МИНИСТЕРТСВО 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**

**ОТЧЕТ**

**о выполнении лабораторной работы №6.1**

по дисциплине «Обработка больших данных»

Работу выполнил \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ студент 39/2 группы, Мандыч Д.И.

Работу проверил\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_преподаватель Шиян В.И.

Краснодар 2023

**Набор данных**

Для выполнения работы был использован набор данных из репозитория машинного обучения UCL.

Информация о наборе данных:

Эти данные отображают достижения учащихся в среднем образовании двух португальских школ. Атрибуты данных включают оценки учащихся, демографические, социальные и школьные характеристики), и они были собраны с помощью школьных отчетов и анкет. Представлены два набора данных, касающихся успеваемости по двум различным предметам: математике (mat) и португальскому языку (por).

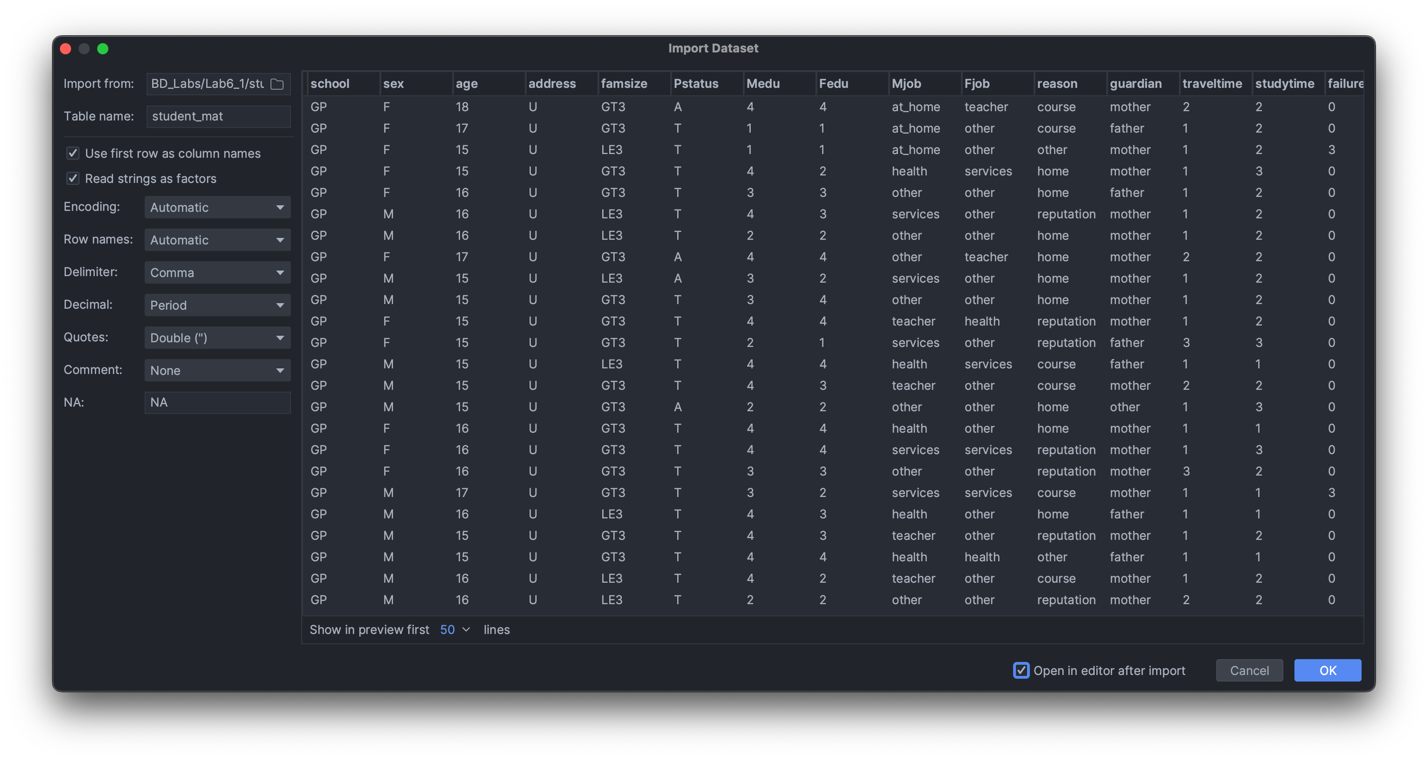
****

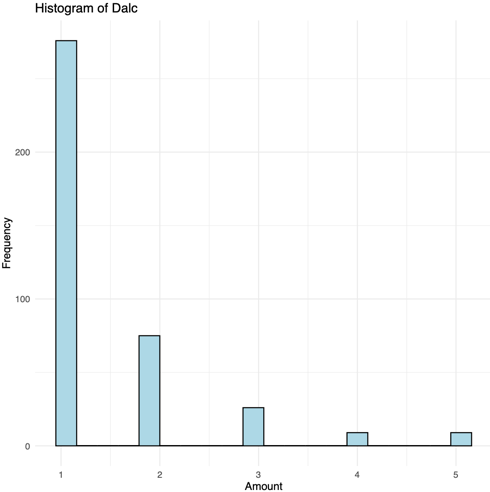
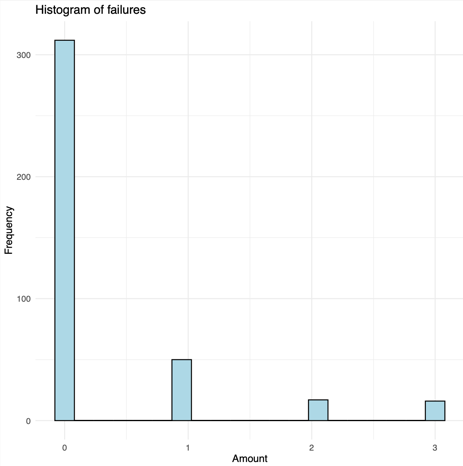
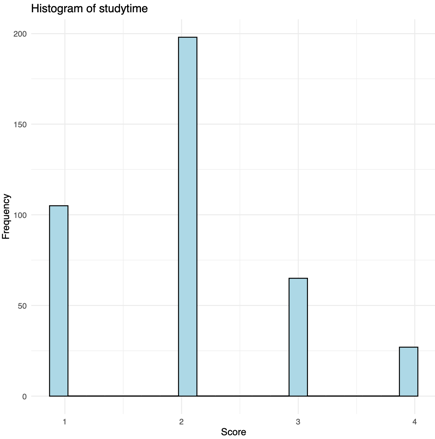
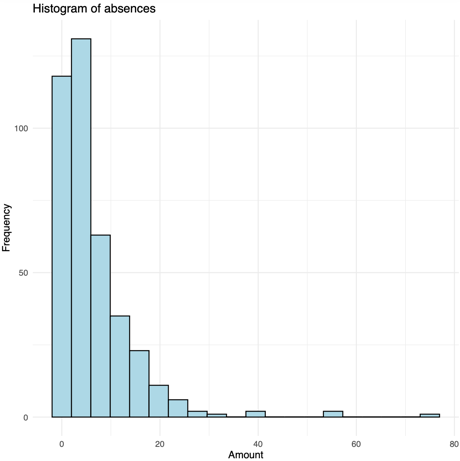
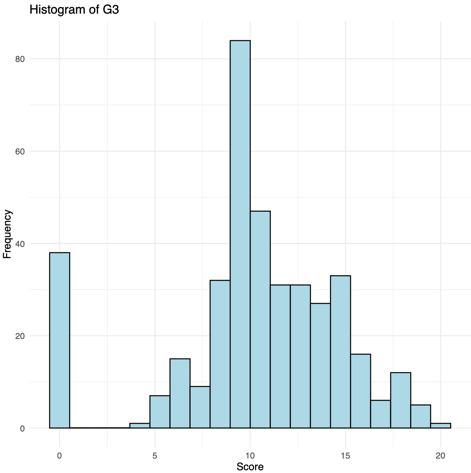
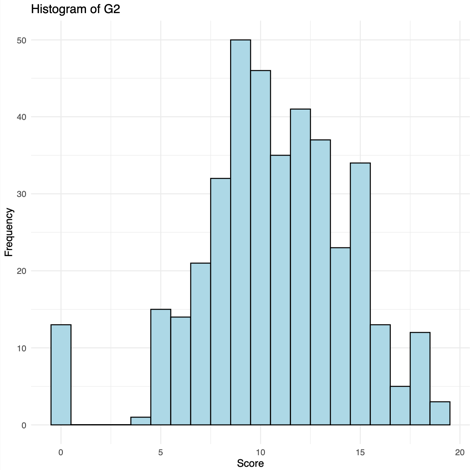
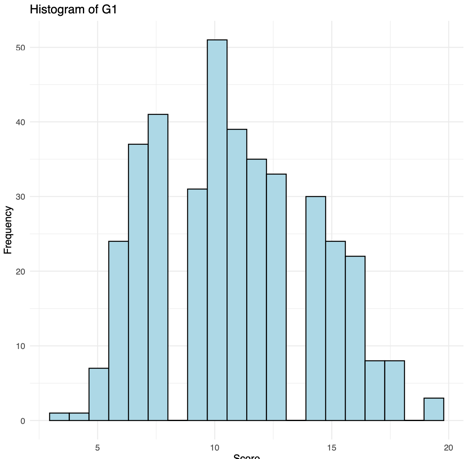
Рисунок 1 – Используемый набор данных

**Анализ данных**

Проведем анализ полученного набора данных.

Для начала проанализируем частоту встречаемости числовых параметров, так как в данной предметной области они играют большую роль.

Представим наиболее важные из них графически.

Chart, bar chart

Description automatically generated

Рисунок 1 – Графическое представление нескольких из основных параметров

**Построение дендограммы**

В результате построения дендограммы можно четко 2–4 основных кластера. Далее будем работать с 3 кластерами.

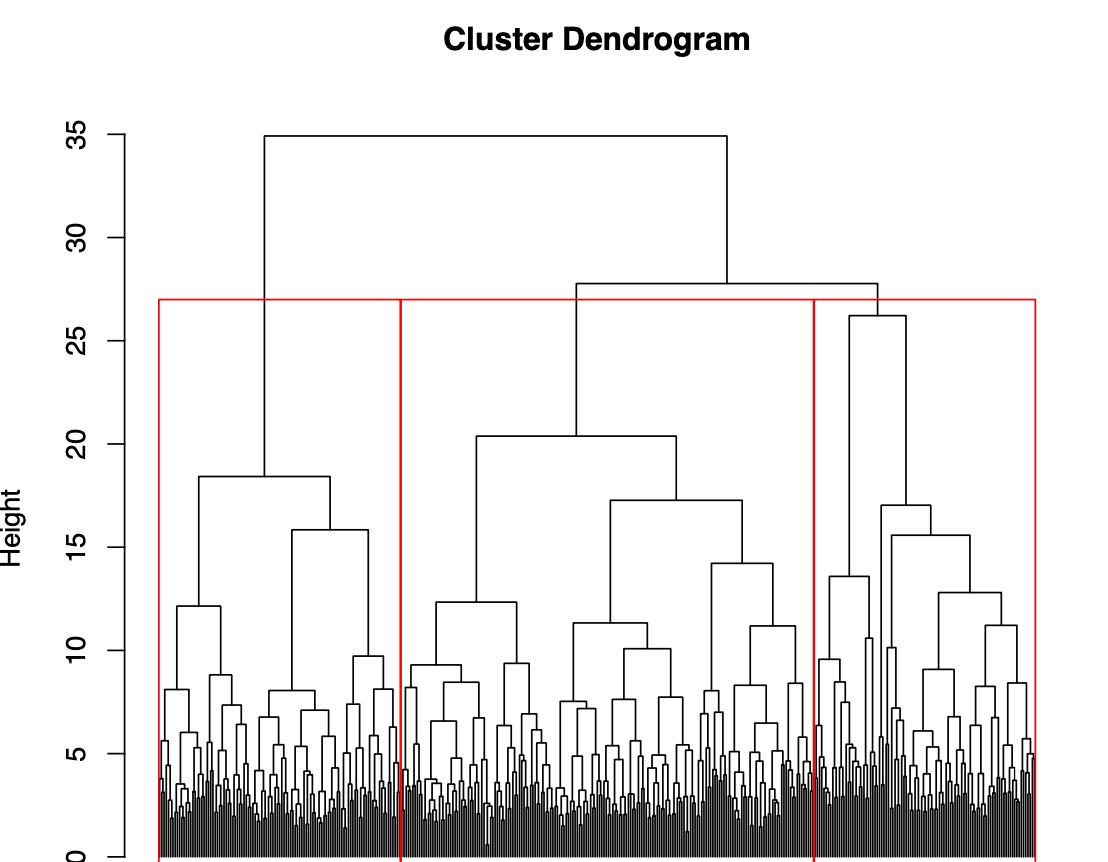


Рисунок 2 – Результат построения дендограммы с выделением выбранных кластеров

**Характеристики кластеров**

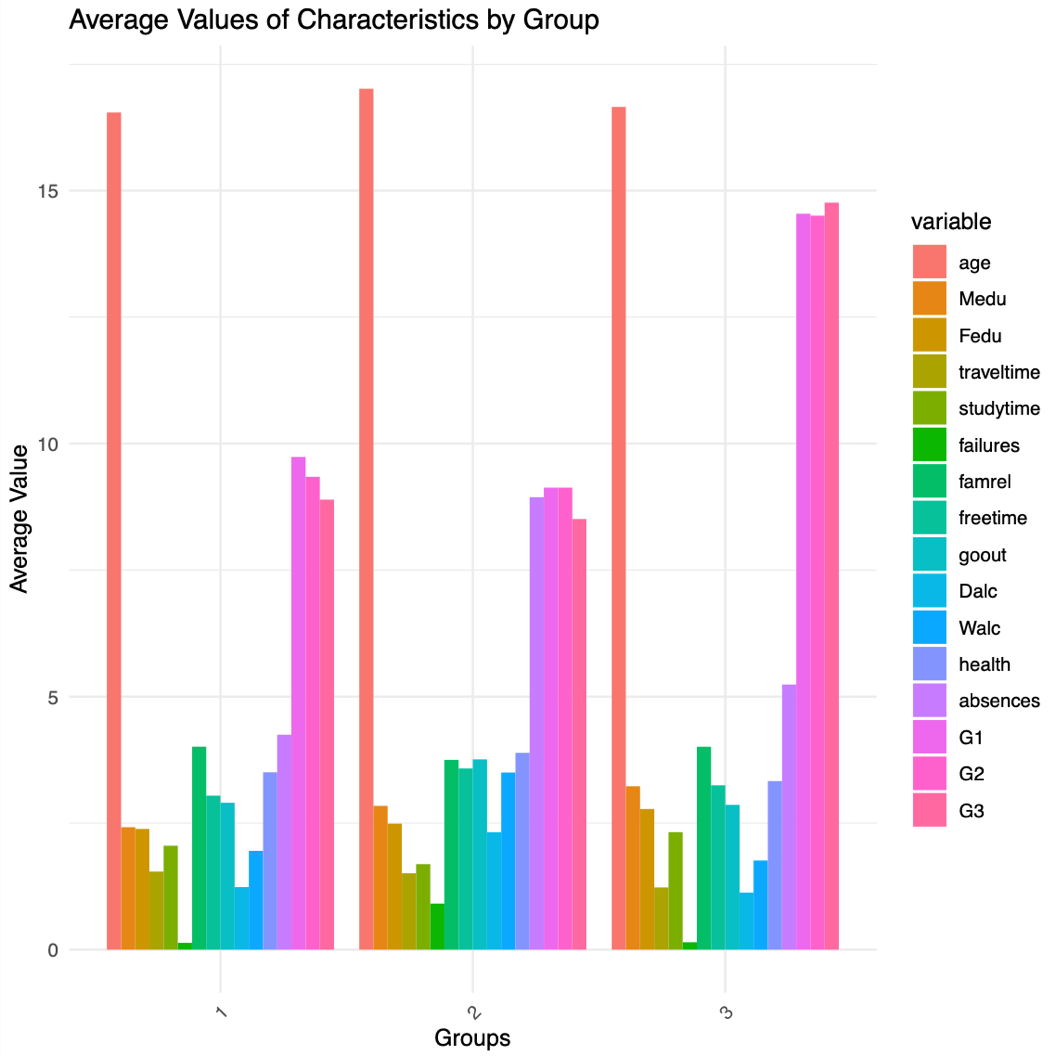


Рисунок 3 – Графическое представление средних характеристик кластеров

Данный график хорошо показывает отличающиеся характеристики кластеров. Первый кластер имеет низкий показатель успеваемости. Второй кластер имеет низкий показатель успеваемости, а также высокие показатели пропусков занятий и количество приемов алкоголя. Третий кластер имеет лучшую успеваемость среди других и, в основном, больше не отличается от первого.

**Диаграмма "Каменная осыпь"**

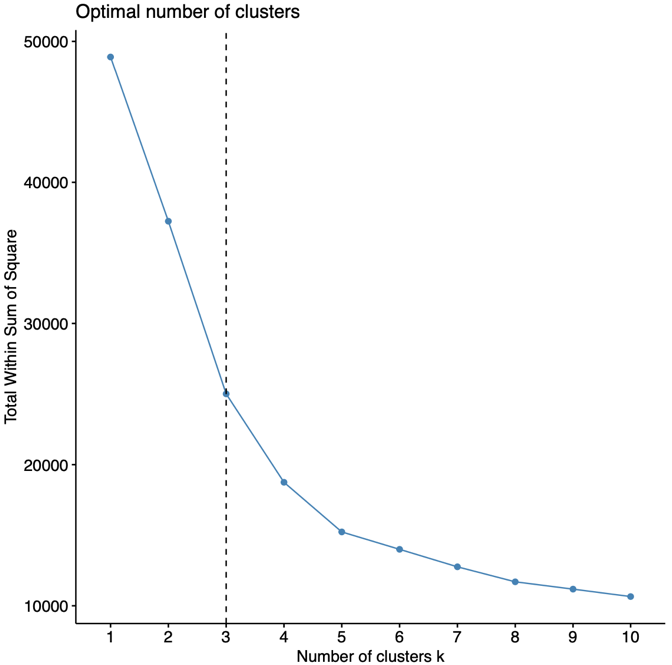
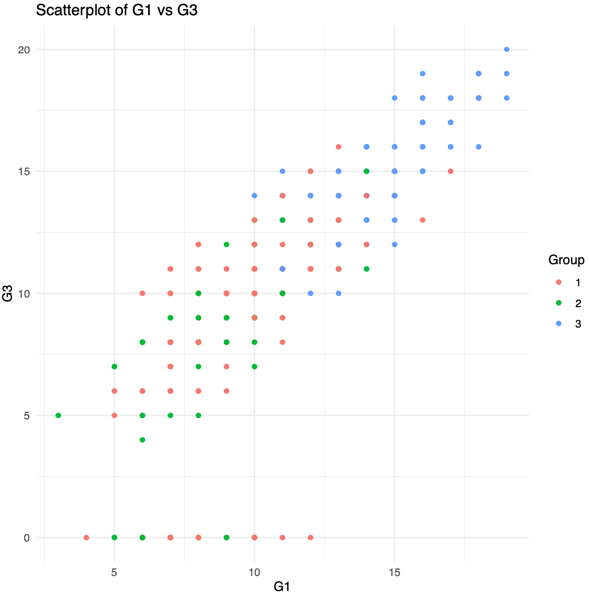
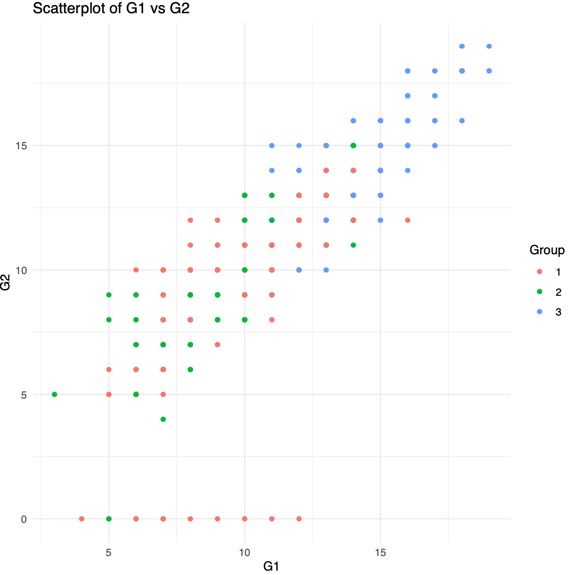
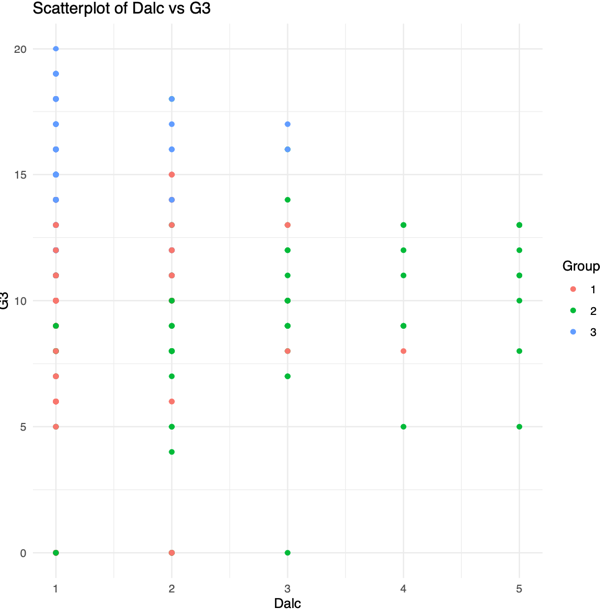
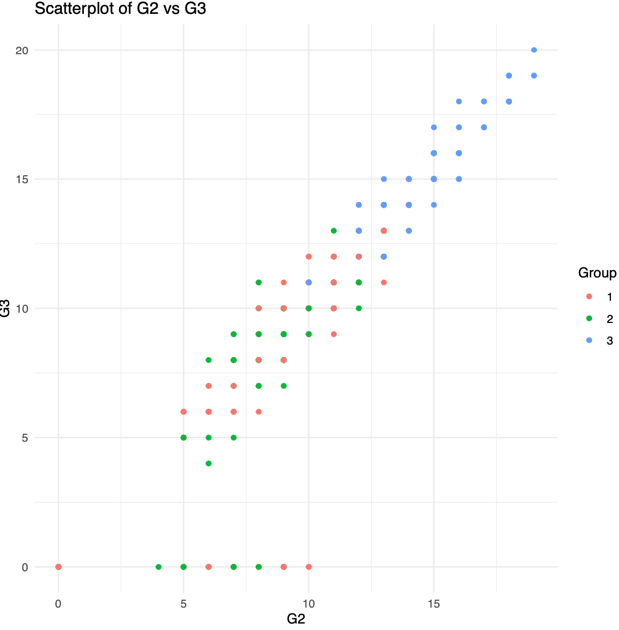


Рисунок 4 – Диаграмма «Каменная осыпь»

Данная диаграмма, подтверждает правильность выбора количества кластеров для рассмотрения.

**Построение scatterplot**



Chart, scatter chart

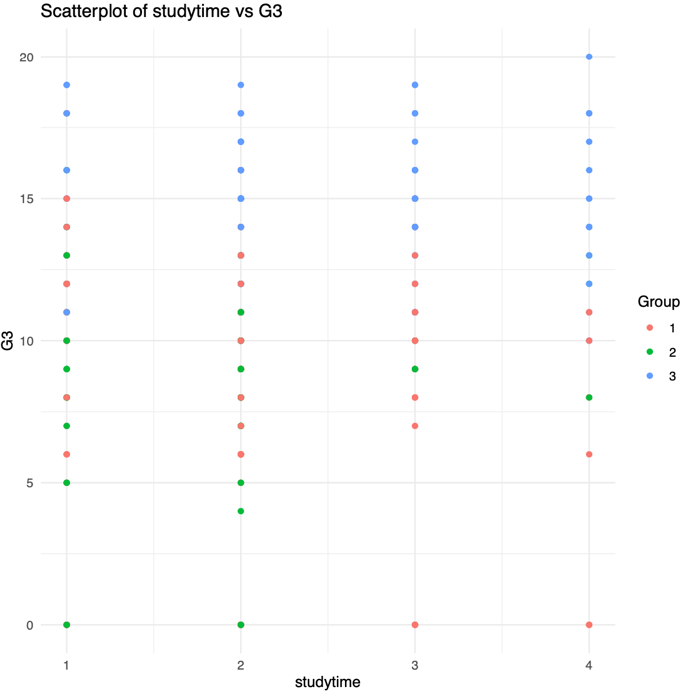
Description automatically generated

Рисунок 5 – Графики scatterplot

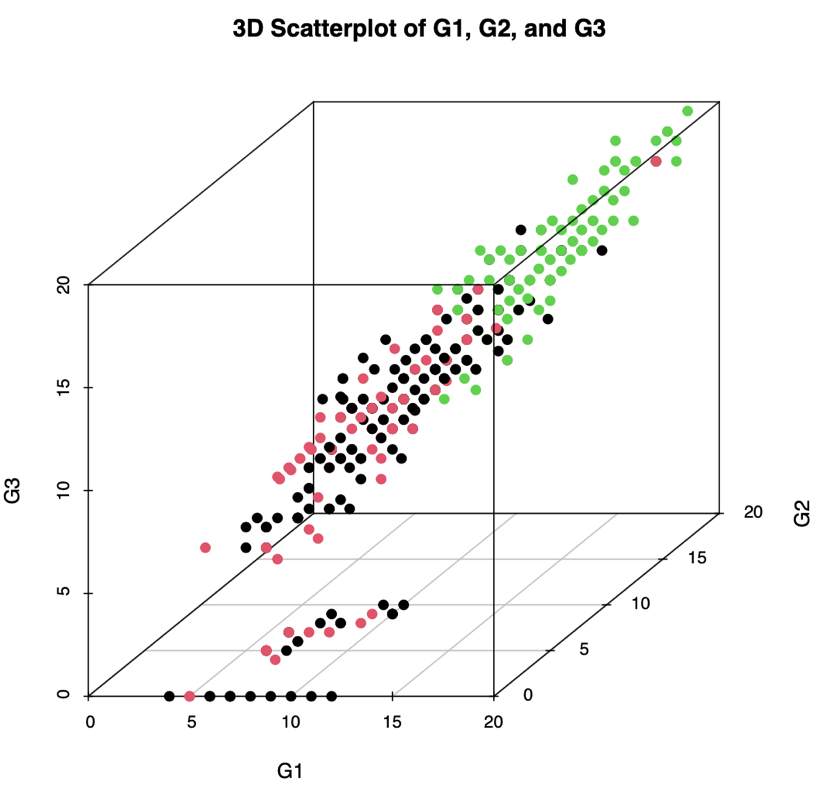


Рисунок 6 – 3D График scatterplot для оценок G1, G2, G3

**Вывод по первой части**

По результатам детального анализа набора данных, можно сделать следующие выводы:

1. G1, G2 и G3 линейно зависимы. При получении высокого балла учеником за G1, G2, есть большая вероятность получения высокого балла G3.
2. Учащиеся, получающие высокий балл, предпочитают не выпивать.
3. Учащиеся, получающие высокий балл, проводят больше времени за учебой.
4. Третий кластер содержит наиболее успевающих учеников.
5. Второй кластер содержит учеников, у которые имеют проблемы с успеваемостью, а также большое количество пропусков занятий и приемов алкоголя.
6. Первый кластер отличается от третьего в основном только в успеваемости: у них она хуже.

**Наивного Байесовского классификатора**

Реализуем наивный байесовского классификатора. Для этого разделим набор данных на тренировочный и тестовый. Тренировочный будет составлять 70%, а тестовый – 30%.

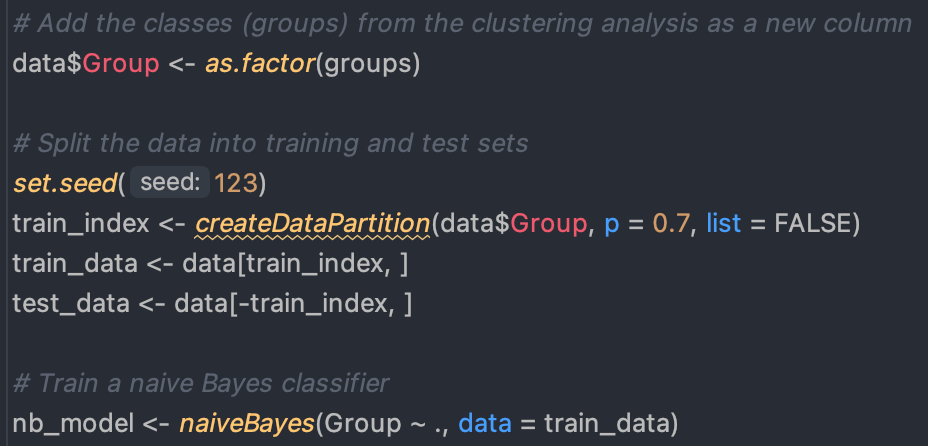


Рисунок 7 – Реализация наивного байесовского классификатора

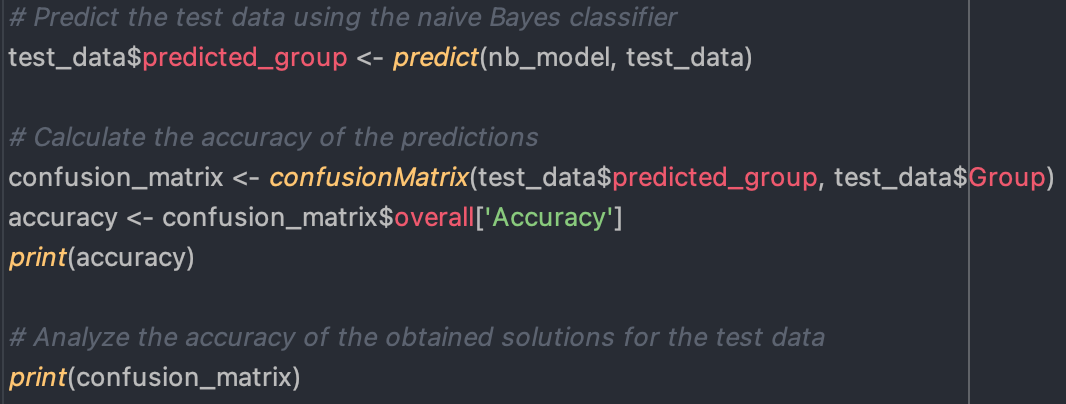


Рисунок 8 – Реализация определения кластера и вывод результатов

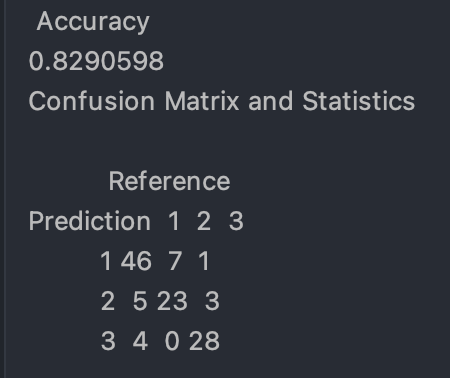


Рисунок 9 – Результаты определения кластера

**Метод деревьев решений**

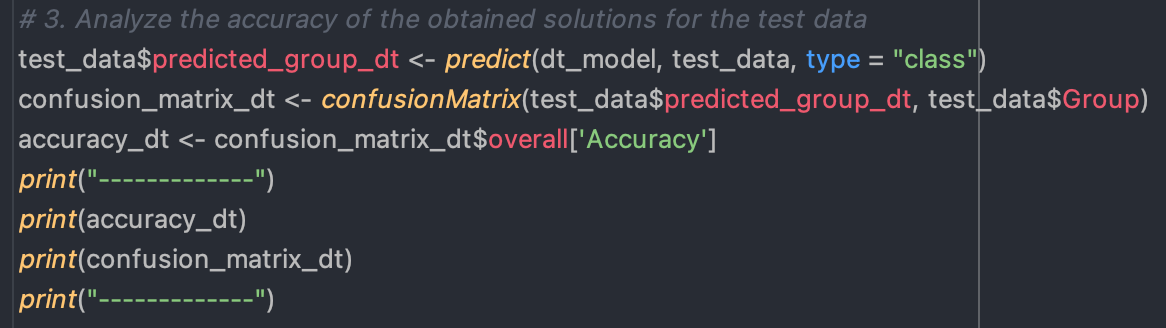


Рисунок 10 – Реализация метода деревьев решений

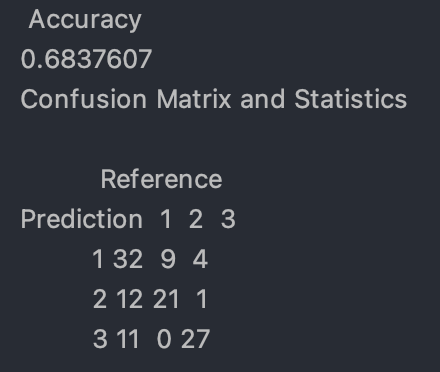


Рисунок 11 – Результаты метода деревьев решений

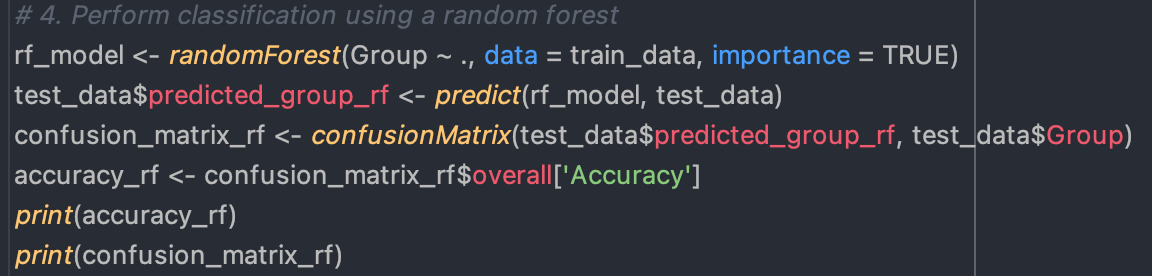


Рисунок 12 – Реализация Random Forest

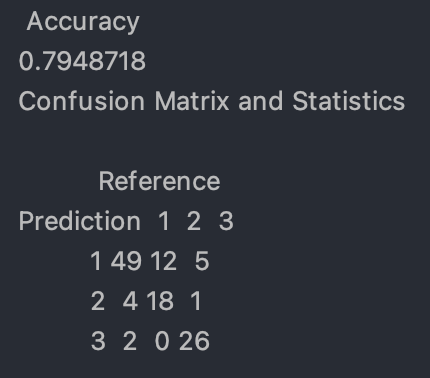


Рисунок 13 – Результат Random Forest

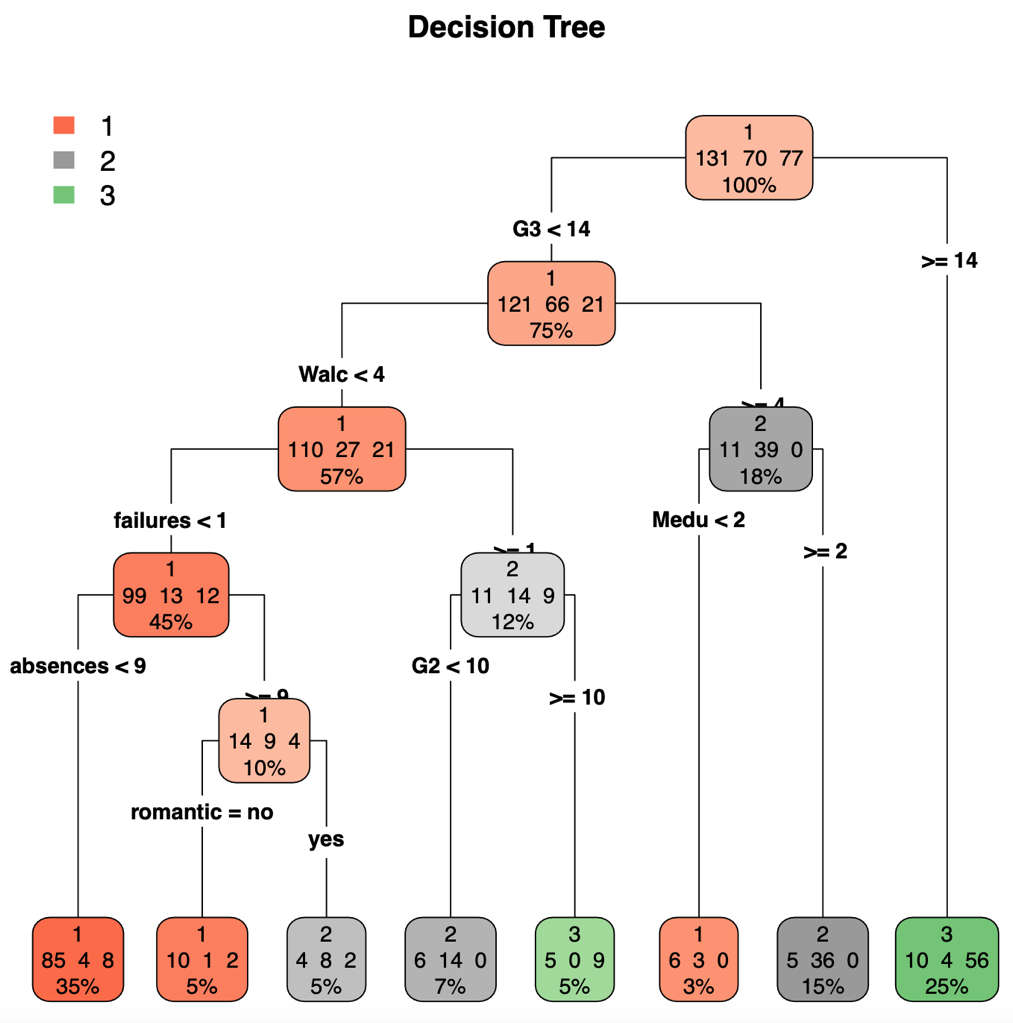


Рисунок 14 – Графическое представление дерева решений

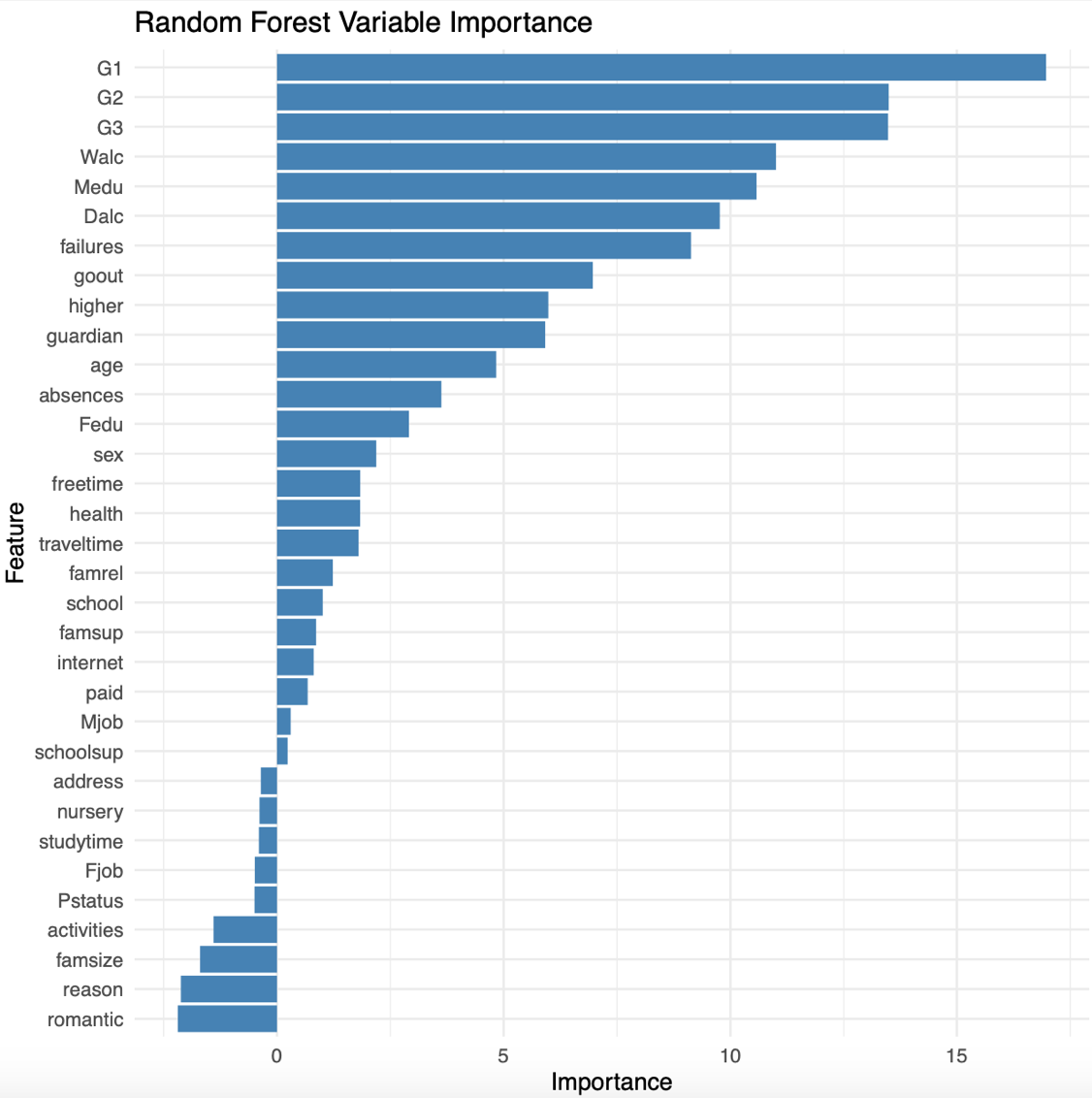


Рисунок 15 – Графическое представление Random Forest