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 №5.2

по дисциплине «Обработка больших данных»

Тема**:** Байесовская классификация и деревья принятия решений на R

Выполнил: ст. гр. 36/2

Решетка Д.В.

Проверил: преподаватель

Яхонтов А.А.

Краснодар

2025

**Цель:** Научиться выполнять классификацию на основе формулы Байеса и деревьев решений.

### **Задание**

Данные берутся из предыдущего набора (ЛР 6 часть 1). Вам необходимо добавить найденные при в результате кластерного анализа классы (groups), как вектор-столбец в предыдущий DATASET, разделить его на две части (обучающую и тестовую) обучить на обучающей выборке классификатор, а затем применить классификатор к тестовым данным.

Для наивного Байесовского классификатора:

1. Решите задачу с помощью наивного Байесовского классификатора;

2. Проанализируйте точность полученных решений для тестовых данных

Для деревьев решений:

1. Примените метод деревьев решений для задачи классификации (для того же набора данных).

2. Исследуйте дерево решений; если позволяет размерность, постройте его график.

3. Проанализируйте точность полученных решений для тестовых данных (с известным значением переменной отклика), сравните результаты с ранее полученными.

4. Выполнить классификацию с помощью случайного леса, сопоставить результат с результатом дерева решения, прокомментировать результат сравнения.

5. Сопоставьте результаты с результатами Байесовского классификатора.

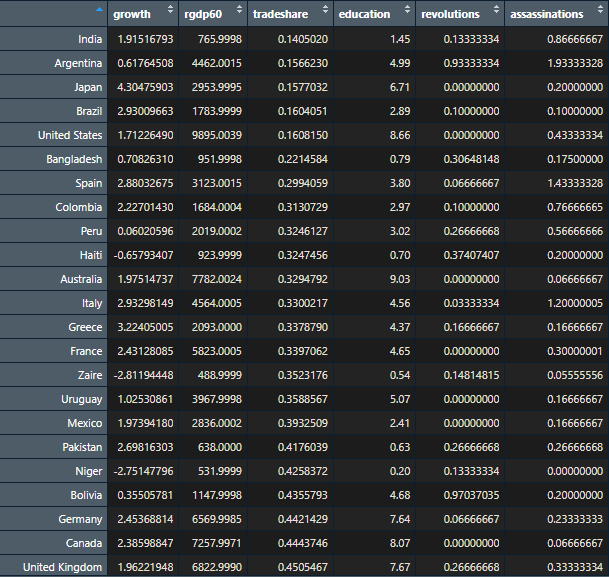
**Вариант**:



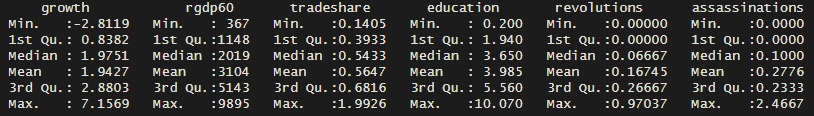
**Ход работы**

1. Исходные данные:

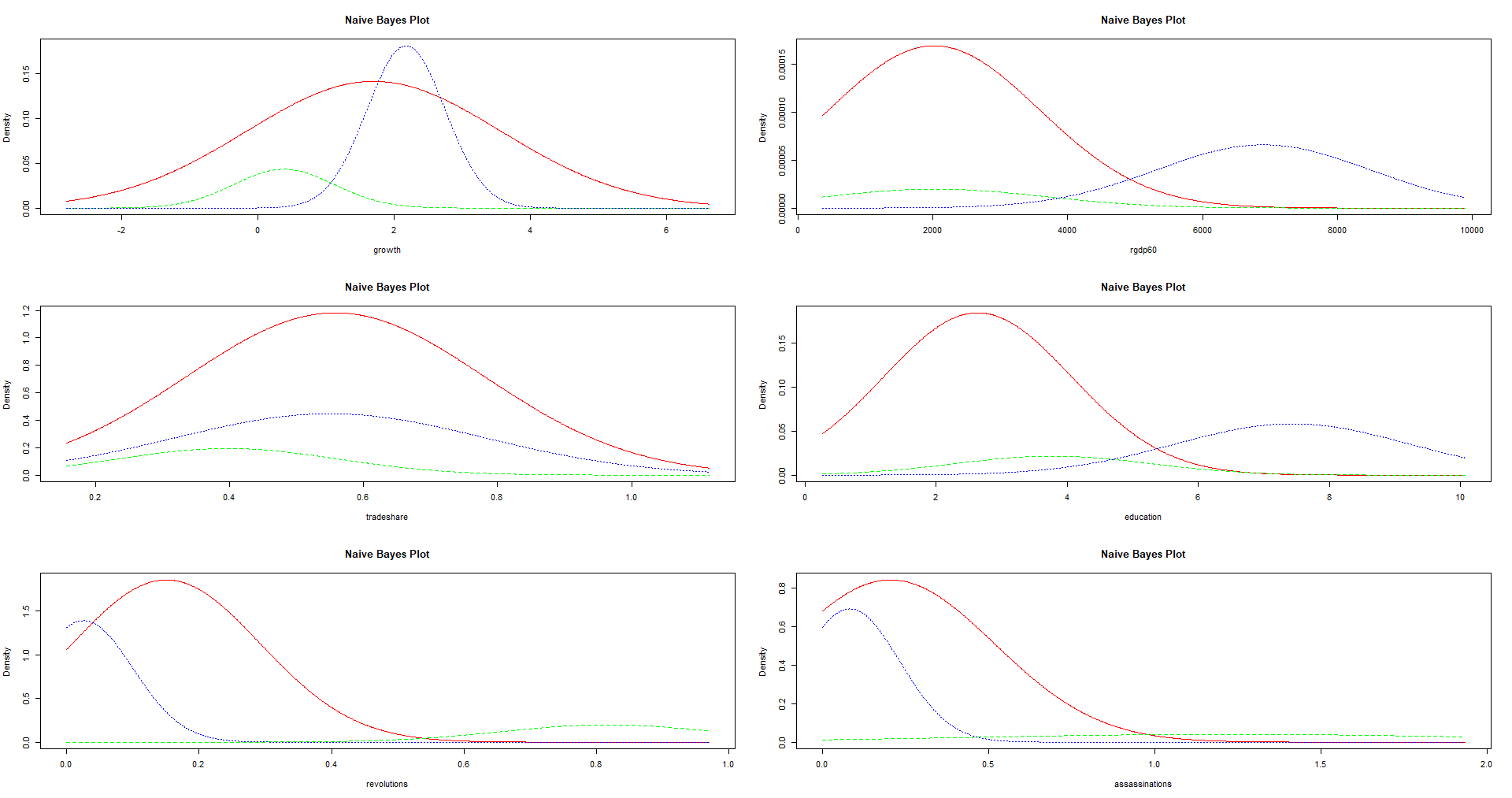
Таблица с метриками стран.



1. Дескриптивный анализ.



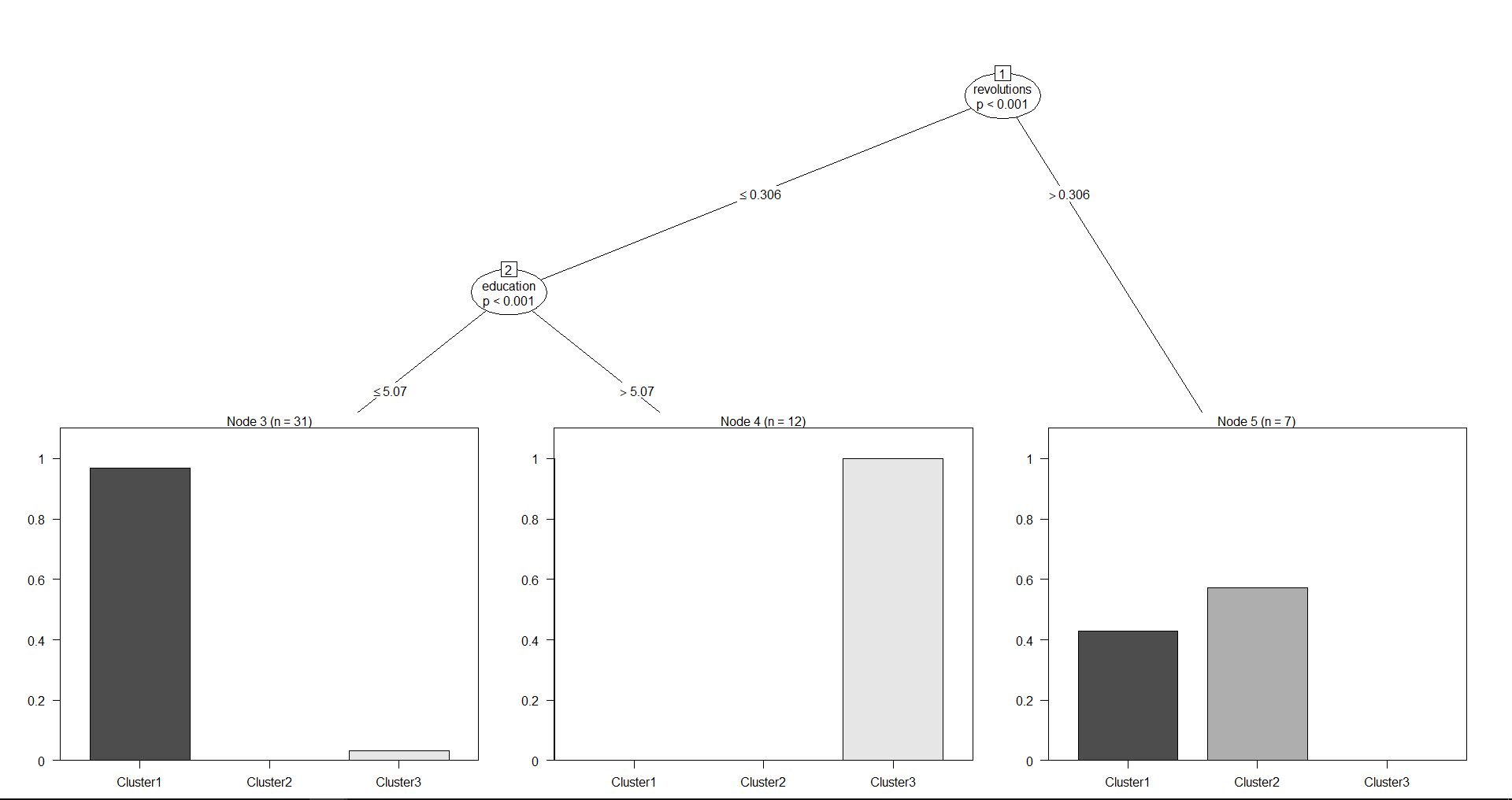
1. Ядерные функции плотности условий вероятности.





По графикам можно увидеть, как значения каждого признака распределяются в каждом кластере. Например, на графике growth видно, что плотность для третьего кластере высокая для значения 2.

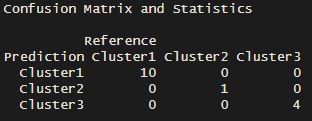
1. Дерево решений.



На этой диаграмме изображено по какому принципу обучалась модель. В эллипсах указаны наиболее важные для классификации параметры (чем меньше значение p, тем важнее параметр). Далее по значениям на ветвях данные делятся и попадают в разные листья.

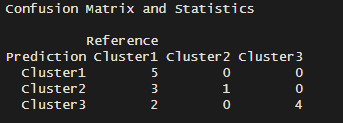
1. Проверка моделей.

**Наивный Байесовский классификатор.**



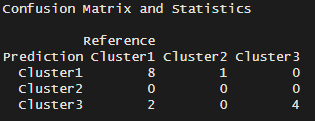
Как видно из результатов, ошибок не было.

**Дерево решений.**



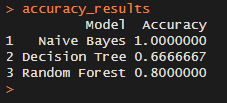
Как видно из результатов, было 5 ошибок.

**Случайный лес.**



Как видно из результатов, было 3 ошибки.

1. Сравнение точностей моделей.



Наивный Байесовский классификатор оказался самым точным, после него идёт случайный лес, а дерево решений оказалось самым неэффективным.

**Вывод**

В ходе выполнения лабораторной работы была проведена работа с классификацией на основе формулы Байеса, деревьев решений и случайного леса.