МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «ЛИПЕЦКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Институт компьютерных наук
Кафедра автоматизированных систем управления

ЛАБОРАТОРНАЯ РАБОТА №3

«Кластеризация данных»

по МАШИННОМУ ОБУЧЕНИЮ

Студенты М-РИТ-25-1 _____ Красиков И.А.
(подпись, дата) Киселев М.С.

Руководитель Сараев П.В.
профессор _____
(подпись, дата)

Липецк 2025 г.

Цель работы – изучение методов кластеризации данных и определения оптимального количества кластеров.

Задание кафедры

1. Постройте кластеризацию данных для определения вида ирисов (база Iris) на основе четырех параметров (без использования информации о классе). Для кластеризации используйте метод k-средних и иерархический метод Уорда. Исследуйте влияние аргументов процедур, реализующих указанные выше методы, на результат кластеризации. Постройте графики.
2. Определите оптимальное количество кластеров на основе метода «локтя».
3. Сравните результаты кластеризации для случая 3-х кластеров с реальными классами, представленными в базе данных Iris.
4. Сделайте выводы о проделанной работе.

Ход работы

1. Загрузка данных и исключение информации о классе (виде), изображена на рисунке 1.

```
[2]: # Загрузка данных
iris = load_iris()
X = iris.data  # 4 параметра: sepal length, sepal width, petal length, petal width
y = iris.target  # реальные классы
feature_names = iris.feature_names
target_names = iris.target_names

print("Информация о наборе данных Iris:")
print(f"Размерность данных: {X.shape}")
print(f"Параметры: {feature_names}")
print(f"Классы: {target_names}")
```

```
Информация о наборе данных Iris:
Размерность данных: (150, 4)
Параметры: ['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)']
Классы: ['setosa' 'versicolor' 'virginica']
```

Рисунок 1 – Загрузка данных и исключение информации о классе

2. Кластеризация методом k-средних. Возьмем 3 кластера т.к. известно, что 3 вида Ирисов, видно на рисунке 2.

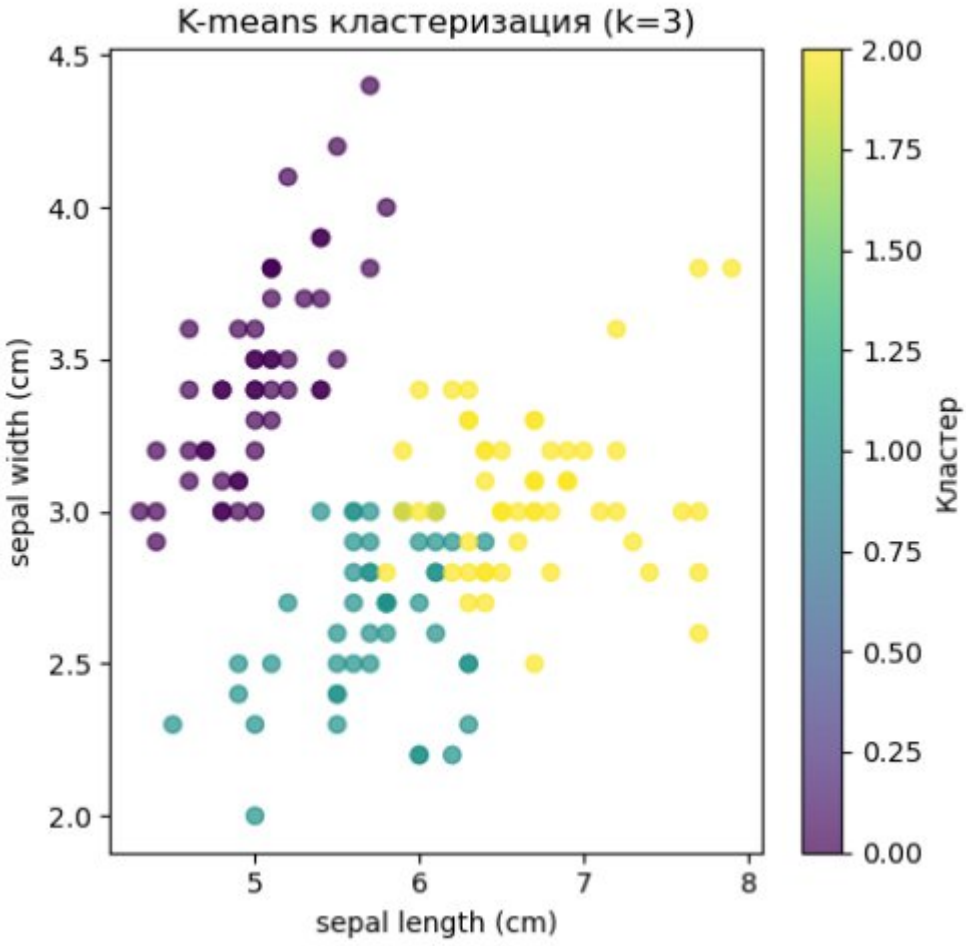

Рисунок 2 – Кластеризация методом k-средних для 3 кластеров.

3. Метод локтя, изображена на рисунке 3. Из графика локтя видно, что изгиб примерно на k = 3 – это оптимальное количество кластеров.

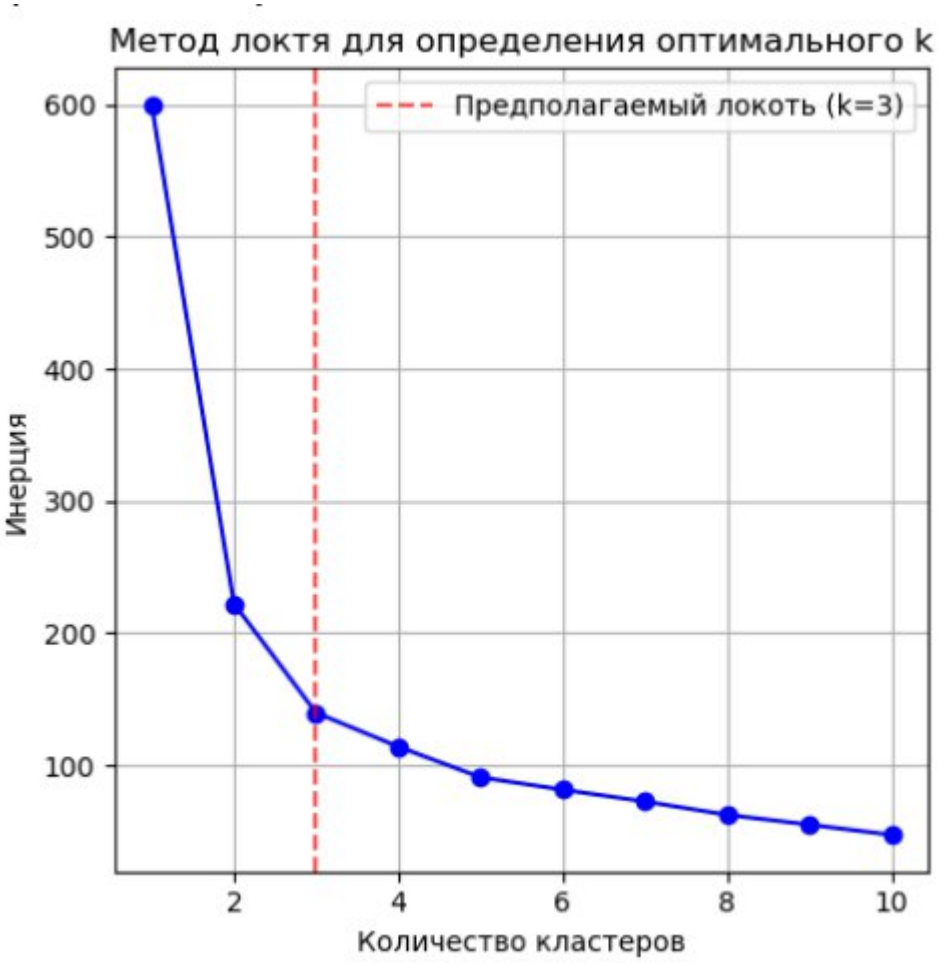

Рисунок 3 – Кластеризация методом Уорда

4. Иерархический метод Уорда изображен на рисунке 4.

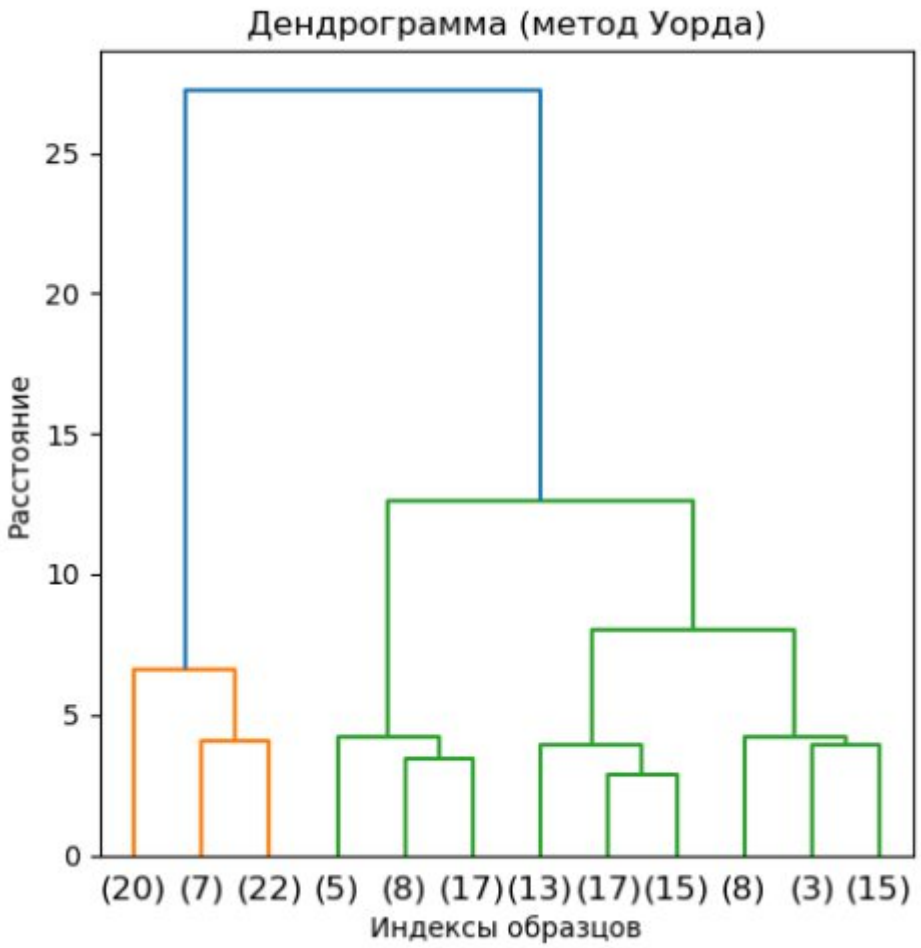

Рисунок 4 – Иерархическая кластеризация методом Уорда

5. Другие иерархические методы 5.

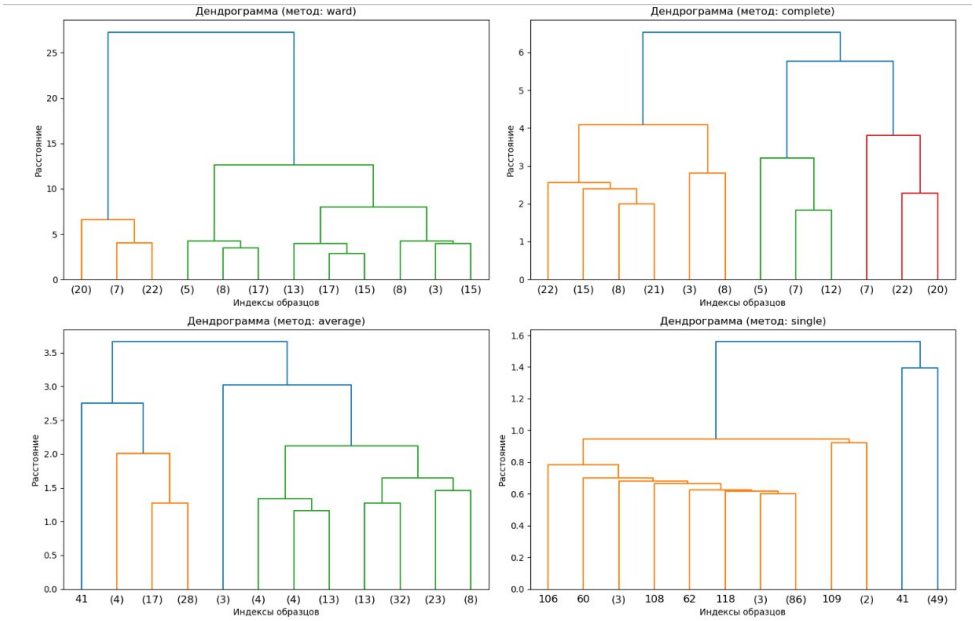

Рисунок 5 – Иерархические методы кластеризации

6. Сравнение методов кластеризация изображено на рисунках 6 и 7.

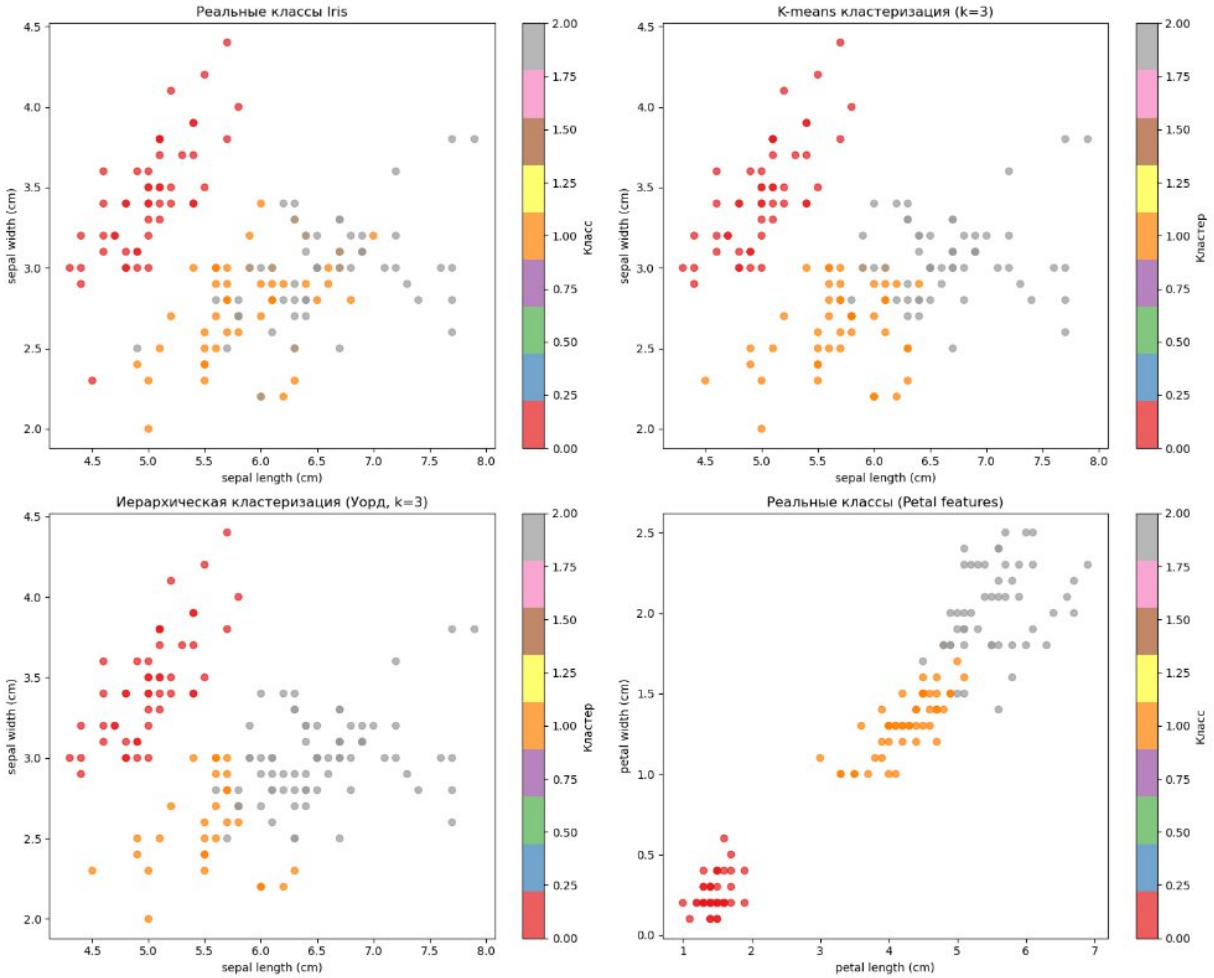

Рисунок 6 – Сравнение методов кластеризации

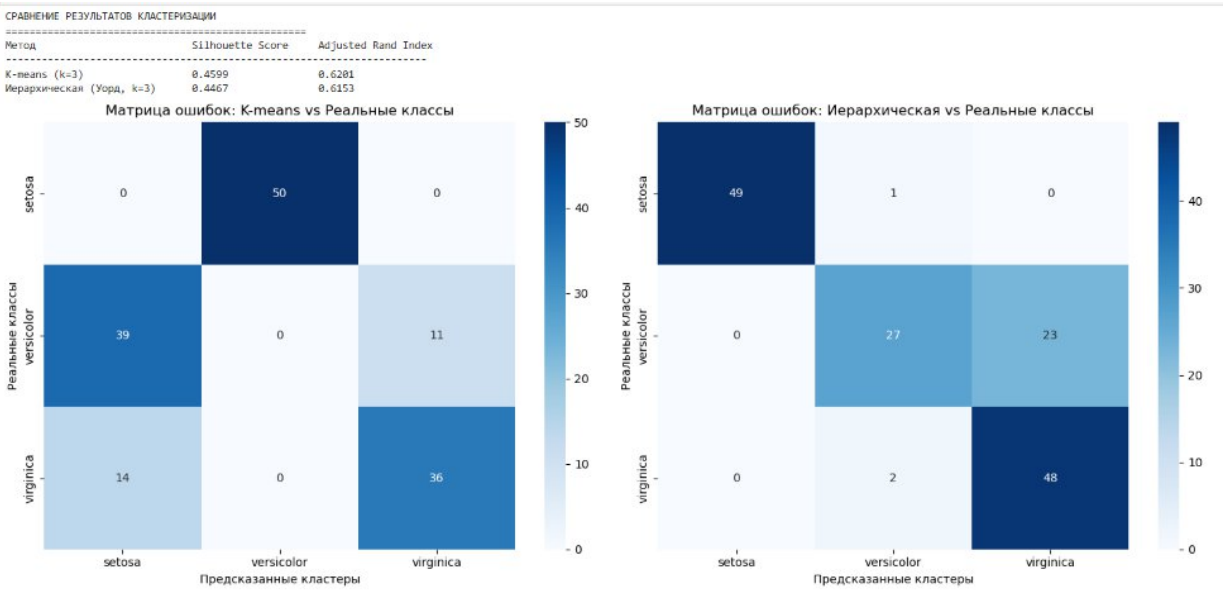

Рисунок 7 – Сравнение методов кластеризации (Матрицы ошибок)

8. Код программы

Репозиторий на github: https://github.com/MicroMolekula/machine-learning/tree/main/lab3

Вывод

Оптимальное количество кластеров (по методу локтя) – k = 3. Метод k-средних хорошо разделил данные, оценка кластеризации примерно 0.4599. Иерархическая кластеризация (Ward) даёт похожий результат, оценка кластеризации примерно 0.4467. В целом результаты стабильно показывают 3 естественных кластера ирисов: Setosa, Versicolor и Virginica.