

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ

Федеральное государственное автономное  
образовательное учреждение высшего образования  
«Санкт-Петербургский национальный исследовательский университет  
информационных технологий, механики и оптики»

**ЛАБОРАТОРНАЯ РАБОТА № 2**

по дисциплине

“Современные инструменты анализа данных”

*Выполнил:*  
Джафари Хоссаин

Студент группы К3340

*Преподаватель:*  
Добренко Наталья Викторовна

Санкт-Петербург, 2025

## **Содержание**

Цель работы.....	3
Краткое описание использованных методов кластеризации.....	3
Описание процесса генерации данных и применения методов .....	4
Визуализация результатов .....	5
Анализ результатов и сравнение эффективности методов .....	12
Выводы.....	14

# Цель работы

Целью данной лабораторной работы является изучение и практическое применение двух алгоритмов кластеризации данных:

1. **K-Means (K-средних)** — центроидный алгоритм кластеризации
2. **DBSCAN** — алгоритм кластеризации на основе плотности

Работа включает:

- Генерацию синтетических данных с известной структурой кластеров
  - Применение алгоритмов к синтетическим данным для проверки их работоспособности
  - Применение алгоритмов к реальному датасету (Clustering Penguins Species)
  - Сравнение эффективности методов на разных типах данных
  - Анализ влияния параметров на качество кластеризации
- 

## Краткое описание использованных методов кластеризации

### 1. K-Means (K-средних)

**K-Means** — итеративный центроидный алгоритм кластеризации. Алгоритм работы:

- Инициализация К центроидов
- Назначение каждой точки ближайшему центроиду
- Пересчет позиций центроидов как среднее точек в кластере
- Повторение шагов 2-3 до сходимости

**Особенности:** требует заранее задать число кластеров K, предполагает сферическую форму кластеров, все точки распределяются по кластерам, быстро работает, чувствителен к выбросам.

**Метрики:** инерция (сумма квадратов расстояний до центроидов), коэффициент силуэта (от -1 до 1). Методы подбора K: метод локтя и коэффициент силуэта.

### 2. DBSCAN

**DBSCAN** — алгоритм кластеризации на основе плотности. Основные понятия:

- **Точка ядра:** в окрестности (радиус eps) находится не менее min\_samples точек

- **Граничная точка:** в окрестности точки ядра, но не является точкой ядра
- **Точка шума:** не является ни точкой ядра, ни граничной

**Параметры:** eps (максимальное расстояние между точками кластера), min\_samples (минимальное число точек для кластера).

**Особенности:** автоматически определяет число кластеров, может находить кластеры произвольной формы, устойчив к выбросам, требует тщательного подбора параметров, чувствителен к масштабу признаков.

---

## Описание процесса генерации данных и применения методов

### 1. Синтетические данные (make\_blobs)

**Генерация:** датасет создан с помощью make\_blobs (scikit-learn): 1000 точек, 2 признака, 4 кластера, cluster\_std=1.0, random\_state=42.

#### K-Means:

- Кластеризация с K=4: инерция 1948.73, силуэт 0.792
- Подбор оптимального K (от 2 до 8): метод локтя и силуэт подтвердили K=4

#### DBSCAN:

- Масштабирование данных (StandardScaler)
- Параметры: eps=0.3, min\_samples=5
- Результат: 3 кластера (вместо 4), силуэт 0.738

### 2. Реальный датасет (Clustering Penguins Species)

#### Предобработка:

- Загрузка из penguins.csv: 344 строки, 5 столбцов
- Выбор числовых признаков (4): culmen\_length\_mm, culmen\_depth\_mm, flipper\_length\_mm, body\_mass\_g
- Удаление пропусков и аномалий: итого 340 строк
- Масштабирование (StandardScaler)

#### K-Means:

- Подбор K: метод локтя — K=2 или K=3, силуэт — K=2 (0.532)
- Выбрано K=3 (соответствует 3 видам пингвинов)
- Результат: распределение 131, 123, 86 точек, инерция 373.93, силуэт 0.451

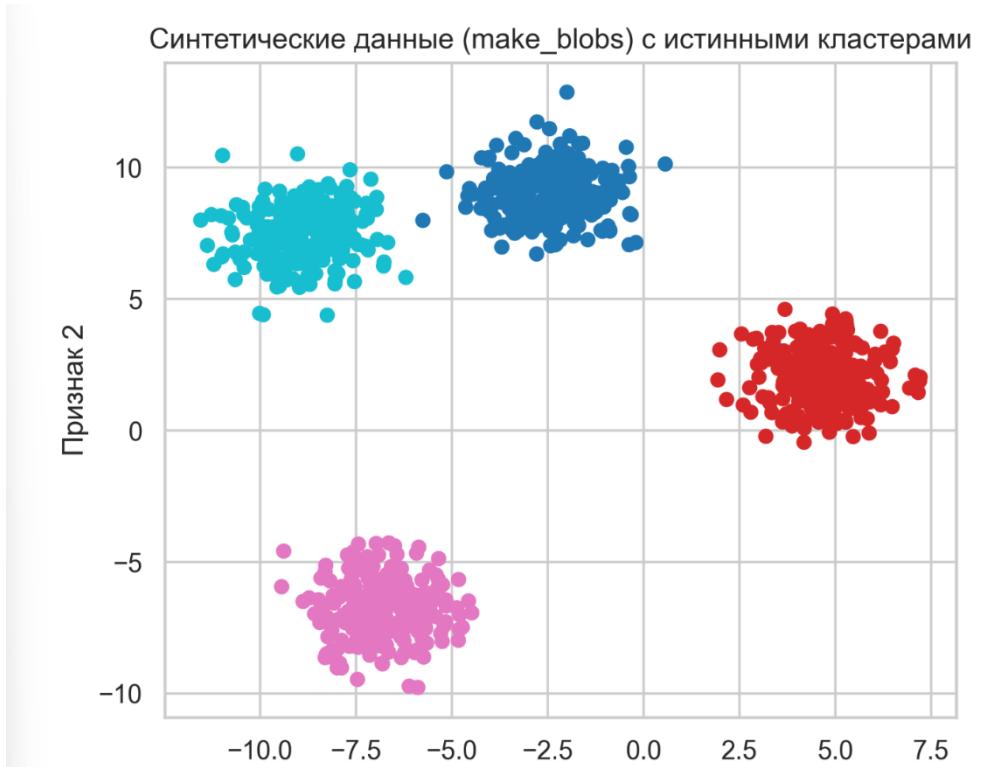
### DBSCAN:

- Параметры:  $\text{eps}=0.5$ ,  $\text{min\_samples}=5$
  - Результат: 4 кластера, 67 точек шума (19.7%), силуэт 0.424
  - Эксперименты: перебор  $\text{eps}$  [0.3-0.7] и  $\text{min\_samples}$  [3,5,7,10], визуализация тепловых карт
- 

## Визуализация результатов

### Синтетические данные

- Исходные данные:



*Рисунок 1: Синтетические данные (`make_blobs`) с истинными кластерами. Видны 4 четко разделенных сферических кластера.*

- Результаты K-Means:

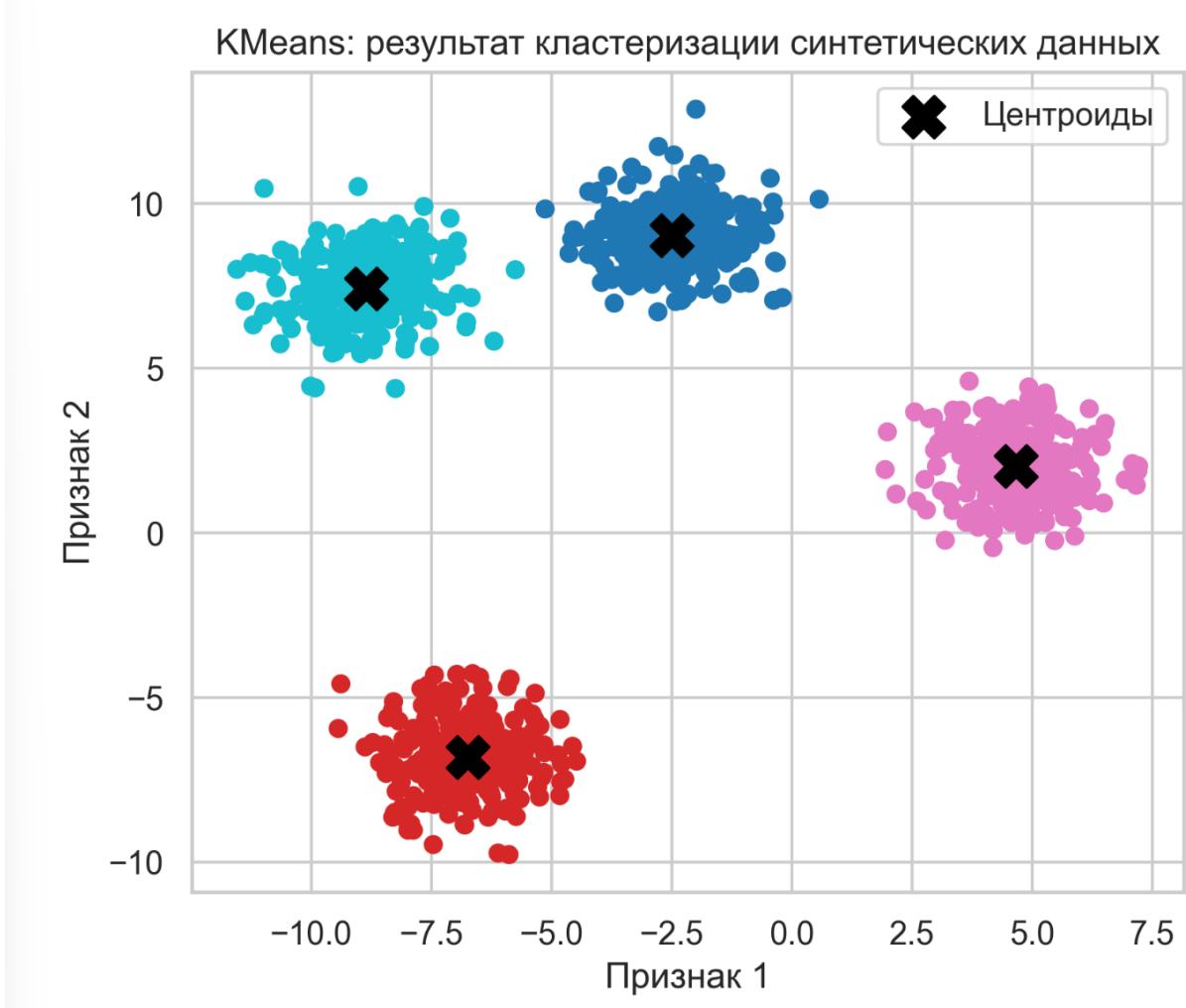


Рисунок 2: Результат кластеризации K-Means на синтетических данных. 4 кластера с центроидами (черные крестики), коэффициент силуэта 0.792.

- Метод локтя и коэффициент силуэта для K-Means

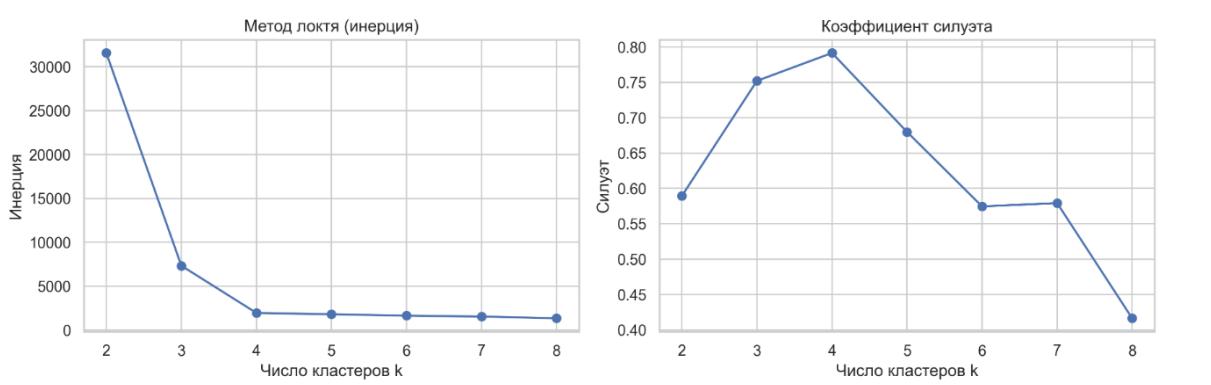


Рисунок 3: Подбор оптимального числа кластеров для K-Means. Слева: метод локтя (инерция), справа: коэффициент силуэта. Оба метода подтверждают оптимальность K=4.

- Результаты DBSCAN:

DBSCAN: результат кластеризации синтетических данных ( $\text{eps}=0.3$ ,  $\text{min\_samples}=5$ )

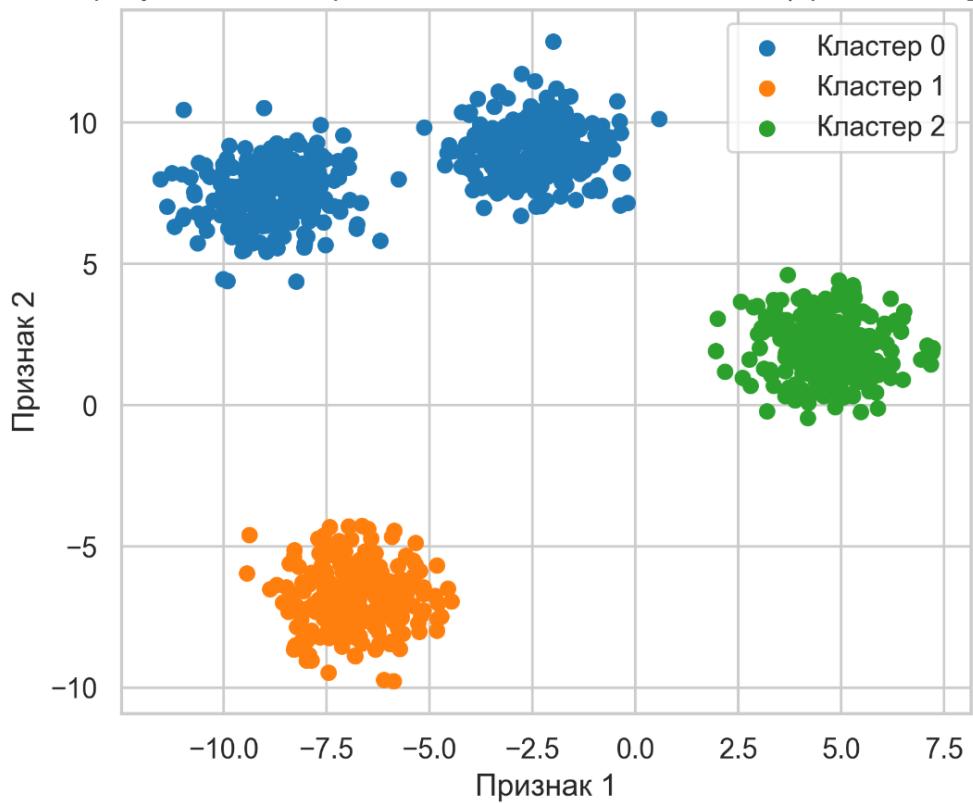


Рисунок 4: Результат кластеризации DBSCAN на синтетических данных. Найдено 3 кластера (вместо 4), точки шума отмечены черным цветом, коэффициент силуэта 0.738.

## Датасет Penguins

- **Предварительный анализ:** гистограммы распределений, матрица scatter plots

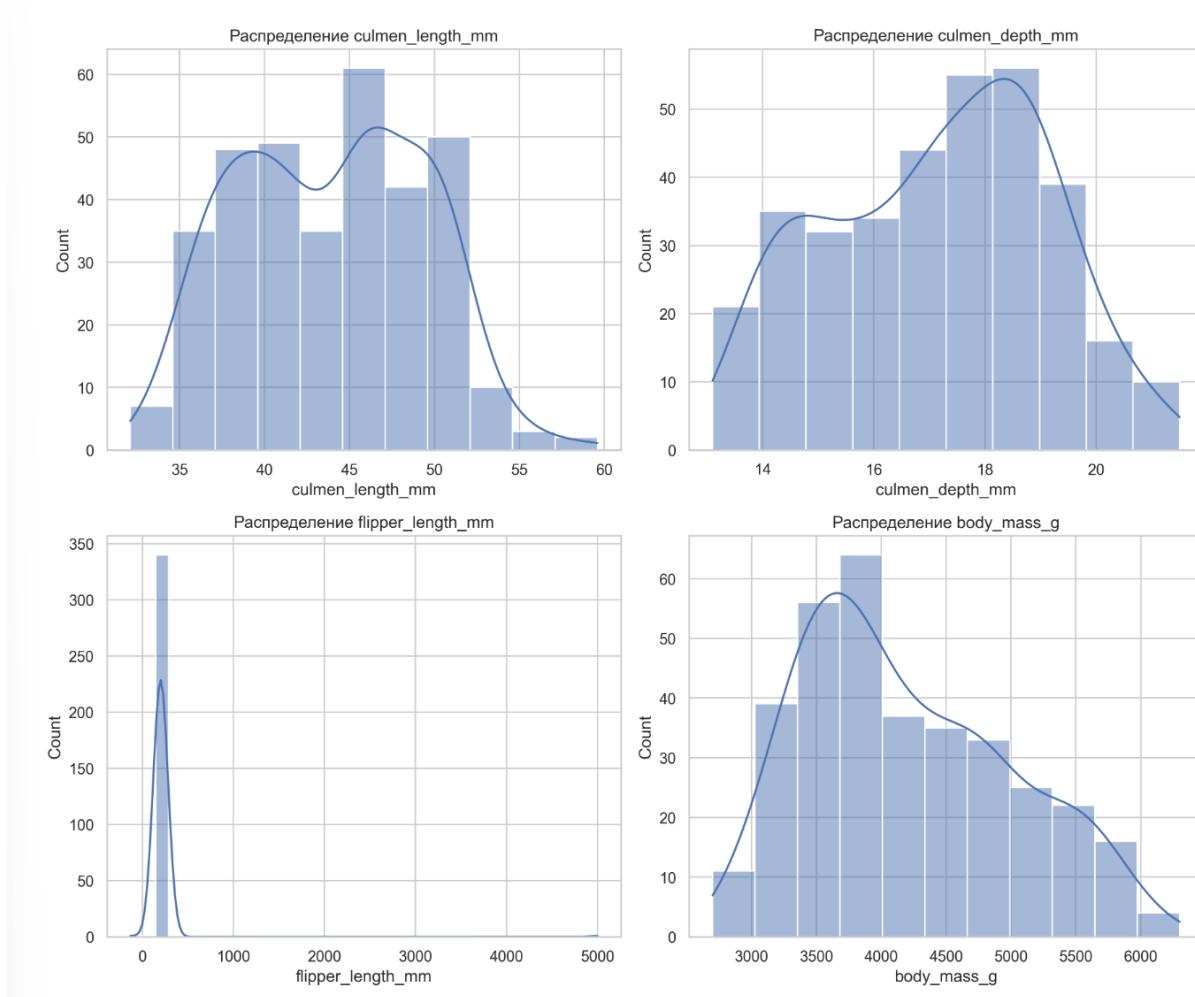


Рисунок 5: Распределения числовых признаков датасета penguins.

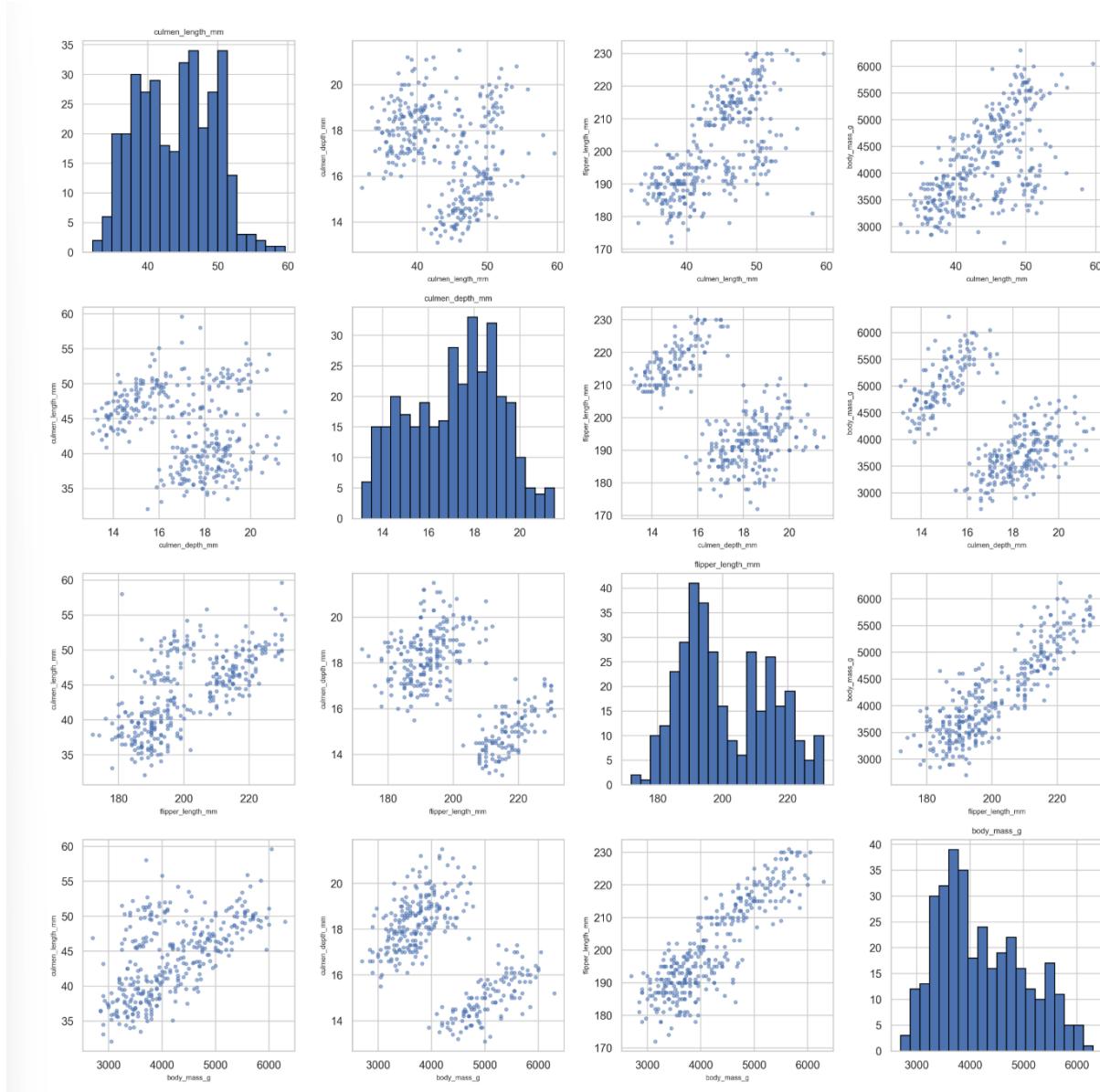


Рисунок 6: Матрица scatter plots для пар признаков датасета penguins.

- Результаты K-Means:

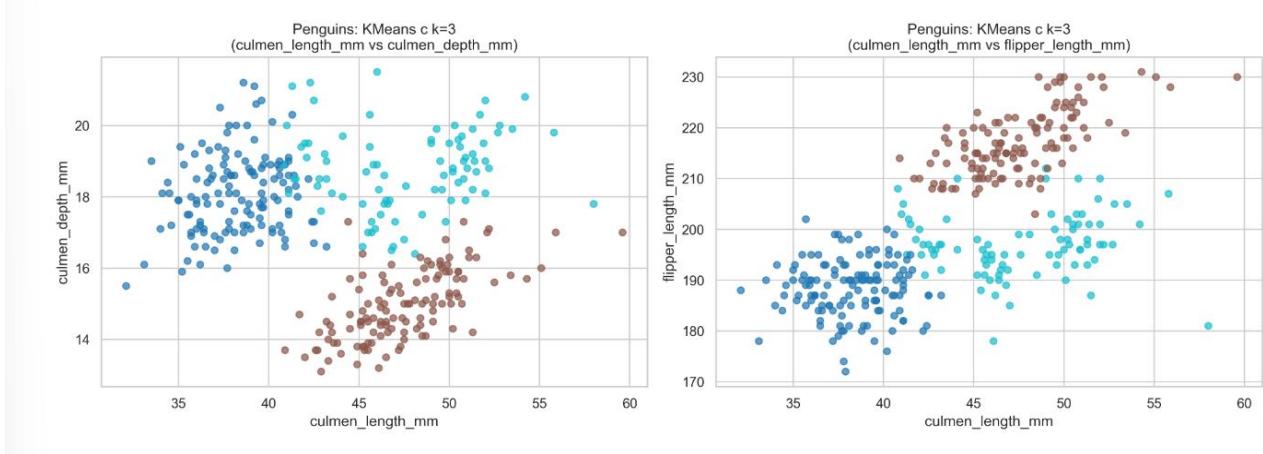


Рисунок 7: Подбор оптимального числа кластеров для K-Means на данных penguins. Метод локтя указывает на  $K=2$  или  $K=3$ , коэффициент силуэта максимизируется при  $K=2$  (0.532).

- Результаты DBSCAN:

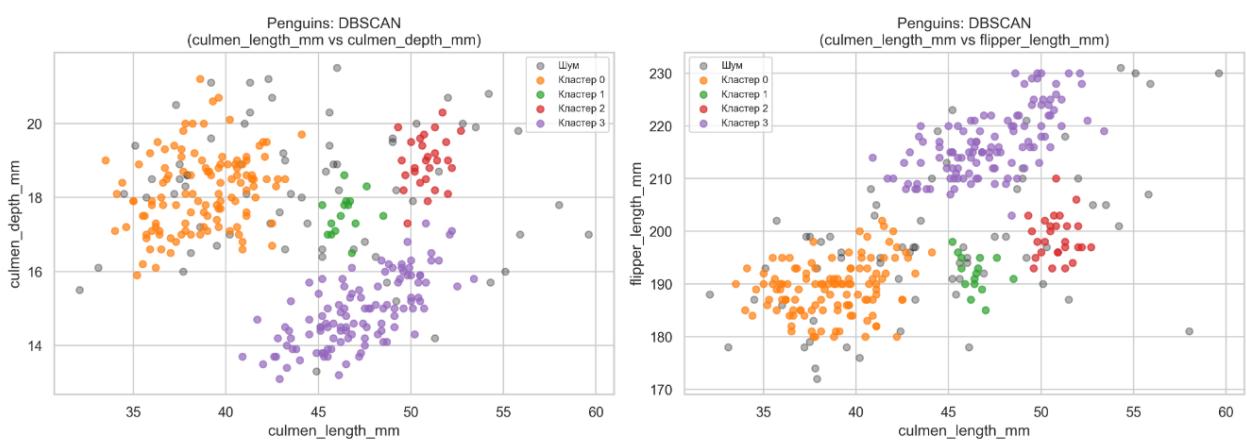


Рисунок 8: Результаты кластеризации DBSCAN на данных penguins. Найдено 4 кластера, 67 точек шума (19.7%), коэффициент силуэта 0.424.

- **Эксперименты DBSCAN:** тепловые карты числа кластеров и процента шума

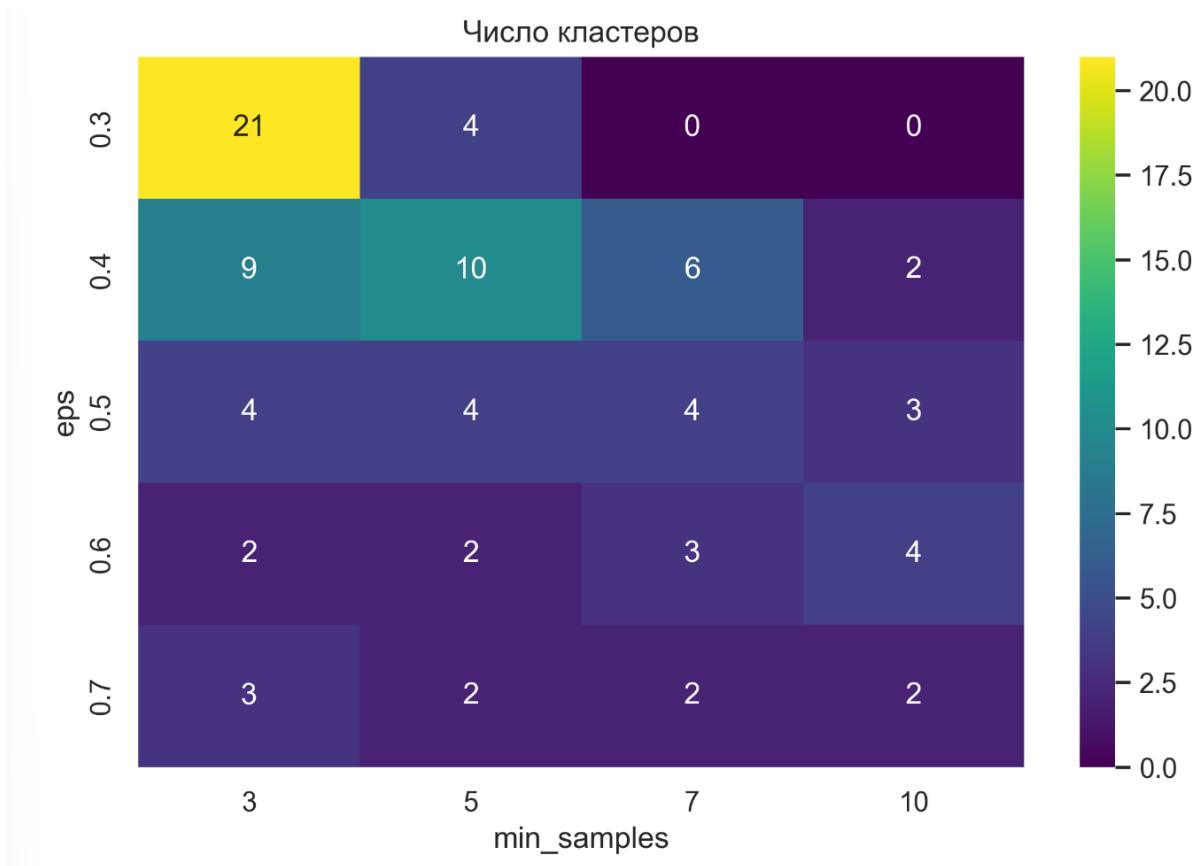


Рисунок 9: Тепловая карта числа кластеров в зависимости от параметров  $\text{eps}$  и  $\text{min\_samples}$  для DBSCAN.

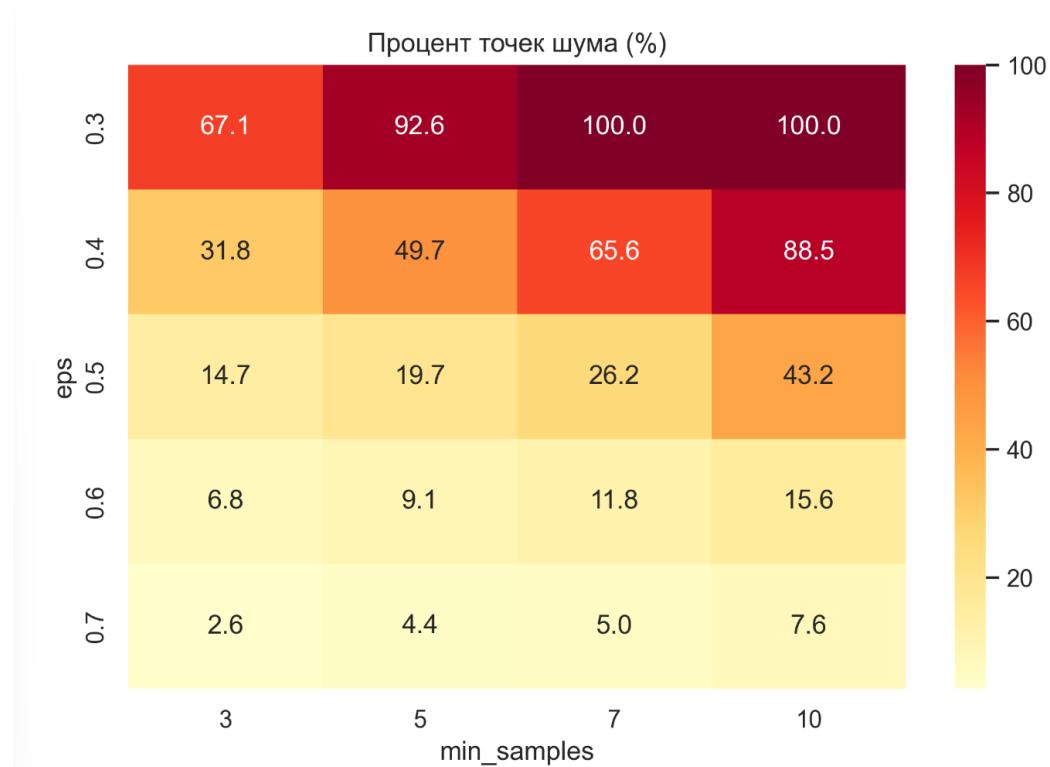


Рисунок 10: Тепловая карта процента точек шума в зависимости от параметров `eps` и `min_samples` для DBSCAN.

---

## Анализ результатов и сравнение эффективности методов

### Синтетические данные

**K-Means:** успешно восстановил 4 кластера (силуэт 0.792), все точки распределены, отличные результаты для сферических кластеров.

**DBSCAN:** нашел 3 кластера вместо 4 (силуэт 0.738), некоторые кластеры объединились, требуется более тщательный подбор параметров.

**Вывод:** K-Means превосходит DBSCAN на синтетических данных со сферическими кластерами.

## Датасет Penguins

Метрика	K-Means	DBSCAN
Число кластеров	3	4
Коэффициент силуэта	0.451	0.424
Инерция	373.93	-
Точки шума	0 (0%)	67 (19.7%)
Соответствие ожидаемому	Да (3 вида)	Нет

**K-Means:** показал лучшие результаты, нашел ожидаемое число кластеров (3 вида), хорошая интерпретируемость.

**DBSCAN:** нашел больше кластеров, высокий процент шума, низкий силуэт, неравномерное распределение.

**Вывод:** K-Means показал лучшие результаты по метрике силуэта и интерпретируемости.

## Общее сравнение методов

### K-Means:

- Простота использования, быстрая работа, все точки в кластерах
- Требует задать K, не находит кластеры произвольной формы, чувствителен к выбросам

### DBSCAN:

- Автоматически определяет число кластеров, находит кластеры произвольной формы, устойчив к выбросам
- Требует подбора параметров, чувствителен к масштабу, может помечать много точек как шум

### Рекомендации:

- **K-Means:** когда известно число кластеров, сферическая структура, нужна скорость
  - **DBSCAN:** когда число кластеров неизвестно, произвольная форма, важно выявить выбросы
-

# Выводы

## 1. На синтетических данных:

- K-Means успешно восстановил истинную структуру (4 кластера) с высоким качеством (силуэт = 0.792)
- DBSCAN показал менее точные результаты (3 кластера вместо 4), требуя более тщательной настройки параметров

## 2. На датасете Penguins:

- K-Means показал лучшие результаты (силуэт = 0.451) и нашел ожидаемое число кластеров (3 вида пингвинов)
- DBSCAN нашел 4 кластера с более низким качеством (силуэт = 0.424) и высоким процентом шума (19.7%)

## 3. Общие выводы:

- Для данных с известной структурой и ожидаемым числом кластеров **предпочтительнее K-Means**
- K-Means лучше подходит для сферических кластеров одинакового размера
- DBSCAN полезен для автоматического определения числа кластеров и выявления выбросов, но требует тщательной настройки
- Оба метода требуют предобработки данных (масштабирование) для корректной работы

## 4. Практические рекомендации:

- При работе с реальными данными рекомендуется использовать оба метода и сравнивать результаты
- Для K-Means важно правильно подобрать число кластеров с помощью метода локтя и коэффициента силуэта
- Для DBSCAN необходимо провести систематический перебор параметров для выбора оптимальных значений
- Масштабирование данных критично для обоих методов, особенно для DBSCAN

---

Ссылка на программу: [https://github.com/343549/new\\_ins\\_analiz\\_data\\_lab2](https://github.com/343549/new_ins_analiz_data_lab2)