|  |
| --- |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение высшего образования **«МИРЭА − Российский технологический университет»**  **РТУ МИРЭА** |

**Институт информационных технологий (ИИТ)**

**Кафедра прикладной математики (ПМ)**

**ОТЧЕТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ №8**

по дисциплине «Языки программирования для статистической обработки данных»

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Студент группы ИНБО-01-17 | *ИМБО-11-23, Журавлев Ф.А.* | (подпись) | |  |
| Преподаватель | *Трушин СМ* | (подпись) | |  |
|  |  | |  | |

Москва 2025 г.

## ЦЕЛЬ И ЗАДАЧИ

**Цель практической работы:**

Научиться проводить кластеризацию данных с использованием методов K-means и иерархической кластеризации в Python, R, а также визуализировать результаты кластерного анализа.

**Задачи практической работы:**

1. Выполнить кластеризацию методом K-means:

• Python: использование библиотеки sklearn.

• R: функция kmeans.

2. Провести иерархическую кластеризацию:

• Python: библиотека scipy.cluster.hierarchy.

• R: функции hclust, dendrogram.

3. Проанализировать результаты кластеризации:

• Интерпретация кластеров (центроиды, количество объектов в каждом кластере).

• Сравнение кластеров, полученных разными методами.

4. Визуализировать результаты кластеризации:

• Python: графики кластеров с использованием matplotlib и

seaborn.

* R: графическое представление дендрограмм и кластеров.

5. Сравнить удобство выполнения кластерного анализа в Python, R.

## РЕЗУЛЬТАТЫ ПРАКТИКИ

**Шаг 1) Кластеризация в Python**

### 1.1) K-means кластеризации в Python.

Сначала загрузим наши данные в питон, отберем только числовые признаки а также масштабируем данные:

*Рисунок 1.1 — Загрузка данных и обработка.*

Изображение выглядит как текст, снимок экрана, Шрифт

Контент, сгенерированный ИИ, может содержать ошибки.

После загрузки исходной таблицы данных в формате .csv, следует провести кластеризацию двумя способами: K-means и иерархическую.

Напишем код, который реализует кластеризацию K-means и иерархическую, перед этим убедившись, что были добавлены все необходимые библиотеки:

*Рисунок 1.2 — K-means кластеризация.*

Изображение выглядит как текст, снимок экрана, программное обеспечение, Шрифт

Контент, сгенерированный ИИ, может содержать ошибки.

Все результаты и итоги подведем в параграфе 3 «Сравнение результатов», а далее напишем код иерархической кластеризации.

Далее напишем код и рассмотрим график, который получился в результате K-means кластеризации указав, что количество кластеров равняется трем. Изображение выглядит как текст, снимок экрана, диаграмма, График

Контент, сгенерированный ИИ, может содержать ошибки.

*Рисунок 1.3 — Код кластеризации K-means.*

### 1.2) Иерархическая кластеризация в Python.

*Изображение выглядит как текст, снимок экрана, компьютер, программное обеспечение

Контент, сгенерированный ИИ, может содержать ошибки.*

*Рисунок 1.4 — Код иерархической кластеризации.*

### 1.3) Визуализация иерархической кластеризация с PCA в Python.

Изображение выглядит как текст, снимок экрана, диаграмма, График

Контент, сгенерированный ИИ, может содержать ошибки.

### 1.4) Анализ кластеров по количеству и средним в Python.

**Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Контент, сгенерированный ИИ, может содержать ошибки.**

Что значит каждый график будет расписано на 3 шаге

**Шаг 2) Кластеризация в Rstudio.**

**2.1) Иерархическая и K-means кластеризация в Rstuido.**

Далее проделаем все тоже самое, но уже с помощью языка программирования R, убедившись, что все необходимые пакеты были успешно установлены.

Изображение выглядит как текст, снимок экрана, Шрифт, веб-страница

Контент, сгенерированный ИИ, может содержать ошибки.

*Рисунок 2.1 – код кластеризации K-means в R.*

**Изображение выглядит как текст, диаграмма, число, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.**

*Рисунок 2.2 – график кластеризации K- means в R.*

Построим иерархическую кластеризацию:

**Изображение выглядит как текст, Шрифт, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.**

*Рисунок 2.3 – код иерархической кластеризации в R.*

И посмотрим на график:

Изображение выглядит как диаграмма, линия, снимок экрана, дизайн

Контент, сгенерированный ИИ, может содержать ошибки.

*Рисунок 2.4 – График иерархической кластеризации в R.*

Проведем визуализацию иерархических кластеров:

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Контент, сгенерированный ИИ, может содержать ошибки.

*Рисунок 2.5 – код визуализации иерархическиз кластеров в R.*

*Изображение выглядит как текст, снимок экрана, диаграмма, карта

Контент, сгенерированный ИИ, может содержать ошибки.*

*Рисунок 2.6 – График визуализации иерархических кластеров в R.*

Проведем анализ кластеров:

Изображение выглядит как текст, программное обеспечение, Шрифт, число

Контент, сгенерированный ИИ, может содержать ошибки.

*Рисунок 2.7 – Анализ кластеров.*

**3 СРАВНЕНИЕ РЕЗУЛЬТАТОВ**

В ходе практической работы были применены два метода кластеризации: **K-means** и **иерархическая кластеризация**. Анализ проводился в средах Python и R. Ниже представлены основные различия и особенности каждого подхода, а также описание всех графиков, полученных в ходе работы.

### Сравнение методов

* **K-means** — это метод, при котором изначально задаётся количество кластеров (например, 3), и алгоритм ищет оптимальное разбиение данных на группы, минимизируя внутрикластерные расстояния. Он работает быстро и хорошо подходит для больших объёмов данных.
* **Иерархическая кластеризация** строит древовидную структуру объединения объектов, начиная с отдельных точек. В отличие от K-means, количество кластеров можно выбрать позже, визуально — на основе дендрограммы.

1. **Интерпретация графиков**
2. **Диаграмма кластеров K-means на основе PCA**

Этот график показывает, как объекты (наблюдения) были разбиты на кластеры методом K-means. Каждая точка — это один объект из выборки. Цвет отражает принадлежность к определённому кластеру. Использование PCA позволяет изобразить многомерные данные на плоскости в виде двух главных компонент, сохранив основную структуру. Если кластеры визуально хорошо разделяются, это означает, что алгоритм нашёл естественные группы в данных.

1. **Дендрограмма (иерархическая кластеризация)**

Этот график показывает процесс последовательного объединения объектов в кластеры. По оси Y откладывается расстояние между объединяемыми группами. Чем выше соединяются ветви — тем менее похожи объединяемые группы. Чтобы выбрать, например, 3 кластера, можно мысленно провести горизонтальную линию, пересекающую три ветви. Дендрограмма удобна тем, что позволяет увидеть структуру данных даже до выбора количества кластеров.

1. **Диаграмма кластеров после иерархической кластеризации (PCA)**

Аналогична графику K-means, но отражает результат иерархического подхода. Отображает, как алгоритм разделил данные на кластеры после «разреза» дендрограммы. График позволяет сравнить полученные группы с теми, что дал K-means: совпадают ли кластеры, есть ли перемешанные точки и т.д.

1. **Таблица средних значений по кластерам**

Эта таблица содержит усреднённые значения всех признаков по каждому кластеру. Она позволяет понять, чем именно различаются группы: например, один кластер может объединять пациентов с высоким пульсом, другой — с повышенным давлением. Такая интерпретация делает результат кластеризации понятным с точки зрения предметной области (в данном случае — медицина).

### Общий вывод

Оба метода — K-means и иерархическая кластеризация — выделили группы в данных. При этом визуализация показала, что результаты кластеризации в целом согласуются: группы перекрываются частично, но в целом отражают структуру данных.

Метод K-means удобен при чётко заданном количестве кластеров, тогда как иерархическая кластеризация более гибкая и даёт лучшее представление о вложенности и близости объектов. Оба подхода успешно применимы к медицинским данным, подобным heart\_cleaned.csv.

**ВЫВОДЫ**

В результате выполнения практической работы мы ознакомились с методами кластеризации данных, включая **K-means** и **иерархическую кластеризацию**. Были изучены основные этапы применения этих алгоритмов, выполнено масштабирование данных, визуализация кластеров и интерпретация полученных результатов. Также была проведена работа в двух средах программирования — **Python и R**, что позволило сравнить их удобство и функциональные возможности при анализе данных.

В завершение были сделаны общие выводы о применимости кластерного анализа к медицинским данным и его роли в выявлении скрытых групп и закономерностей в выборке.