Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Факультет компьютерных технологий и прикладной математики**

**Кафедра вычислительных технологий**

**ЛАБОРАТОРНАЯ РАБОТА №6.1**

**Дисциплина: Обработка больших данных**

Работу выполнил: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_А.Д. Воробьев

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Т.А. Приходько

**Цель работы:** закрепить знания об алгоритмах классификации и кластеризации данных, ознакомиться с некоторыми функциями языка R, осуществляющими этот вид анализа, принципами их работы. Научиться визуализировать результаты работы функций кластерного анализа и классификаторов, интерпретировать полученные результаты. Научиться выполнять классификацию на основе формулы Байеса и деревьев решений. Также провести сравнение результатов кластеризации двух датасетов по одной тематике за различные периоды.

**Часть 1.** **Подготовка данных**

В качестве датасета была выбрана статистика по покупателям

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

Рисунок 1 – Фрагмент исходной таблицы с данными

Далее был проведён стандартный описательный анализ для данных

Изображение выглядит как текст, снимок экрана, Шрифт, меню

Автоматически созданное описание

Рисунок 2 – Описательный анализ сырых данных

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рисунок 3 – Описательный анализ подготовленных данных

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

Автоматически созданное описаниеРисунок 4 – Датасет с нормализованными данными

Теперь данные приведены в форму, которую пригодно использовать для дальнейшей кластеризации и сравнения.

**Часть 2.** **Кластеризация**

Далее различными способами определим наилучшее количество кластеров

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Рисунок 5 – Метод среднего локтя

Изображение выглядит как текст, снимок экрана, дисплей, Прямоугольник

Автоматически созданное описание

Рисунок 6 – Диаграмма «Каменная осыпь» для 24 года.

Изображение выглядит как текст, линия, График, снимок экрана

Автоматически созданное описание

Рисунок 7 – Метод среднего силуэта

Теперь применим полученный коэффициент схожести для отделения классов друг-от друга на дендрограмме

Изображение выглядит как рисунок, зарисовка, дизайн

Автоматически созданное описание

Рисунок 8 – Дендрограмма

Далее представим распределения характеристик по всем кластерам.

Изображение выглядит как снимок экрана, Красочность, текст, Прямоугольник

Автоматически созданное описание

Рисунок 13 – Распределение характеристик

Исходя из стековой диаграммы сразу видится разделение по возрасту и расходам в магазине

Изображение выглядит как текст, снимок экрана, Красочность, диаграмма

Автоматически созданное описание Изображение выглядит как текст, снимок экрана, Красочность, Прямоугольник

Автоматически созданное описаниеРисунок 14 – Распределение характеристик по группам по возрасту и доходу

Исходя из распределения параметров по группам видно, что алгоритм диверсифицирует людей по возрасту, а по уровню доходов не сильно

Изображение выглядит как снимок экрана, текст, Красочность, диаграмма

Автоматически созданное описание Изображение выглядит как текст, снимок экрана, диаграмма, Красочность

Автоматически созданное описаниеРисунок 15 – Распределение характеристик по группам по расходам и опыту работы

Исходя из распределения параметров по группам видно, что алгоритм диверсифицирует людей по расходам, и по опыту работы

Изображение выглядит как текст, снимок экрана, Красочность, диаграмма

Автоматически созданное описание Рисунок 16 – Распределение характеристик по размеру семьи

По этому признаку алгоритм тоже отделяет группы

Изображение выглядит как снимок экрана, текст, линия, Прямоугольник

Автоматически созданное описание

Рисунок 20 – Диаграмма рассеивания по рабочему опыту и доходам. Очевидно, что алгоритм определяет людей с высоким общим доходом и малым опытом работы в отдельный класс

Изображение выглядит как текст, снимок экрана, линия, диаграмма

Автоматически созданное описание

Рисунок 21 – Диаграмма соотношения расходов к размеру семьи

Очевидно, что модель довольно чётко классифицирует по соотношению расходов и размеру семьи

Изображение выглядит как снимок экрана, текст, диаграмма, линия

Автоматически созданное описание

Рисунок 22 – Диаграмма соотношения расходов к опыту работы

Очевидно, что модель довольно чётко классифицирует по соотношению расходов и опыту работы

Изображение выглядит как снимок экрана, Красочность, текст

Автоматически созданное описание

Рисунок 23 – Диаграмма соотношения расходов к заработку. Очевидно, что чем больше человек зарабатывает, тем больше он тратит.

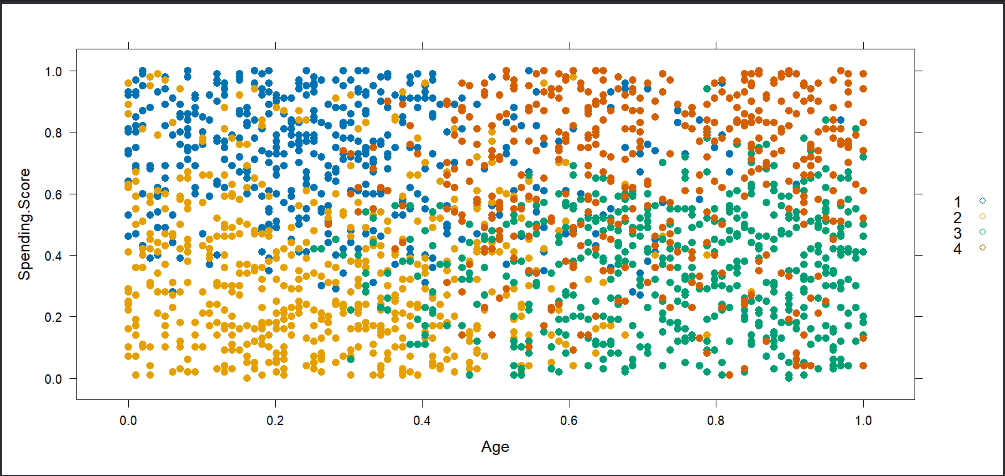


Рисунок 24 – Диаграмма соотношения расходов к возрасту. Явной зависимости нет, но алгоритм чётко разделяет на молодых и не тратящих, старых не тратящих, молодых тратящих и старых не тратящих.

Изображение выглядит как диаграмма, карта

Автоматически созданное описание

Рисунок 26 – 3Д график, показывающий распределение дохода, возраста и опыта работы

Изображение выглядит как снимок экрана, Красочность, диаграмма, дизайн

Автоматически созданное описание

Рисунок 27 – 3Д график, показывающий распределение расходов, доходов и размера семьи

Из графика виден наибольший расход среди 3 и 4 группы (Более возрастные и с большим опытом работы)

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