Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования «Кубанский государственный университет»

Кафедра вычислительных технологий

**ОТЧЕТ**

о выполнении лабораторной работы № 5

по дисциплине «Обработка больших данных»

Тема:задачи классификации и кластеризации

.

Выполнил: ст. гр. 36/2

Агаджанян А. С.

Проверил: преп. кафедры ВТ

Яхонтов А. А.

Краснодар

2025

Набор данных при изначальном прочтении выглядел как показано на рисунке 1.

Изображение выглядит как текст, снимок экрана, меню, число

Контент, сгенерированный ИИ, может содержать ошибки.

Рис.1 – изначальный набор данных

Далее была проведена нормализация (масштабирование) и набор данных стал выглядеть так, как показано на рисунке 2.

Изображение выглядит как текст, снимок экрана, число, меню

Контент, сгенерированный ИИ, может содержать ошибки.

Рис.2 – измененный набор данных

Был проведен дескриптивный анализ датасета, результаты которого видны на рисунке 3 и 4.

Изображение выглядит как текст, Шрифт, число, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Рис.3 Дескриптиынй анализ (1 часть)

Изображение выглядит как текст, снимок экрана, Шрифт

Контент, сгенерированный ИИ, может содержать ошибки.

Рис.4 Дескриптивный анализ (2 часть)

Для выявления оптимального количества кластеров, на которые надо разбить датасет, были применены следующие методы: “метод локтя” (рисунок 5), ”метод силуэта” (рисунок 6), ”метод статистики разрыва” (рисунок 7), ”алгоритм на основе консенсуса” (рисунок 8).

Изображение выглядит как текст, линия, График, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.

Рис.5 – метод локтя

Метод локтя показывает, что оптимальное число кластеров – 4, потому что дальше график перестает существенно уменьшаться.

Изображение выглядит как текст, линия, диаграмма, График

Контент, сгенерированный ИИ, может содержать ошибки.

Рис. 6 – метод силуэта

Метод силуэта показывает, что оптимальное число кластеров – 2.

Изображение выглядит как текст, диаграмма, линия, График

Контент, сгенерированный ИИ, может содержать ошибки.

Рис.7 – метод статистики разрыва

Метод статистики разрыва показывает, что оптимальное число кластеров – 5.

Изображение выглядит как текст, снимок экрана, диаграмма, График

Контент, сгенерированный ИИ, может содержать ошибки.

Рис.8 – алгоритм на основе консенсуса

Алгоритм на основе консенсуса показывает, что оптимальное число кластеров – 5.

Разобьем набор данных на 5 кластеров и нарисуем дендрограмму (рис. 9)

Изображение выглядит как текст, диаграмма, Технический чертеж, План

Контент, сгенерированный ИИ, может содержать ошибки.

Рис. 9 – дендрограмма

Изображение выглядит как текст, снимок экрана, диаграмма, Шрифт

Контент, сгенерированный ИИ, может содержать ошибки.

Рис. 10 – разделили выборку на 5 кластеров

Выполним класстеризацию по k-means:

Изображение выглядит как текст, снимок экрана, диаграмма, График

Контент, сгенерированный ИИ, может содержать ошибки.

Рис. 11 – 5 групп кластеров с легендой

Изображение выглядит как текст, диаграмма, График, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Рис. 12 – изменена палитра легенды и эллипсоидное очертание

Скаттерплоты, отражающие характеристики классов:

Изображение выглядит как текст, диаграмма, чек

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, диаграмма, снимок экрана, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Построение трехмерной кластеризации по scatterplot3d (были выбраны 3 характеристики из всей выборки – это “education”, “tradeshare”, “rgdp60”):

Изображение выглядит как диаграмма, линия, График, Параллельный

Контент, сгенерированный ИИ, может содержать ошибки.

**Часть 2. Байесовская классификация и деревья принятия решений на R**

Добавим фактор-столбец group с метками кластеров из части 1 в датасет. Посмотрим центры кластеров:

Изображение выглядит как текст, Шрифт, снимок экрана

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, снимок экрана, диаграмма, Параллельный

Контент, сгенерированный ИИ, может содержать ошибки. Изображение выглядит как текст, снимок экрана, диаграмма, Параллельный

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, диаграмма, кроссворд

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, диаграмма, снимок экрана, План

Контент, сгенерированный ИИ, может содержать ошибки.

Составим матрицу ошибок (confusion matrix):

Изображение выглядит как текст, снимок экрана, Шрифт, число

Контент, сгенерированный ИИ, может содержать ошибки.

Всего ошибок – 0. Точность классификации: 100%.

Далее проведем классификацию методом Decision Tree. Перед этим разобьём датасет на обучающую и тестовую выборку с соотношением 70% и 30% соответственно.

Изображение выглядит как текст, Шрифт, снимок экрана, Цвет электрик

Контент, сгенерированный ИИ, может содержать ошибки.

Матрица ошибок для **обучающей** выборки:

Изображение выглядит как текст, снимок экрана, Шрифт, дизайн

Контент, сгенерированный ИИ, может содержать ошибки.

Матрица ошибок для **тестовой** выборки:

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, диаграмма, План, Технический чертеж

Контент, сгенерированный ИИ, может содержать ошибки.

Используем алгоритм Random Forest (случайный лес). Построим лес из 100 деревьев.

Изображение выглядит как снимок экрана, Шрифт, текст, число

Контент, сгенерированный ИИ, может содержать ошибки.

У алгоритма 6 ошибок.

Вывод: были закреплены знания об алгоритмах классификации и кластеризации данных, ознакомлен с некоторыми функциями языка R, осуществляющими этот вид анализа. Научился визуализировать результаты работы функций кластерного анализа и классификаторов, интерпретировать полученные результаты и выполнять классификацию на основе формулы Байеса и деревьев решений.