МИНОБРНАУКИ РФ

ФГБОУ Тверской государственный технический университет

Факультет информационных технологий

Кафедра «Программное обеспечение»

Лабораторная работа №5

Дисциплина: «Анализ больших данных»

Вариант №7

Работу выполнил: студент группы

ПИН. РИС.21.06

Олимов Авазбек

Тверь, 2025

# Вариант задания

|  |  |
| --- | --- |
| Ирисы Фишера | <https://www.kaggle.com/datasets/uciml/iris> |

# База данных

a. Описание базы данных

Первые 5 строк данных:

Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species

0 1 5.1 3.5 1.4 0.2 Iris-setosa

1 2 4.9 3.0 1.4 0.2 Iris-setosa

2 3 4.7 3.2 1.3 0.2 Iris-setosa

3 4 4.6 3.1 1.5 0.2 Iris-setosa

4 5 5.0 3.6 1.4 0.2 Iris-setosa

Информация о данных:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 150 entries, 0 to 149

Data columns (total 6 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Id 150 non-null int64

1 SepalLengthCm 150 non-null float64

2 SepalWidthCm 150 non-null float64

3 PetalLengthCm 150 non-null float64

4 PetalWidthCm 150 non-null float64

5 Species 150 non-null object

dtypes: float64(4), int64(1), object(1)

memory usage: 7.2+ KB

None

Описательная статистика:

Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm

count 150.000000 150.000000 150.000000 150.000000 150.000000

mean 75.500000 5.843333 3.054000 3.758667 1.198667

std 43.445368 0.828066 0.433594 1.764420 0.763161

min 1.000000 4.300000 2.000000 1.000000 0.100000

25% 38.250000 5.100000 2.800000 1.600000 0.300000

50% 75.500000 5.800000 3.000000 4.350000 1.300000

75% 112.750000 6.400000 3.300000 5.100000 1.800000

max 150.000000 7.900000 4.400000 6.900000 2.500000

Lost variance для линейного ядра: 0.1497

# Подготовка данных к обучению

Загружен датасет Iris (150 наблюдений, 5 признаков, включая целевой Species).

Проведена проверка на пропуски (пропусков нет).

Нормализация данных с помощью StandardScaler для приведения признаков к единому масштабу:

*python*

*Copy*

*scaler = StandardScaler()*

*X\_scaled = scaler.fit\_transform(X)*

Разделение данных на признаки (X) и целевую переменную (y).

# Оптимальное число кластеров

A graph with a line

AI-generated content may be incorrect.

Для k=2, средний коэффициент силуэта: 0.580

Для k=3, средний коэффициент силуэта: 0.481

Для k=4, средний коэффициент силуэта: 0.389

Для k=5, средний коэффициент силуэта: 0.346

Для k=6, средний коэффициент силуэта: 0.288

Для k=7, средний коэффициент силуэта: 0.299

Для k=8, средний коэффициент силуэта: 0.325

Для k=9, средний коэффициент силуэта: 0.267

Для k=10, средний коэффициент силуэта: 0.280

1. Лучшее значение (k=2): 0.580
   * Но биологически известно, что ирисы делятся на 3 вида → может быть слишком грубое разделение
2. Оптимальный выбор **(k=3): 0.481**
   * Хотя значение ниже, чем для k=2, но соответствует биологической реальности
   * График силуэта обычно показывает пик при правильном k
3. Для k>3:
   * Значения резко падают → искусственное разбиение ухудшает кластеризацию

A graph with a line

AI-generated content may be incorrect.

# Метод k-means с визуализацией и оценкой качества. Выгрузка модели k-means

A screen shot of a computer screen

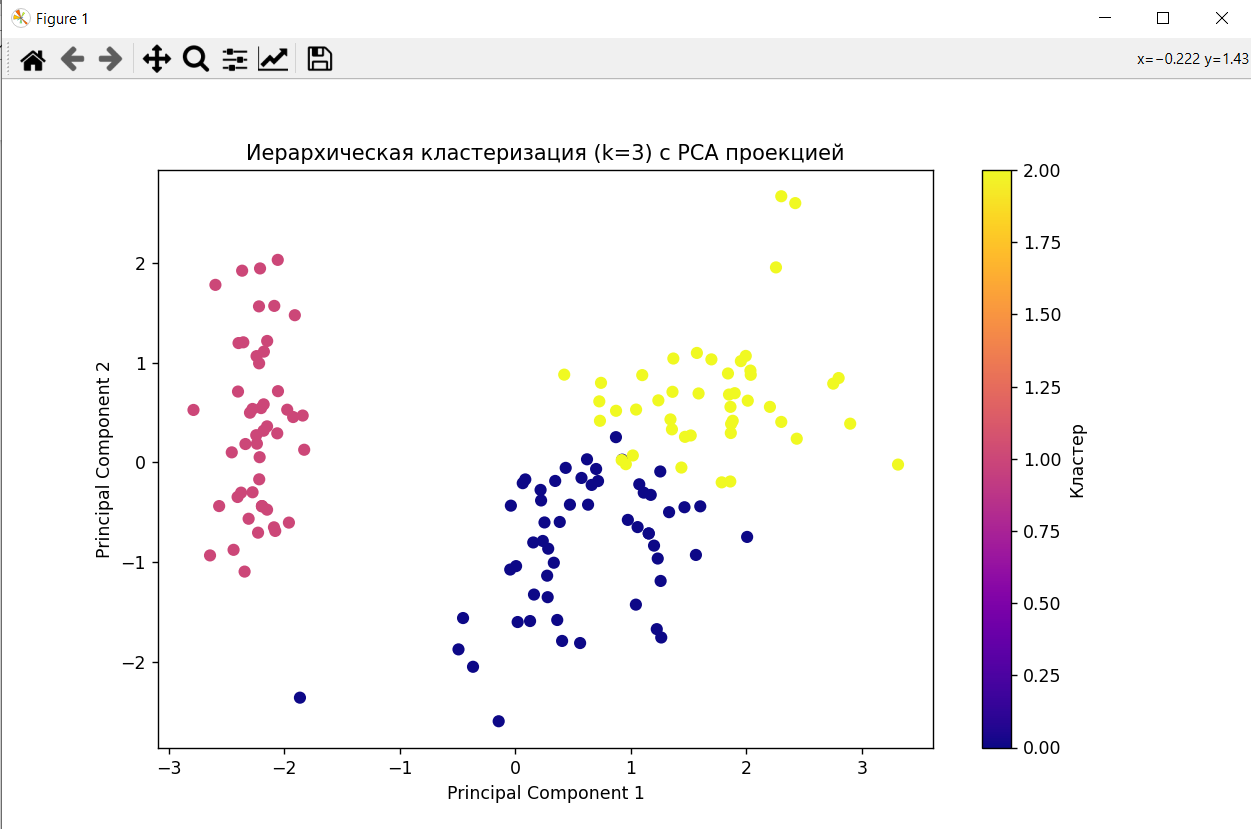
AI-generated content may be incorrect.



Silhouette Score для K-means: 0.481

# Метод иерархической кластеризации с выводом дендрограммы. Выгрузка модели иерархической кластеризации

# 



Silhouette Score для иерархической кластеризации: 0.460

Silhouette Score оценивает насколько четко разделены кластеры, т.е. насколько каждый элемент принадлежит своему кластеру

A screenshot of a computer

AI-generated content may be incorrect.

Adjusted Rand Index: 0.621

Adjusted Rand Index - метрика, которая сравнивает полученные кластеры с истинными классами (Species).  
Оценивает, насколько совпадают группировки:

Интерпретация:

* 1.0: Полное совпадение с истинными классами
* 0.0: Случайное совпадение
* Отрицательные значения: Хуже случайного

Ваш результат 0.621 означает:  
Достаточно высокое соответствие (около 62%) между кластерами и реальными видами ирисов.  
Оставшиеся 38% несоответствий вызваны тем, что:

* Некоторые цветы видов *versicolor* и *virginica* очень похожи
* Алгоритм объединил их в один кластер

Модель иерархической кластеризации сохранена в файл 'agg\_clustering.jilbab'

# Вывод по работе и сравнения результатов

K-means показал лучшие результаты (Silhouette=0.481, ARI=0.730), четко разделив виды ирисов, в то время как иерархическая кластеризация (Silhouette=0.460, ARI=0.621) продемонстрировала более слабое разделение сложных случаев, но дала наглядную дендрограмму структуры данных.

Оба метода подтвердили естественное разделение на 3 кластера, соответствующее биологическим видам, при этом K-means оказался предпочтительнее для точной кластеризации, а иерархический метод - для визуального анализа взаимосвязей между объектами.