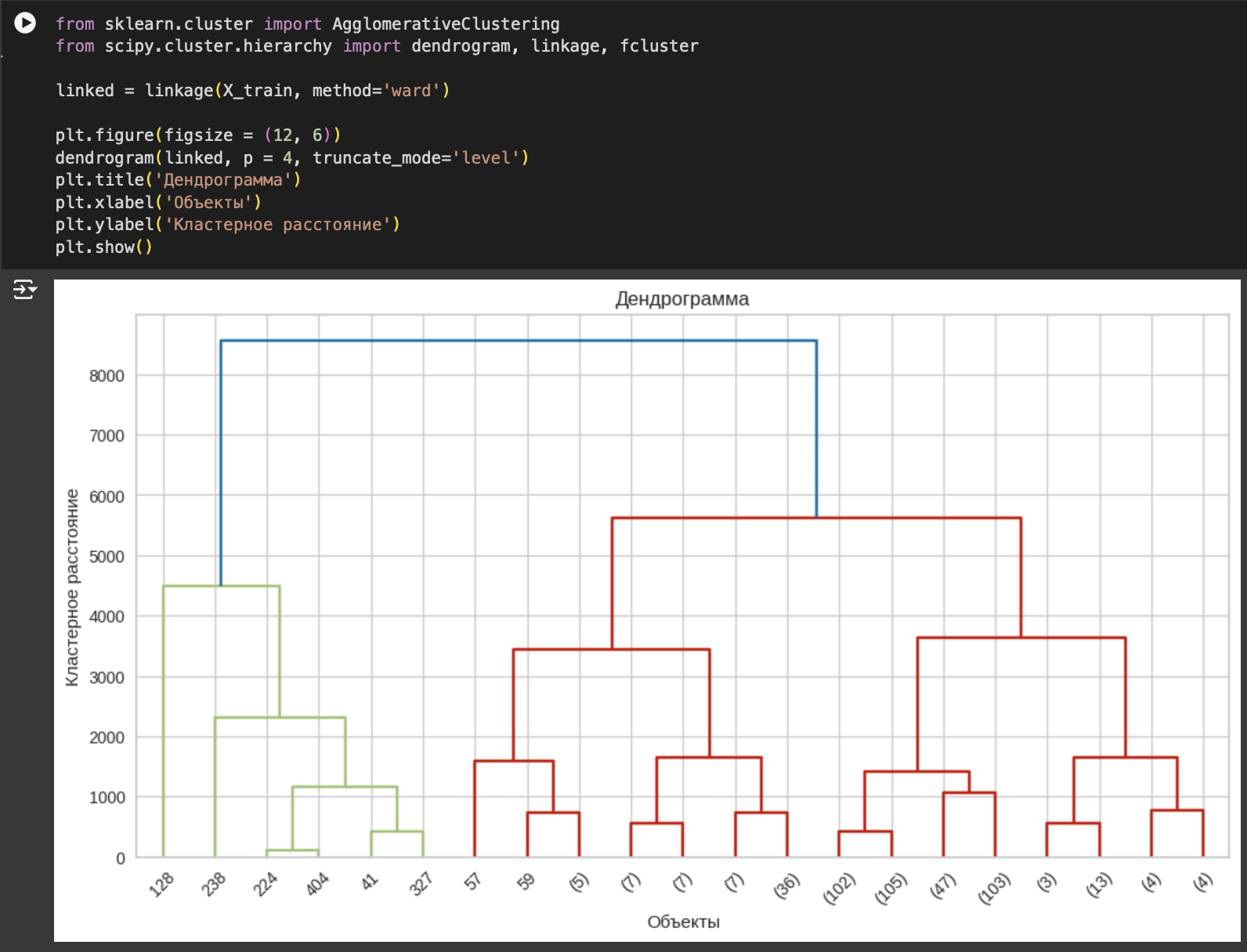


Рисунок 35 - Дендрограмма

Дендрограмма показывает, как наблюдения объединялись в кластеры. На ее основе делается вывод, что наблюдения относятся к 2 кластерам, так как на этом уровне произошел резкий скачек дендрограммы.

Далее выполняется кластеризация иерархическим агломеративным методом. Подсчитываются средние показатели по каждому признаку каждого кластера и вычисляются несколько метрик качества (рис. 36).

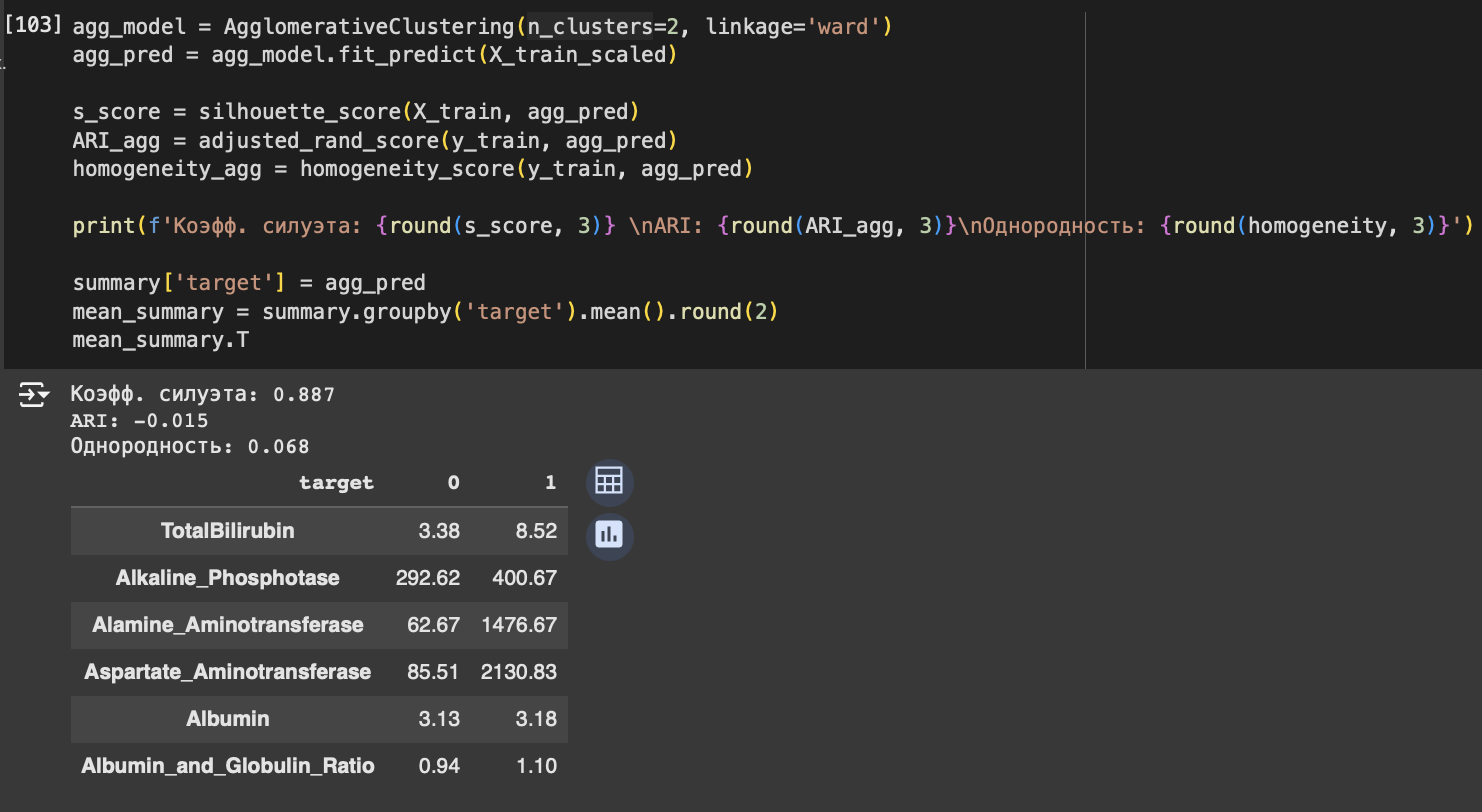


Рисунок 36 – Кластеризация иерархическим агломеративным методом.

На основе сравнительной таблицы делаются выводы:

* класс 0 имеет низкий уровень почти всех показателей, кроме 'Albumin\_and\_Globulin\_Ratio', что скорее всего указывает на то, что в эту группу попали здоровые люди;
* класс 1 имеет, наоборот, высокие показатели, что может говорить о болезни.

Вывод по метрикам качества:

* коэффициент силуэта (0.887) - это высокий показатель, что говорит о хорошем разделении кластеров. Объекты внутри кластеров довольно похожи;
* ARI (-0.015) близкое к нулю указывает на то, что результат кластеризации не совпадает с истинными метками классов;
* очень низкая однородность означает, что внутри кластеров данные очень разнообразны и не похожи друг на друга.

Модель кластеризации в целом разделяет данные на два кластера с явными различиями по значениям признаков, но низкая однородность и низкий ARI показывают, что кластеры могут быть не полностью различимы моделью.

**Вывод по работе**

В ходе выполнения лабораторной работы были изучены методы и алгоритмы кластеризации данных на практике, с целью выявления закономерностей.

В первой части работы была выполнена кластеризация на основе сгенерированного набора данных. Была выполнена стандартизация данных, обучение модели методом К-средних и предсказание для двух, трех и четырех кластеров. По результатам анализов, метрик и при помощи «метода локтя» было выявлено, что данные лучше всего делятся на 3 кластера.

Во второй части работы была произведена работа по кластеризации датасета с информацией о пациентах и статусе их заболевания (болезни печени).

Первым шагом была выполнена предварительная обработка данных, затем проанализированы зависимости между признаками, выявлены "шумы" и коллинеарные признаки, которые не учитывались при дальнейшем обучении модели. Признаки были разделены на обучающую и тестовую выборки, после чего стандартизированы.

Далее было определено кол-во кластеров методом К-средних, обучение модели и предсказание. В среднем метрики качества модели показали довольно низкие значения, что могло объясняться некорректностью данных, либо тем, что признаки плохо описывают целевой признак.

Далее для этого же датасета количество кластеров определялось с помощью построения дендрограммы. Проводилась кластеризация иерархическим агломеративным методом.

По результатам работы, лучшие показатели продемонстрировала модель кластеризации иерархическим агломеративным методом для 2 кластеров, нежели чем модель обученная методом К-средних для 4 кластеров.

По показателям метрик качества и сравнительным таблицам со средними значениями всех признаков каждого кластера было выявлено, что данные скорее всего содержат именно 2 кластера. Но ни одна модель с высокой точностью не определила правильный целевой признак.

**Выполнение дополнительного задания**

Задание 6: выполнить подсчёт 4 любых метрик для разного количества кластеров (от 2 до 10). Построить график - по оси х - кол-во кластеров, по у - значение метрики (отразить все 4 метрики на одном графике разным цветом).

Выполнение задания продемонстрировано на рисунке 37.

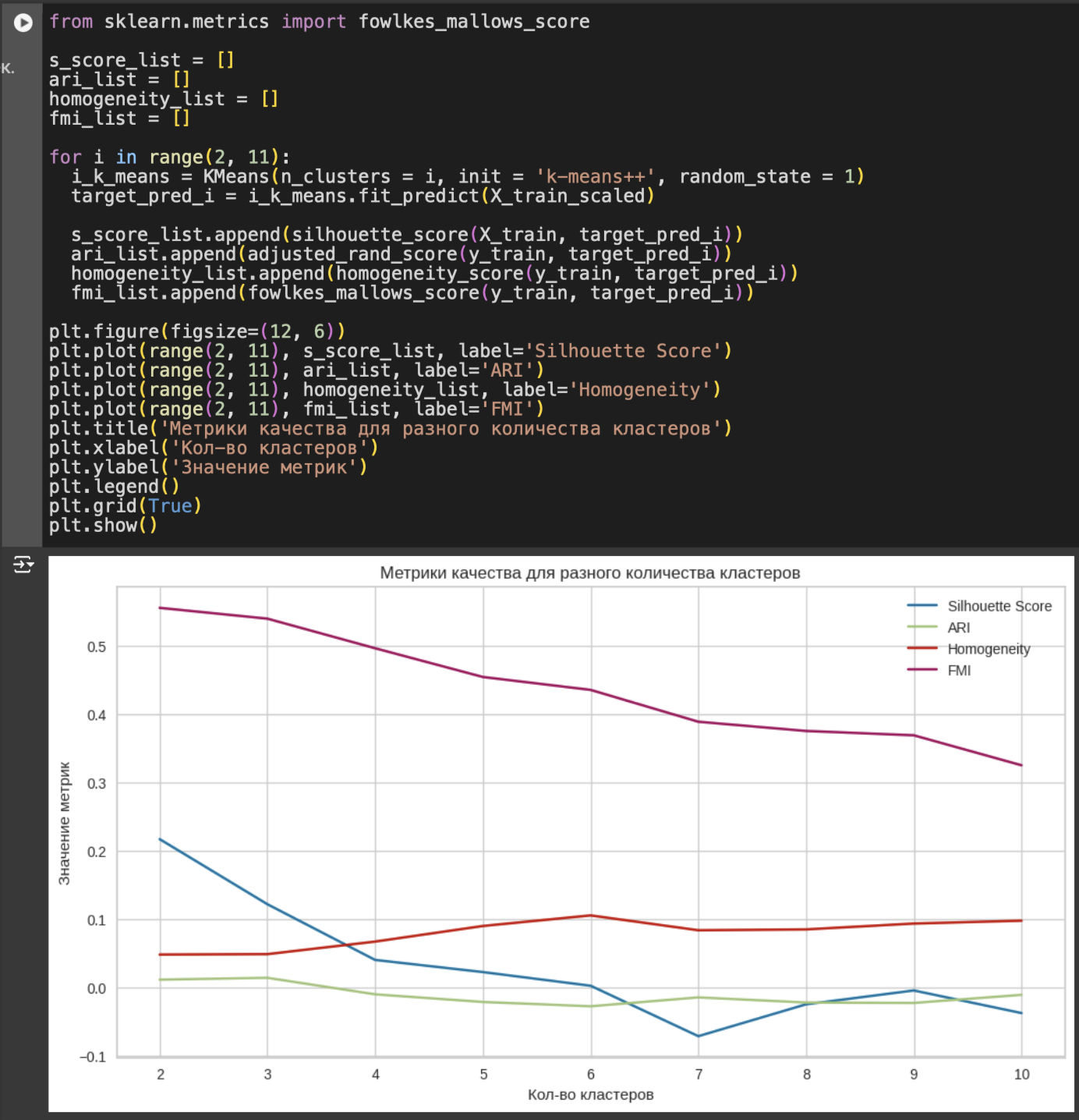


Рисунок 37 – Подсчет метрик для разного кол-ва кластеров

На основе анализа метрик качества модели К-средних, оптимальное количество кластеров для данной набора данных составляет 2, так как при 2 кластерах наблюдается самый высокий коэффициент силуэта и FMI, что говорит о самых сильных различиях в наблюдений одного кластера от наблюдений другого и о самой высокой доле правильно кластеризованных наблюдений именно при разделении наблюдений на 2 кластера.