МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Факультет прикладной математики и информатики
Кафедра математического моделирования и анализа данных

Кластерный анализ неоднородных данных

Отчет по лабораторной работе №5
студентки 3 курса 7 группы
Летецкой М.С.

Преподаватель
Малюгин В.И.

Минск, 2024

УСЛОВИЕ ЗАДАНИЯ

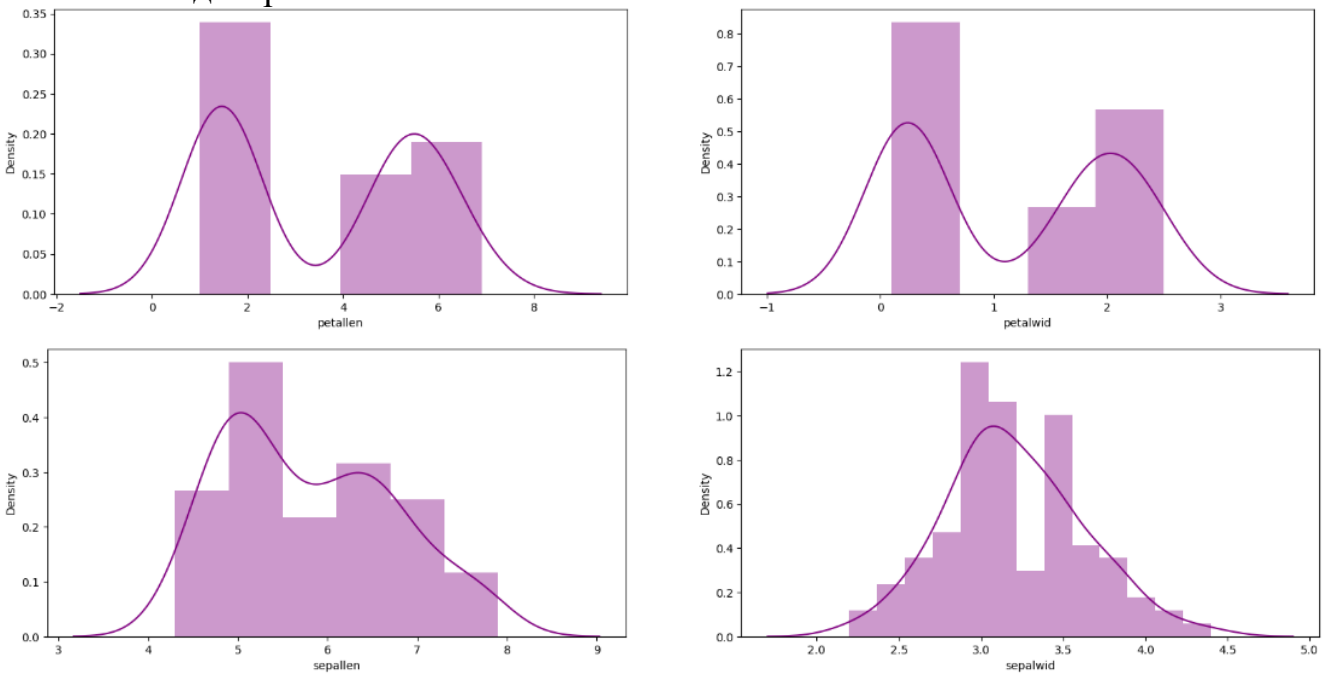
Указание: использовать неклассифицированную выборку из смеси распределений (13) для всех переменных;

использовать алгоритмы: К-средних и иерархический кластерный анализ, провести сравнительный анализ результатов.

АЛГОРИТМ К-СРЕДНИХ

Алгоритм k-средних (k-means) – это алгоритм неконтролируемого обучения, который используется для разделения набора данных на k кластеров (групп). Основная цель k-средних - найти центры кластеров (центроиды), которые минимизируют суммарное расстояние от точек данных до центроидов их кластеров.

Метод к-средних подходит для хорошо разделённых данных. Чтобы понять, соответствуют данные этому условию или нет, были построены столбчатые диаграммы:



Так как наблюдается явное разделение по признакам petallen и petalwid, то в дальнейшем работа будет проводиться на их основе.

Пусть количество кластеров неизвестно. Для его нахождение будет использован метод локтя.

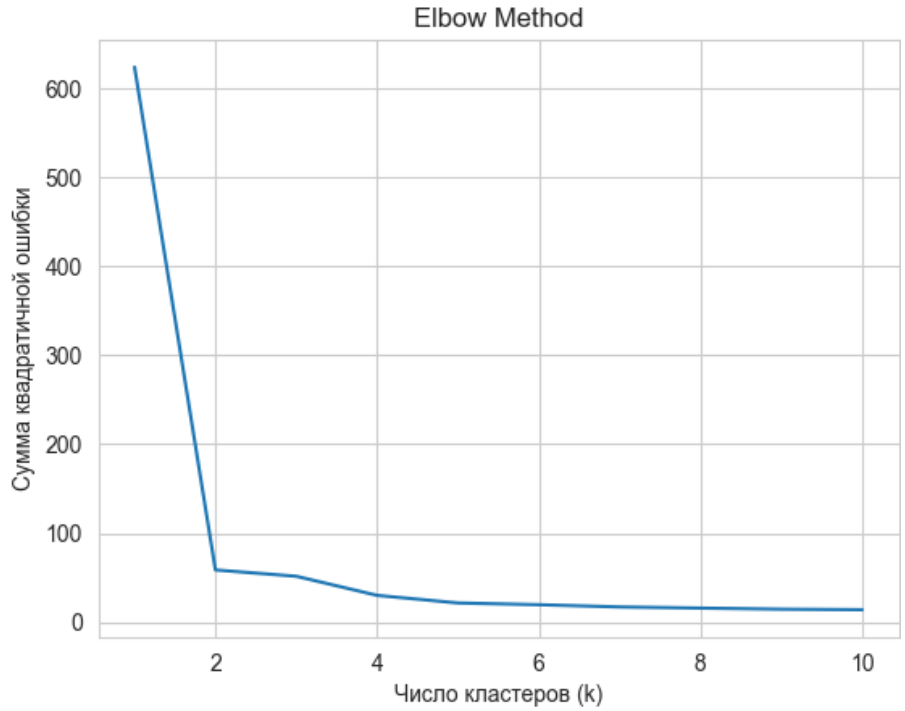
Метод локтя (Elbow Method) — это эвристический метод, используемый для определения оптимального числа кластеров (k) в алгоритме k-средних (и других методах кластеризации). Он основан на визуальном анализе графика, отображающего зависимость суммы квадратов расстояний (SSE — Sum of Squared Errors) от числа кластеров.

Схема работы метода локтя:

1. Запуск k-means для различных значений k: Алгоритм k-средних запускается несколько раз для разных значений k (числа кластеров), например, от 1 до 10 или больше. Для каждого значения k алгоритм сходится, и вычисляется SSE.

2. Вычисление SSE: SSE — это сумма квадратов евклидовых расстояний между каждой точкой данных и центроидом кластера, к которому она принадлежит. Чем меньше SSE, тем лучше кластеризация (в идеале, все точки находятся очень близко к своим центроидам).
3. Построение графика: Строится график, где по оси X откладывается число кластеров (k), а по оси Y — соответствующее значение SSE.
4. Поиск “локтя”: График обычно имеет вид, напоминающий локоть руки. “Локоть” — это точка на графике, где уменьшение SSE с ростом k начинает замедляться, то есть скорость уменьшения SSE резко снижается. Это значение k и считается оптимальным числом кластеров.

В соответствии с этим методом был построен график:



Явная точка сгиба наблюдается при числе кластеров, равном 2.

Метод локтя не всегда работает, т.к. на графике точка сгиба может быть не видна, поэтому был использован метод силуэта.

Метод силуэта предоставляет количественную оценку того, насколько хорошо каждая точка данных соответствует своему кластеру и насколько она отличается от других кластеров. Эта оценка основана на вычислении силуэтного коэффициента (silhouette coefficient) для каждой точки данных.

Ключевые идеи метода силуэта:

1. Силуэтный коэффициент: для каждой точки данных вычисляется силуэтный коэффициент, который показывает, насколько хорошо эта точка “вписывается” в свой кластер и насколько она “отделена” от других кластеров.
2. Значение от -1 до +1: силуэтный коэффициент может принимать значения от -1 до +1:
    - +1: означает, что точка находится далеко от соседних кластеров и хорошо соответствует своему кластеру. Это идеальная ситуация.

- 0: означает, что точка находится близко к границе между двумя кластерами, то есть неясно, к какому кластеру ее отнести.
- -1: означает, что точка, вероятно, отнесена к неверному кластеру. Она, скорее, должна была быть в другом кластере.

3. Средний силуэтный коэффициент: после вычисления силуэтных коэффициентов для всех точек данных, вычисляется средний силуэтный коэффициент для каждого кластера и для всего набора данных.
4. Оптимальное k: оптимальное число кластеров (k) выбирается на основе значений среднего силуэтного коэффициента. Обычно выбирают то значение k, при котором средний силуэтный коэффициент является максимальным.

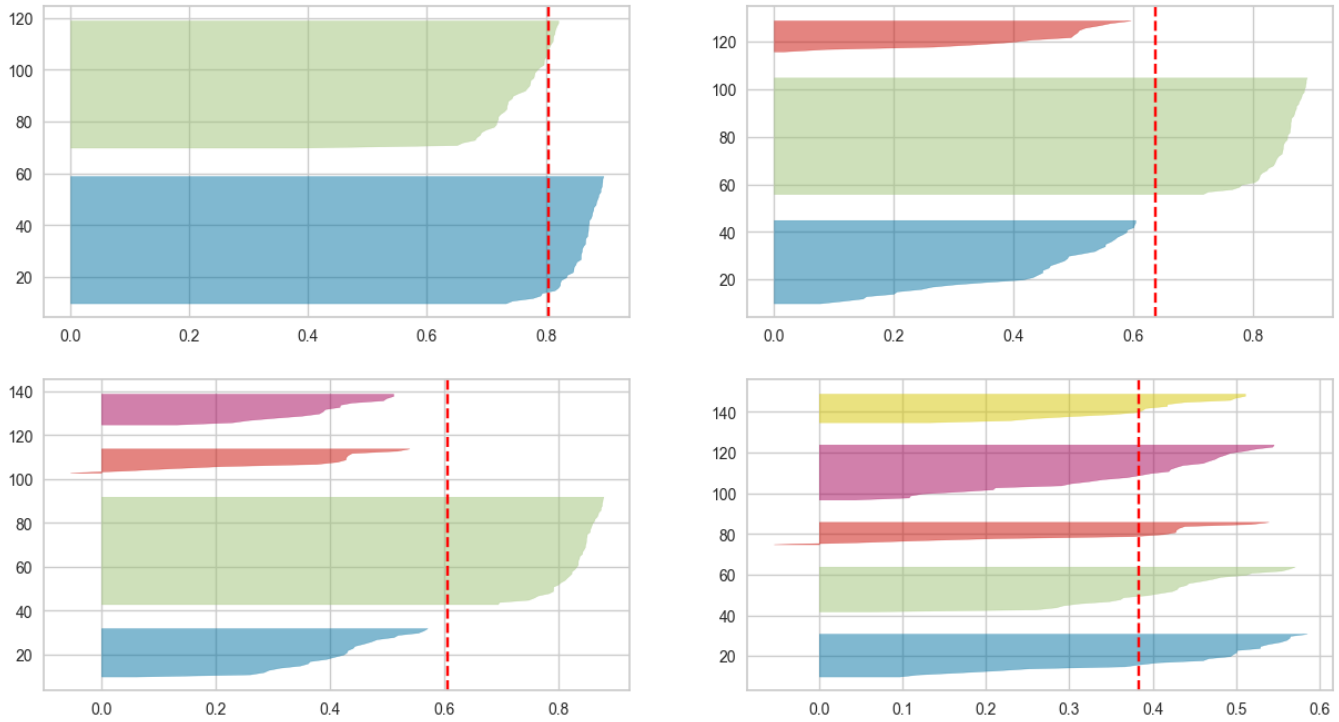
Для точки данных *i* силуэтный коэффициент *s(i)* вычисляется следующим образом:

s(i) = (b(i) - a(i)) / max(a(i), b(i))

где:

- a(i): Среднее расстояние от точки i до всех других точек в том же кластере, к которому принадлежит i. (Средняя внутрикластерная дистанция)
- b(i): Минимальное среднее расстояние от точки i до всех точек в любом другом кластере, кроме того, к которому принадлежит i. (Минимальная межкластерная дистанция)

Был получен график (silhouette plot):



Графики показывают коэффициенты силуэта для разного количества кластеров, от 2 до 5. Каждый график соответствует определенному количеству кластеров (k), используемых в алгоритме кластеризации. Ось X представляет

значения коэффициента силуэта (в диапазоне от -1 до 1, где 1 — наилучшее значение, а -1 — наихудшее), причем большие значения находятся справа. Ось Y обозначает отдельные точки, сгруппированные по кластерам. Вертикальная красная пунктирная линия на каждом графике представляет собой средний балл силуэта для данного количества кластеров.

Принимая во внимание средний балл силуэта и визуальный анализ, выбор двух кластеров (k=2) дает более последовательную и четко определенную кластеризацию.

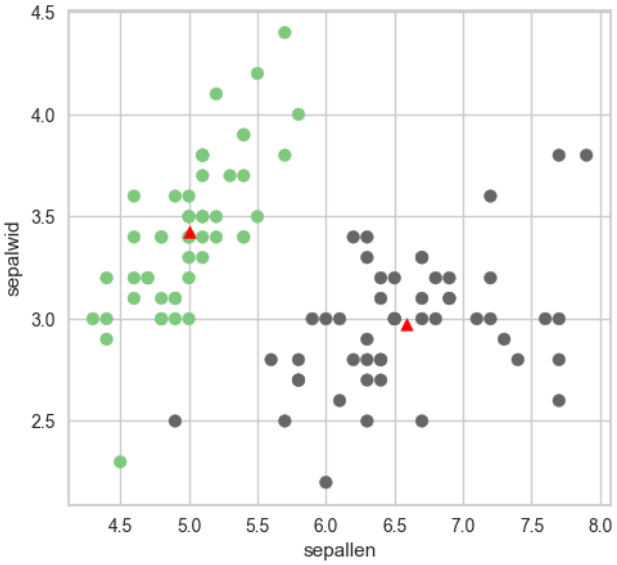
Оба метода определения количества кластеров сошлись на том, что оптимальное количество групп – 2.

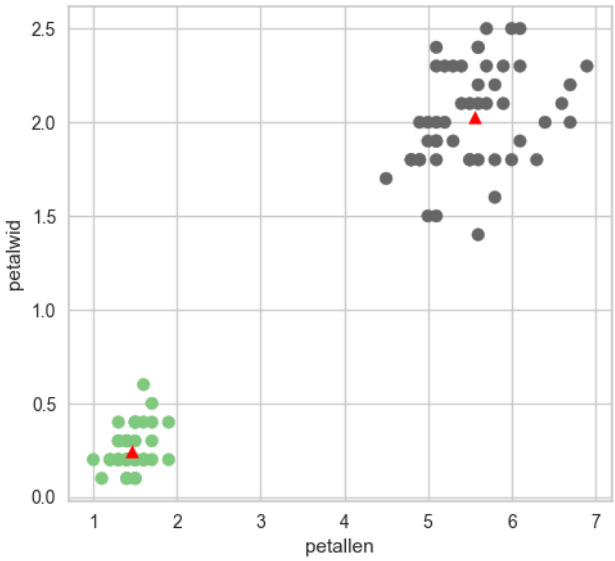
Была создана модель к-средних, к=2:

```
kmeans = KMeans(n_clusters = 2, random_state = 2)
kmeans.fit(X)
```

```
KMeans
KMeans(n_clusters=2, random_state=2)
```

В результате была проведена кластеризация данных, для визуализации ниже приведены диаграммы рассеивания:





ИЕРАРХИЧЕСКИЙ КЛАСТЕРНЫЙ АНАЛИЗ

Иерархический кластерный анализ — это метод кластеризации, который строит иерархию кластеров, представленную в виде древовидной структуры, называемой дендрограммой . Основная идея состоит в том, чтобы итеративно объединять или разделять кластеры на основе некоторых метрик расстояния или связности.

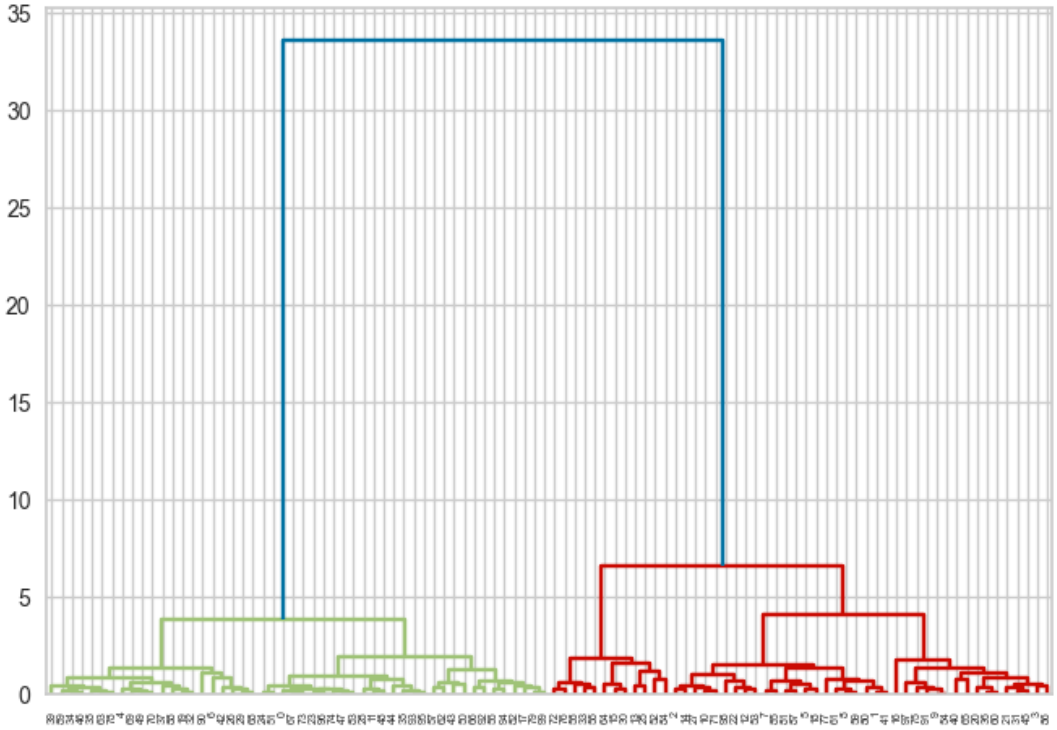
Основные идеи иерархического кластерного анализа:

1. Иерархия кластеров: Результатом работы иерархического кластерного анализа является иерархия кластеров, где кластеры на более низких уровнях вкладываются в кластеры на более высоких уровнях.

2. Дендрограмма: Иерархия кластеров представлена в виде дендрограммы, которая отображает процесс объединения или разделения кластеров.
3. Агломеративный и дивизионный подходы: существуют два основных подхода к иерархической кластеризации:
    - Агломеративный: начинается с отдельных точек и постепенно перемещается в кластеры (снизу вверх).
    - Дивизионный: начинается со всего набора данных и рекурсивно разделяется на более мелкие кластеры (сверху вниз).

Дендрограмма — это дерево, которое показывает, как кластеры объединяются или разделяются на каждом этапе итеративного процесса. По вертикальной оси дендрограммы отображают расстояние между кластерами, а по горизонтальной — точки расположения или кластеры.

В ходе проведения данного анализа была построена дендрограмма:



В результате проведения иерархического кластерного анализа данные были разбиты на два кластера.

ВЫВОДЫ

Сходства методов:

- Методы кластеризации: оба являются методами кластеризации данных, предназначенными для группировки похожих объектов в кластеры.
- Неконтролируемое обучение: оба относятся к методам неконтролируемого обучения, то есть не требуются размеченные данные для обучения.
- Используют метрику расстояния: оба метода используют метрики расстояния (например, евклидово расстояние) для измерения расстояния между объектами.

Лучше использовать:

- Иерархический кластерный анализ:
  - Когда нужно понять иерархическую структуру данных.
  - Когда не известно заранее количество кластеров.
  - Когда размер набора данных относительно небольшой.
- Метод k-средних:
  - Когда нужно быстро кластеризовать большой набор данных.
  - Когда известно или можно оценить приблизительное количество кластеров.
  - Когда выбросы не являются большой проблемой.