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**

**Дисциплина: Обработка больших данных**

Работу выполнил: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_К.В.Стасюк

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ В.И.Шиян

**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

Вариант 13 - GooglePlay

**Ход работы:**

**Часть 1.** **Задачи классификации и кластеризации**

Из исходного csv файла данные были импортированы в RStduio. Ниже представлен фрагмент таблицы, всего в ней содержится 10841 строк и 13 столбцов.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 1 – Фрагмент исходной таблицы с данными

После удаления одинаковых строк в таблице их осталось 10358.

После чего с использованием значений из таблицы был проделан некоторый анализ:

Изображение выглядит как диаграмма, круговая диаграмма

Автоматически созданное описание

Рисунок 2 – Круговая диаграмма

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 3 – Столбчатая диаграмма

Для анализа нам необходимо оставить столбцы Rating, Reviews, Size и Installs, а также заменить некорректные значения и удалить строки со значением NA.После чего осталось 7424 строки.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 4 – Удалены пропущенные значения.

В данной задаче переменные существенно различны, поэтому необходимо выполнить нормализацию данных. Для стандартизации исключим 1 колонку.

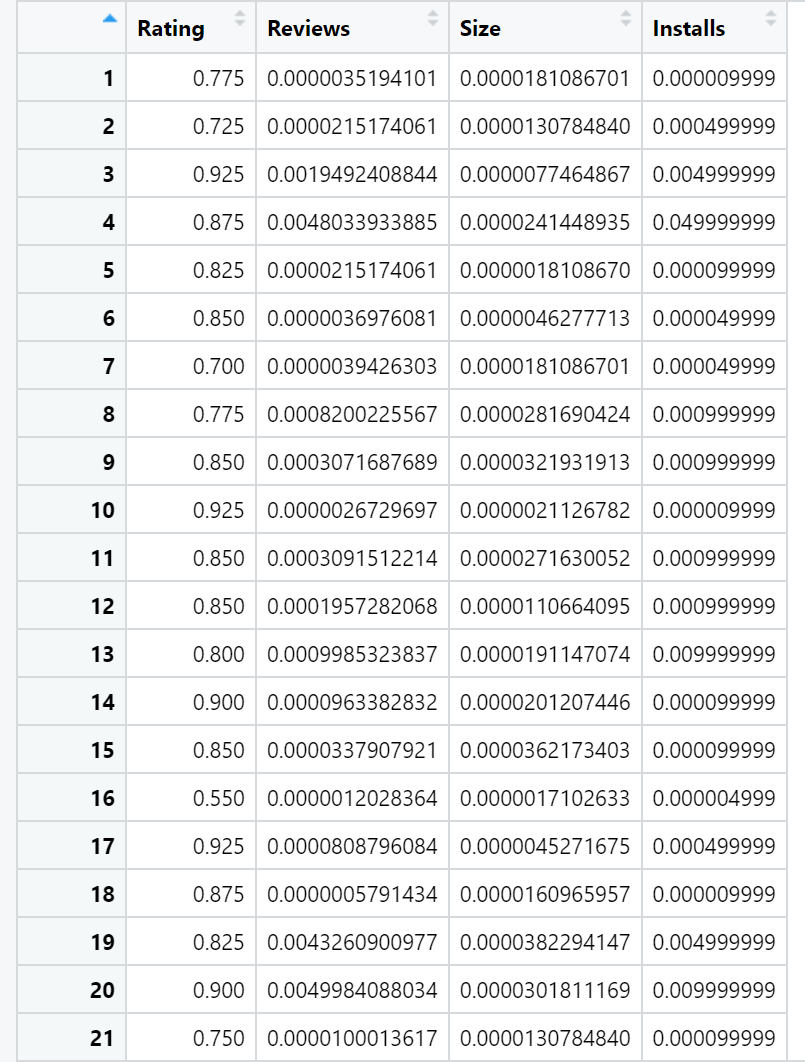


Рисунок 5 – Фрагмент стандартизированной таблицы с данными

По полученным данным построим дендрограмму.

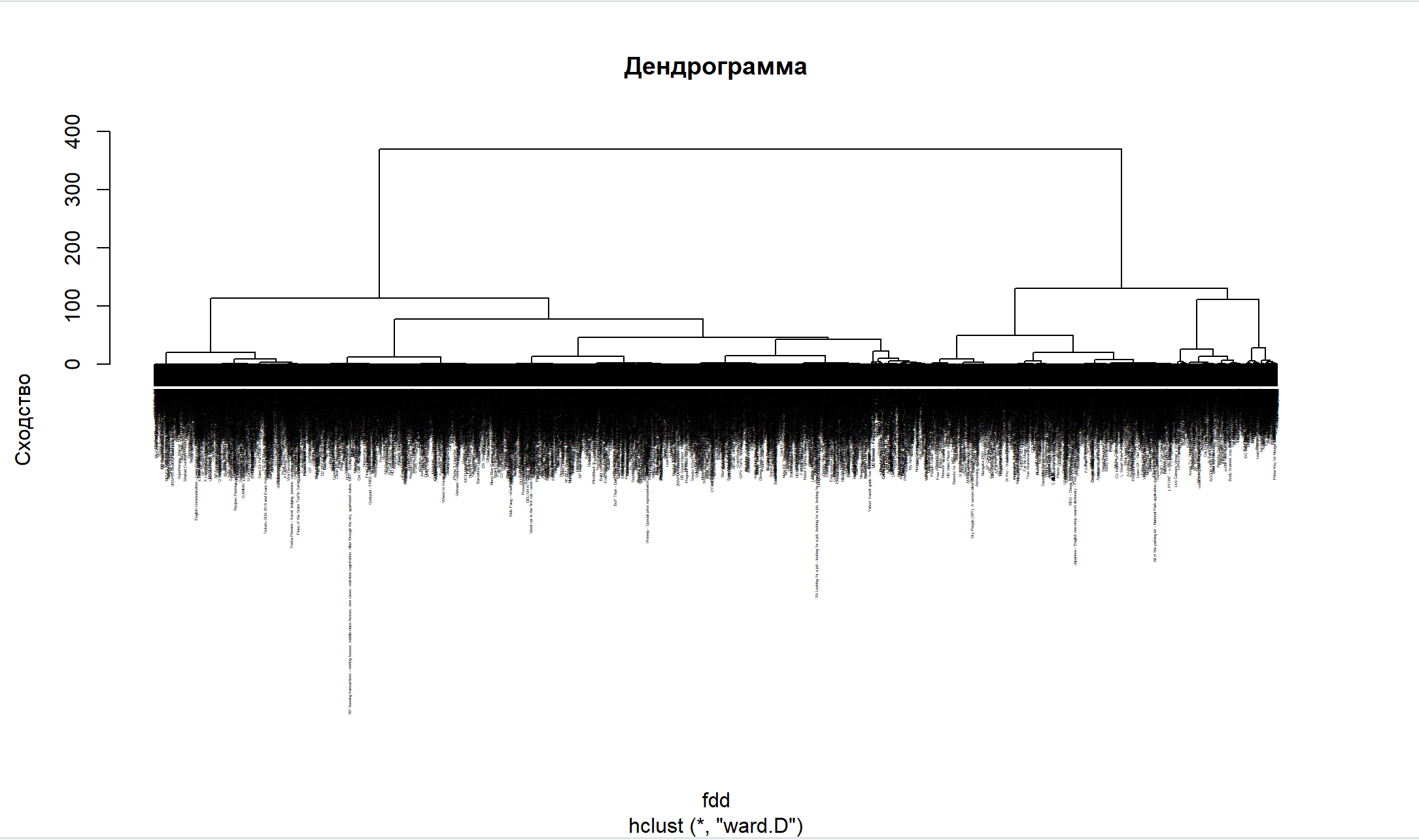


Рисунок 6 – Дендограмма, построенная по данным

Целесообразнее выделить 4 кластера, выделим их красным цветом.

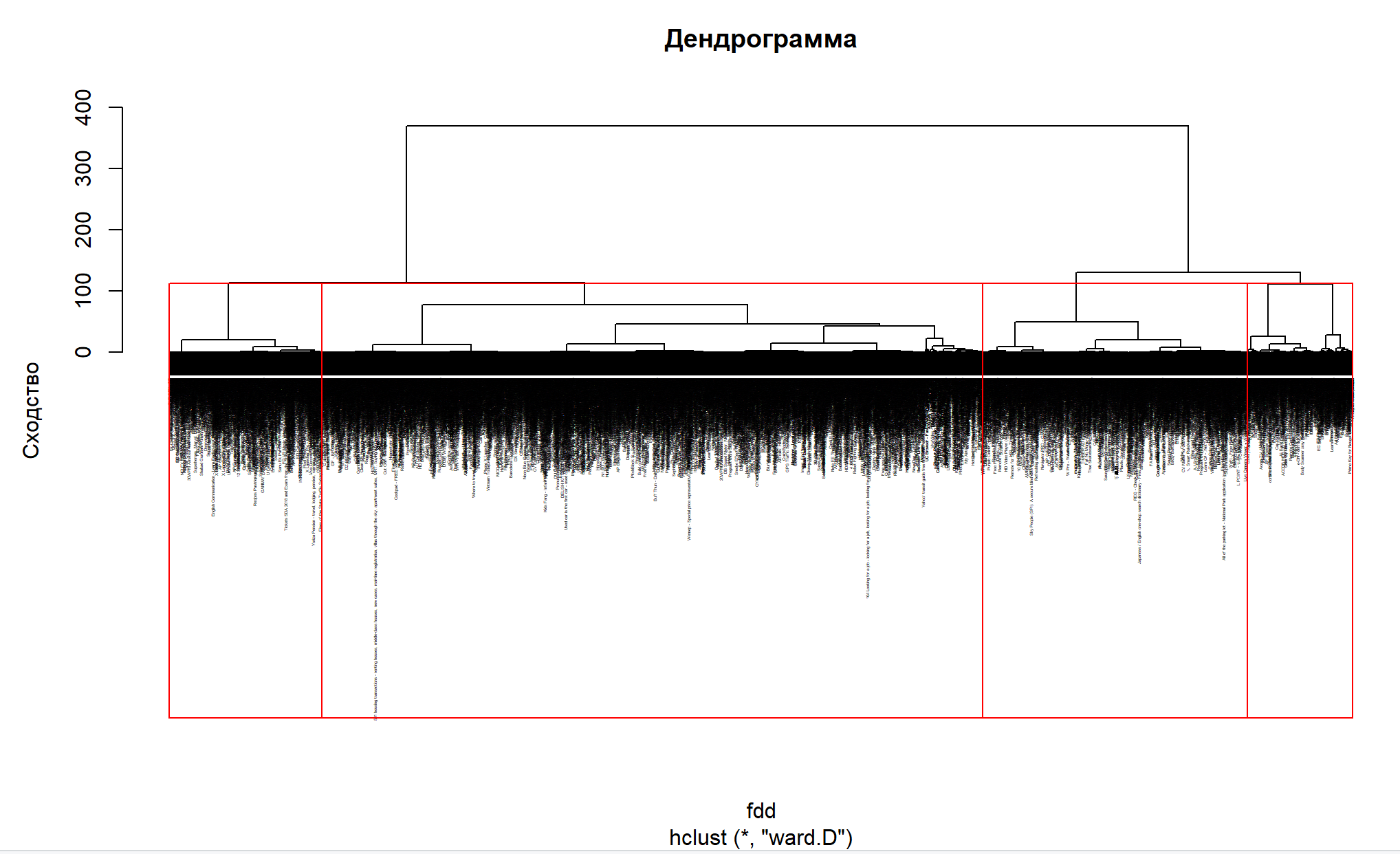


Рисунок 7 – Дендрограмма с кластерами.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 8 – Диаграмма «Каменная осыпь».

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 9 – Раскрашенная дендрограмма

Далее представим пример части списка приложений группы 3.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 10 – Приложения, разбитые по кластерам

Далее представим распределения характеристик по всем кластерам.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 11 – Распределение характеристик в каждом кластере.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 12 – Распределение характеристики Рейтинг.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 13 – Распределение характеристики Отзывы.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 14 – Распределение характеристики Размер.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 15 – Распределение характеристики Скачивание.

Рейтинг-общий рейтинг пользователя приложения.

Отзывы-количество отзывов пользователей для приложения.

Размер-размер приложения.

Скачивание- количество пользовательских загрузок / установок для приложения.

На рисунке 10 изображена диаграмма рассеивания, выражающая зависимость между индексами Рейтинга и Скачиваний.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 16 – Диаграмма рассеивания для Рейтинга и Скачиваний.

После были построены «ящики с усами», показанные на рисунке 17.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 17 – Диаграмма ящик с усами.

На нем мы можем видеть, как распределяются индексы рейтинга по кластерам. Можно заметить, что в третьем кластере этот индекс превышает все остальные.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 18 – График, показывающий распределение размера и рейтинга

Можно сделать вывод, что размер приложения не сильно влияет на его рейтинг.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 19 – Трехмерный график

Данный график классифицирует приложения по индексу рейтинга, скачиваний и количество отзывов. Можно заметить что количество отзывов не влияет на рейтинг и скачивания. Рейтинг так же не сильно влияет на количество скачиваний.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 20 – Зависимость количества отзывов от скачиваний в кластерах.

На данном шаге была произведена кластеризация k-means. Данная кластеризация позволяет отследить взаимосвязь двух непрерывных переменных. В данном случае – скачиваний больше, когда отзывов меньше, что странно.

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

Сначала строим таблицы вероятностей по всем характеристикам. В первом столбце – средние значения параметров, во втором столбце – их стандартные отклонения.

**Изображение выглядит как текст, чек

Автоматически созданное описание**

Рисунок 21 –Таблица вероятностей

Далее строим функции плотности для характеристик. Распределение на графике подтверждает правильность распределения на кластеры.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 22 – Функции плотности

Далее проведем классификацию Decision Tree.

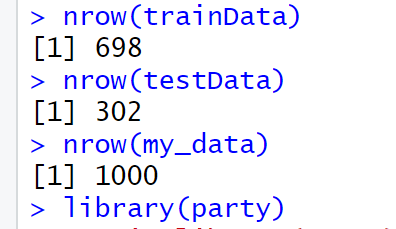


Рисунок 23 – Распределение филиалов

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 24 – Дерево условного вывода с 7 конечными узлами

Следующим шагом строим модель, используя функцию ctree(), передав ей в качестве параметра нашу формулу и выборку для обучения модели (в нашем случае - это trainData).

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 25 – Модель

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 26 – Таблица с классификацией

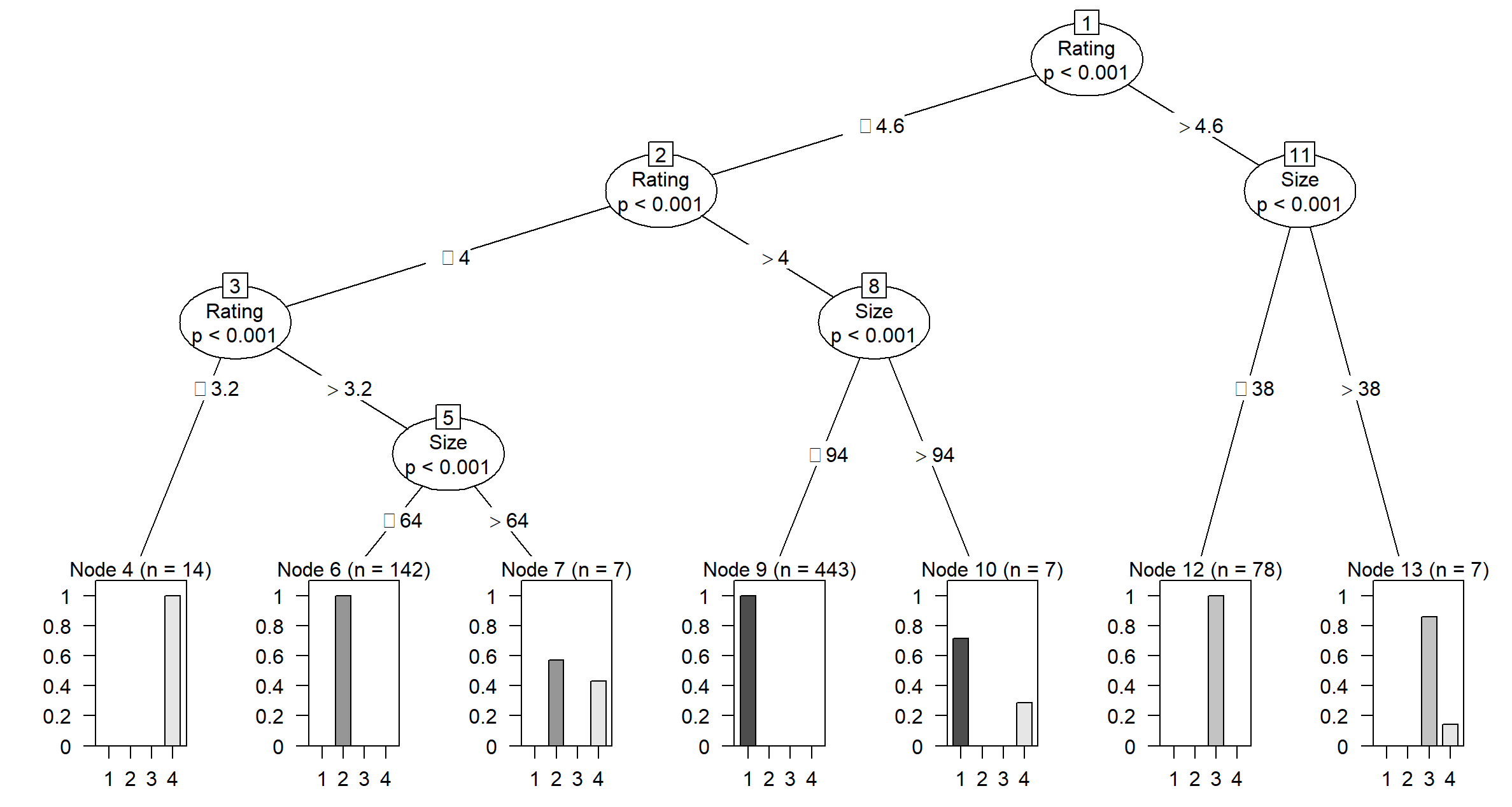


Рисунок 19 – Дерево.

Попробуем использовать более точный алгоритм Random Forest.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 20 – Результат работы алгоритма Random Forest

**Вывод**

В ходе работы я закрепила знания об алгоритмах классификации и кластеризации данных, ознакомилась с некоторыми функциями языка R, осуществляющими этот вид анализа, принципами их работы. Научилась визуализировать результаты работы функций кластерного анализа и классификаторов, интерпретировать полученные результаты. Научилась выполнять классификацию на основе формулы Байеса и деревьев решений