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 №5.**

**Дисциплина: Обработка больших данных**

Работу выполнил:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Вавакин В.О.

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

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

Краснодар

2025

**Тема**: Задачи классификации и кластеризации. Байесовская классификация и деревья принятия решений в R.

**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 5:

|  |  |
| --- | --- |
| Вавакин Владислав Олегович | Страны по уровню жизни |

**Ход работы**

1. Приведём фрагмент начальных данных и решаемую задачу.

Фрагмент исходного датасета:

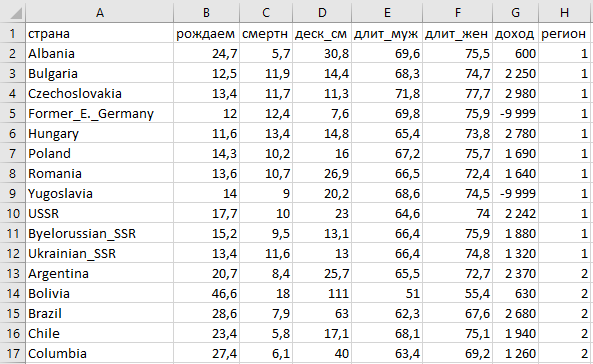


Рисунок 1. Фрагмент исходного датасета

Задача: провести кластерный анализ наблюдений. Постараться найти такой метод кластеризации, чтобы полученные кластеры соответствовали регионам. Из условий следует, что при проведении кластерного анализа переменная *регион* должна быть исключена из анализа.

2. Выполним дескриптивный анализ по каждому из числовых атрибутов датасета. Он будет включать минимальное и максимальное значение и соответствующие им страны, среднее арифметическое, медиану, моду, а также боксплот.

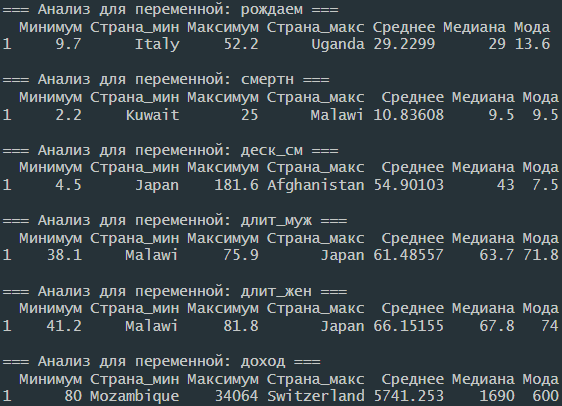


Рисунок 2. Числовые результаты дескриптивного анализа

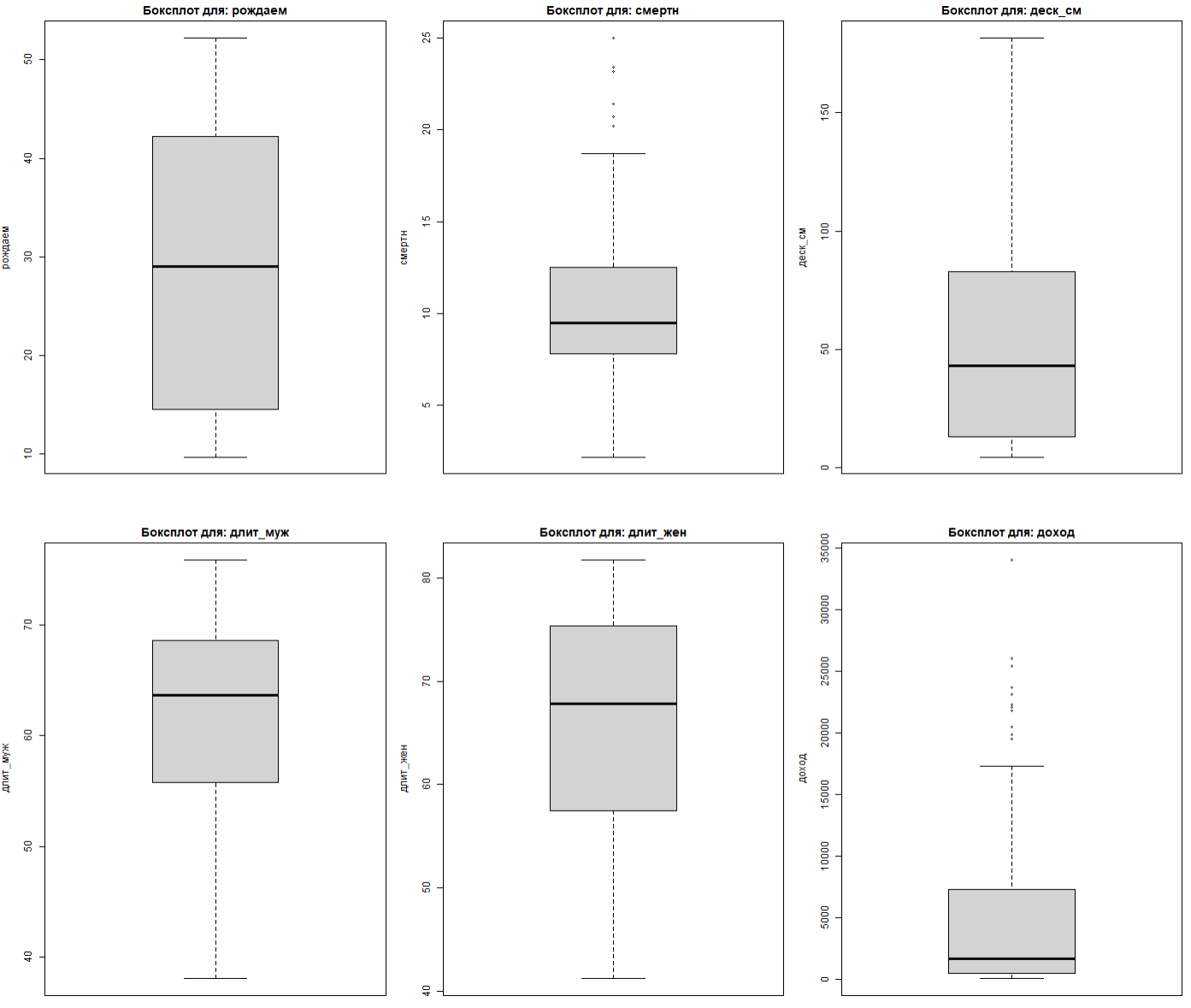


Рисунок 3. Графические результаты дескриптивного анализа

3. Оценим оптимальное количество кластеров. Для этого воспользуемся четырьмя различными методами:

1) Метод локтя: рассматривает общую сумму квадратов внутри кластера относительно количества кластеров.

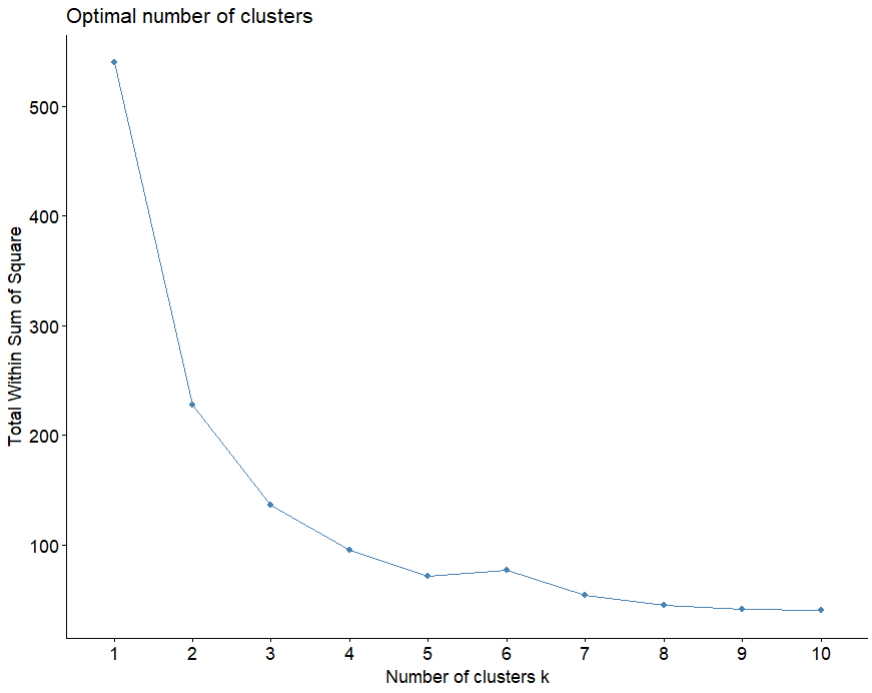


Рисунок 4. График внутригруппового разброса по методу локтя

Согласно данному графику, внутригрупповой разброс замедляется после отметки в 7 кластеров, и вовсе увеличивается при 6 кластерах.

2) Метод среднего силуэта: определяет насколько хорошо каждая точка лежит в пределах своего кластера.

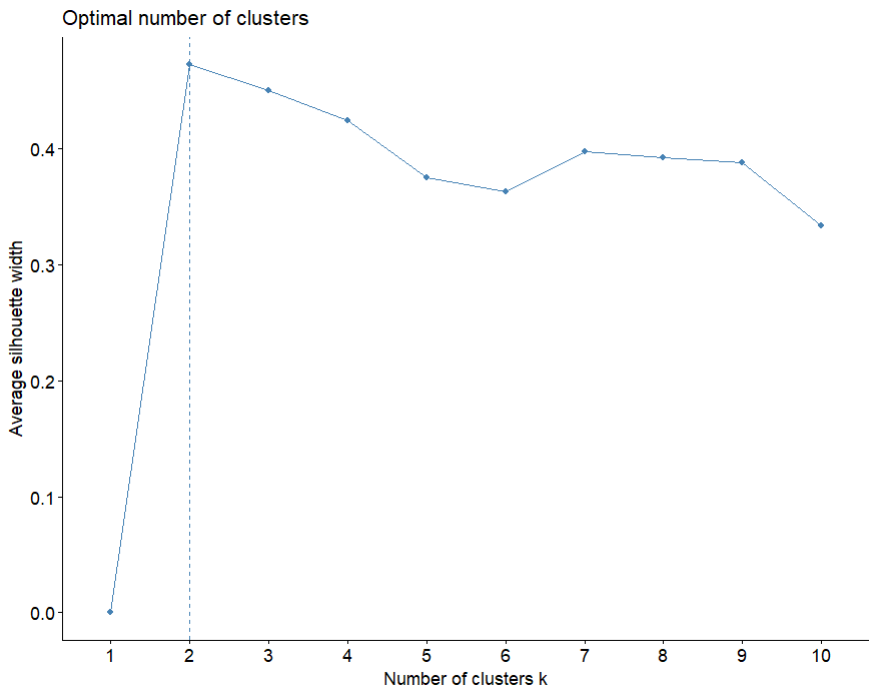


Рисунок 5. График ширины силуэта по методу среднего силуэта

Исходя из этого графика, разделение на 2 кластера наиболее оптимально, но при 3 и 4 кластерах значение ширины силуэта остаётся достаточно хорошим. При 5 и 6 кластерах расстояние между кластерами уменьшается, а от 7 до 9 – снова увеличивается.

3) Статистика разрыва: сравнивает общую внутрикластерную дисперсию для различных значений k с их ожидаемыми значениями для распределения без кластеризации.

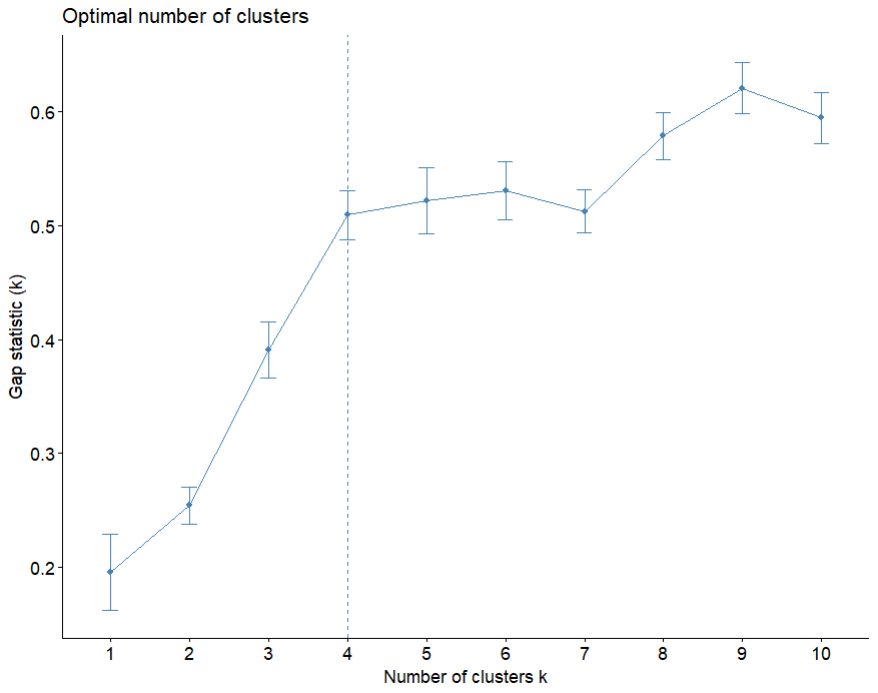


Рисунок 6. График статистики разрыва

Согласно результатам данного метода, наиболее оптимальное значение кластеров – 4.

4) Алгоритм на основе консенсуса: выполняет множество различных методов выбора количества кластеров и собирает по ним статистику, показывая наиболее согласованное число.

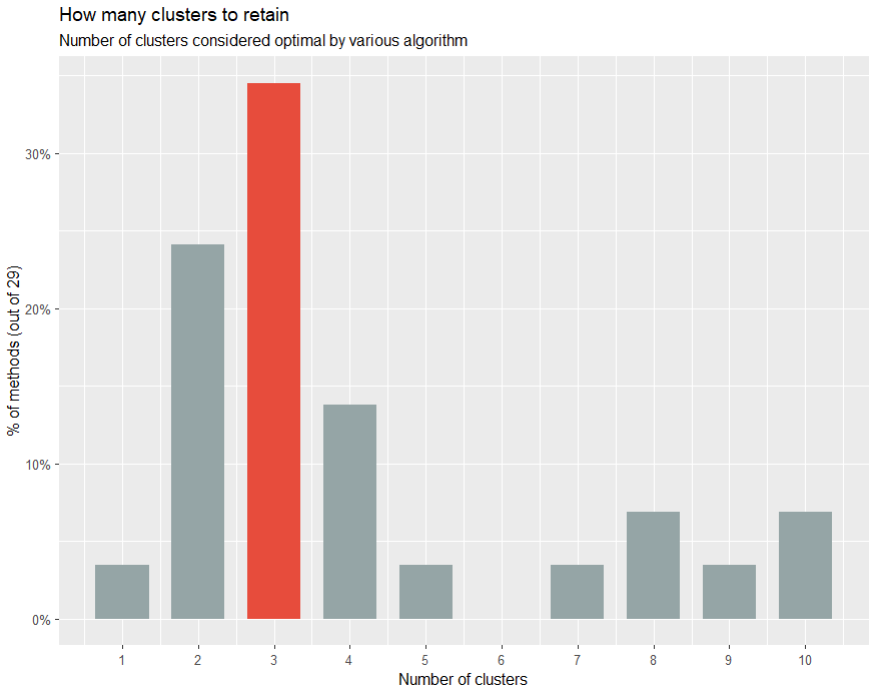


Рисунок 7. Диаграмма консенсуса методов выбора оптимального количества кластеров

Согласно консенсусу, наиболее оптимальное число кластеров – 3. Чуть менее оптимальными являются значения в 2 и 4 кластера, а среди большого числа выделяются разделения в 8 и 10 кластеров.

4. Выполним кластеризацию набора данных, построив дендрограмму.

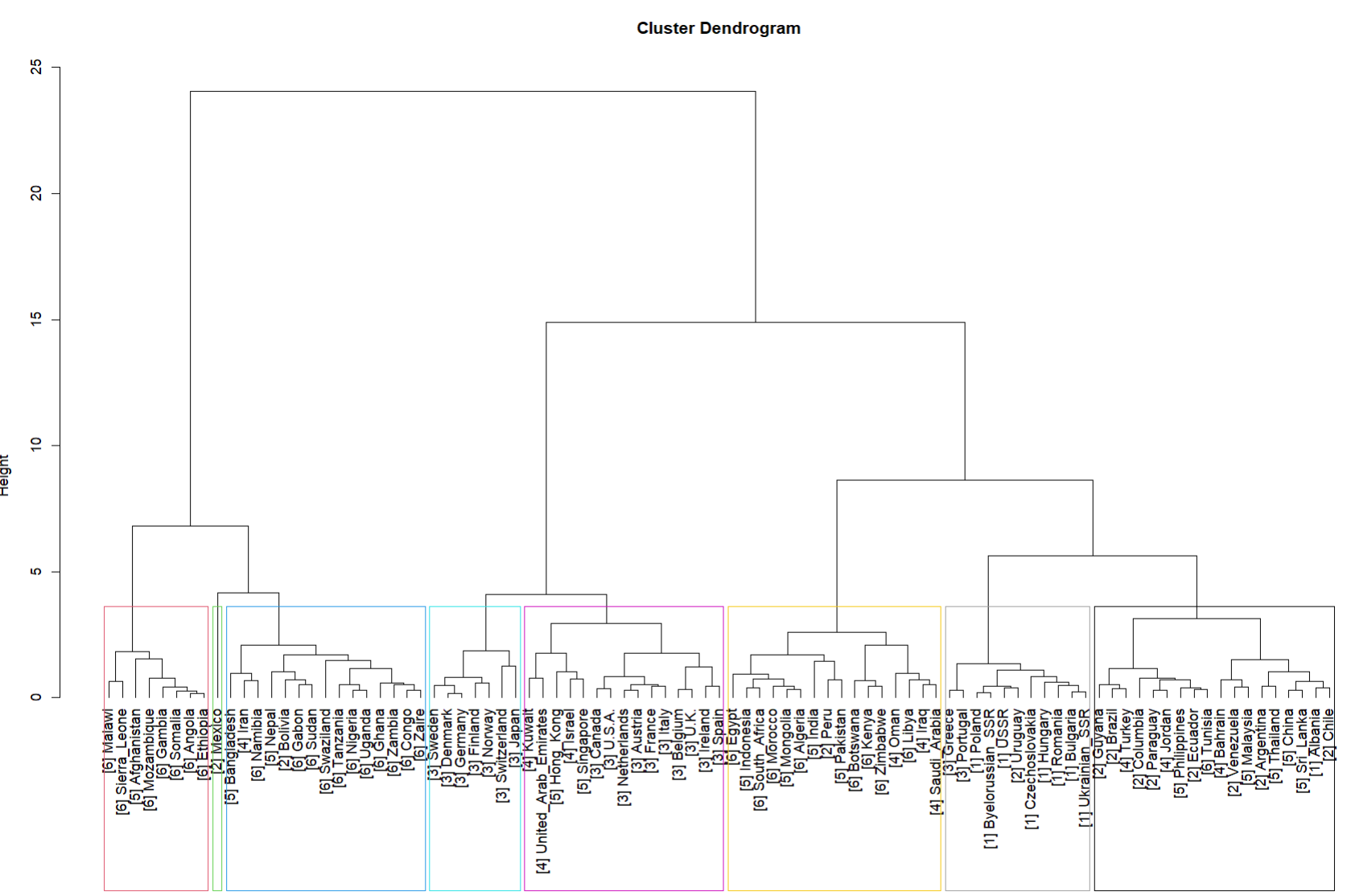


Рисунок 8. Дендрограмма иерархической кластеризации набора данных

В квадратных скобках перед названиями стран отображаются их регионы, указанные в поле *регион* исходного датасета.

Кластеризация была проведена на 8 кластеров, поскольку, согласно условию, кластеры должны соответствовать регионам, то есть кластеров должно быть хотя бы 6. При этом некоторые регионы, такие как Азия, Африка и Европа содержат большое разнообразие экономик, приводящих к небольшим группам, схожим по уровню жизни, такие как страны Скандинавии, страны ЕС и высокоразвитые малые экономики (ОАЭ, Кувейт, Гонконг, Сингапур и Израиль), которые либо пересекаются с группами из других регионов, либо выбиваются в отдельные кластеры.

5. Построим столбчатые диаграммы и боксплоты средних значений параметров для всех кластеров, полученных в дендрограмме.

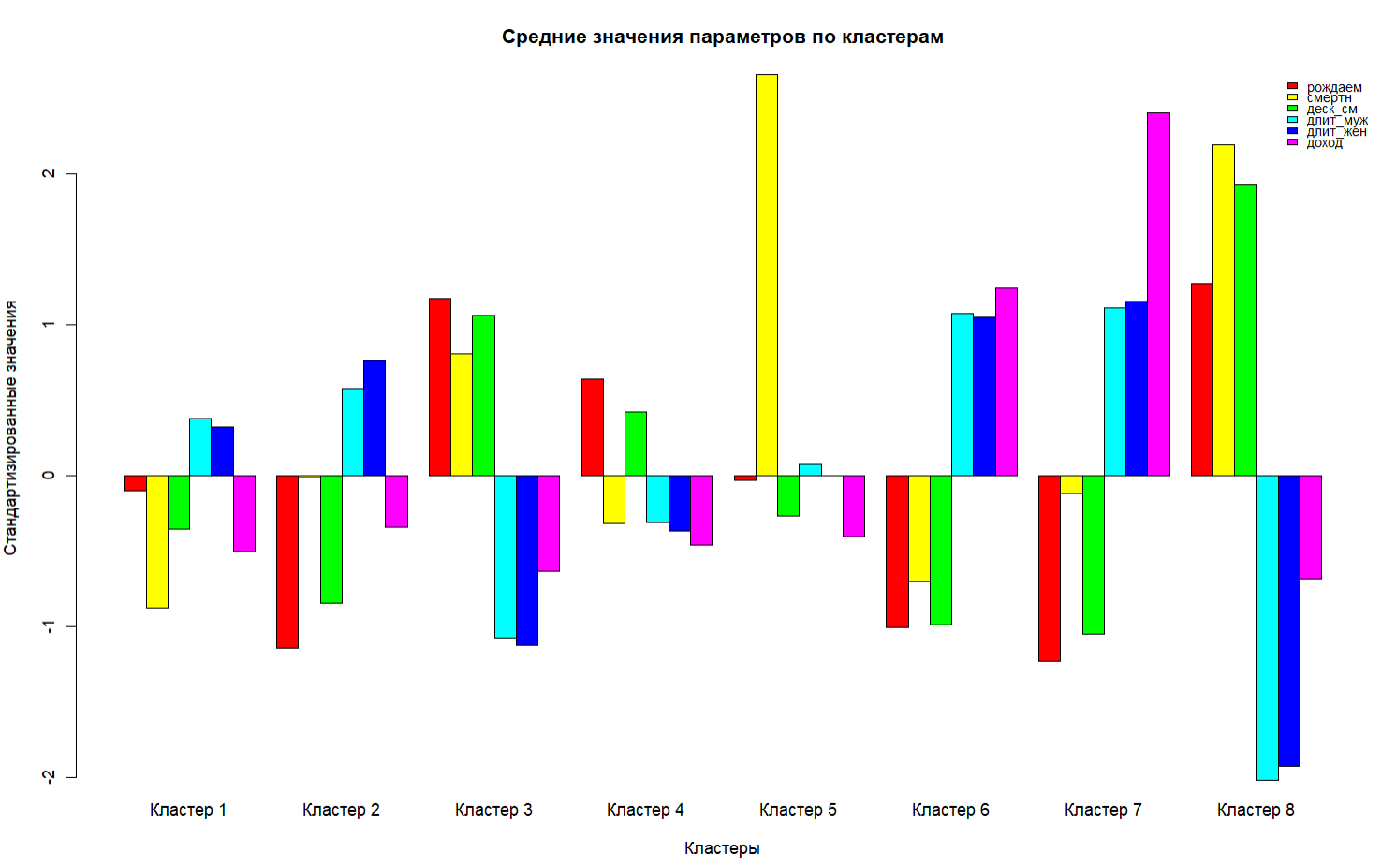


Рисунок 9. Столбчатая диаграмма кластеров дендрограммы

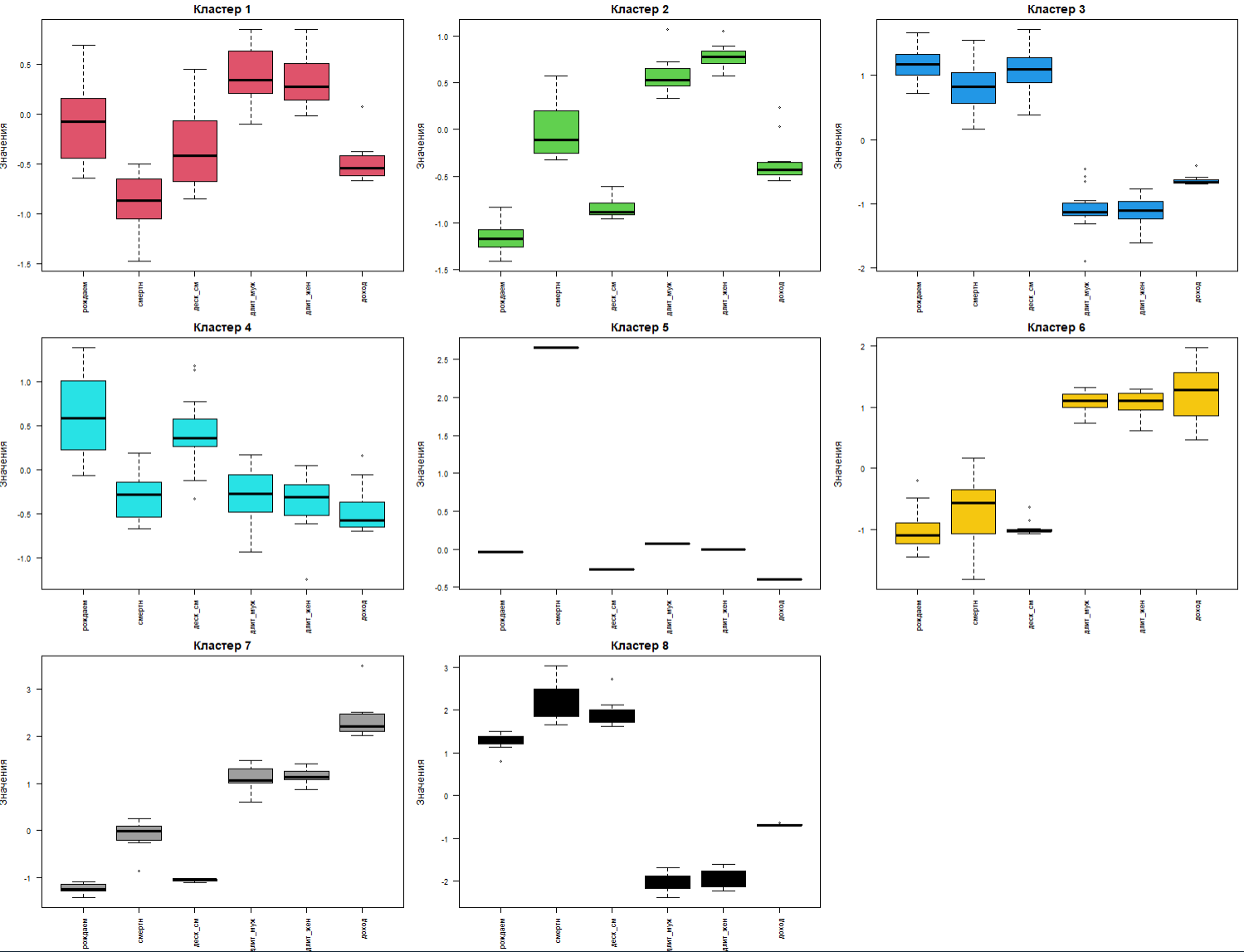


Рисунок 10. Боксплоты кластеров дендрограммы

Проанализировав данные диаграммы и сопоставив их с входящими в кластеры странами, можем следующим образом охарактеризовать каждый из кластеров:

Кластер 1: страны среднего уровня жизни с низкими уровнем доходов и рождаемостью и высокой длительностью жизни. К ним относятся государства из различных регионов, но в основном это страны Южной Америки и Юго-Восточной Азии.

Кластер 2: преимущественно страны Восточной Европы. Отличаются очень низкой рождаемостью, высокой длительностью жизни и доходом ниже среднего.

Кластер 3: бедные страны (преимущественно Африки). Имеют высокие рождаемость и смертность, а также низкие продолжительность жизни и уровень дохода.

Кластер 4: страны среднего уровня жизни с низкими уровнем доходов и длительностью жизни и высокой рождаемостью. К ним относятся прогрессивные страны Африки и некоторые страны Ближнего Востока и Центральной Азии.

Кластер 5: Мексика. Отличается крайне высокой смертностью, связанной с высоким уровнем преступности. В совокупности с низкой детской смертностью и рождаемостью ниже среднего Мексика не определяется ни в 3, ни в 8 кластер, к которым она наиболее близка.

Кластер 6: страны с развитой экономикой. Отличаются высокими доходами и продолжительностью жизни и низкими рождаемостью и смертностью. Данный кластер состоит преимущественно из стран с регионом 3 из исходного набора данных.

Кластер 7: страны с наиболее развитой экономикой. Отличаются очень низкой рождаемостью и очень высоким уровнем дохода. К ним относятся страны Скандинавии, а также Германия, Швейцария и Япония.

Кластер 8: наиболее бедные страны с большой смертностью и очень низкими уровнем дохода и длительностью жизни. Данные этих стран аналогичны кластеру 3, но имеют более экстремальные значения.

6. Выполним кластеризацию набора данных методом К-средних.

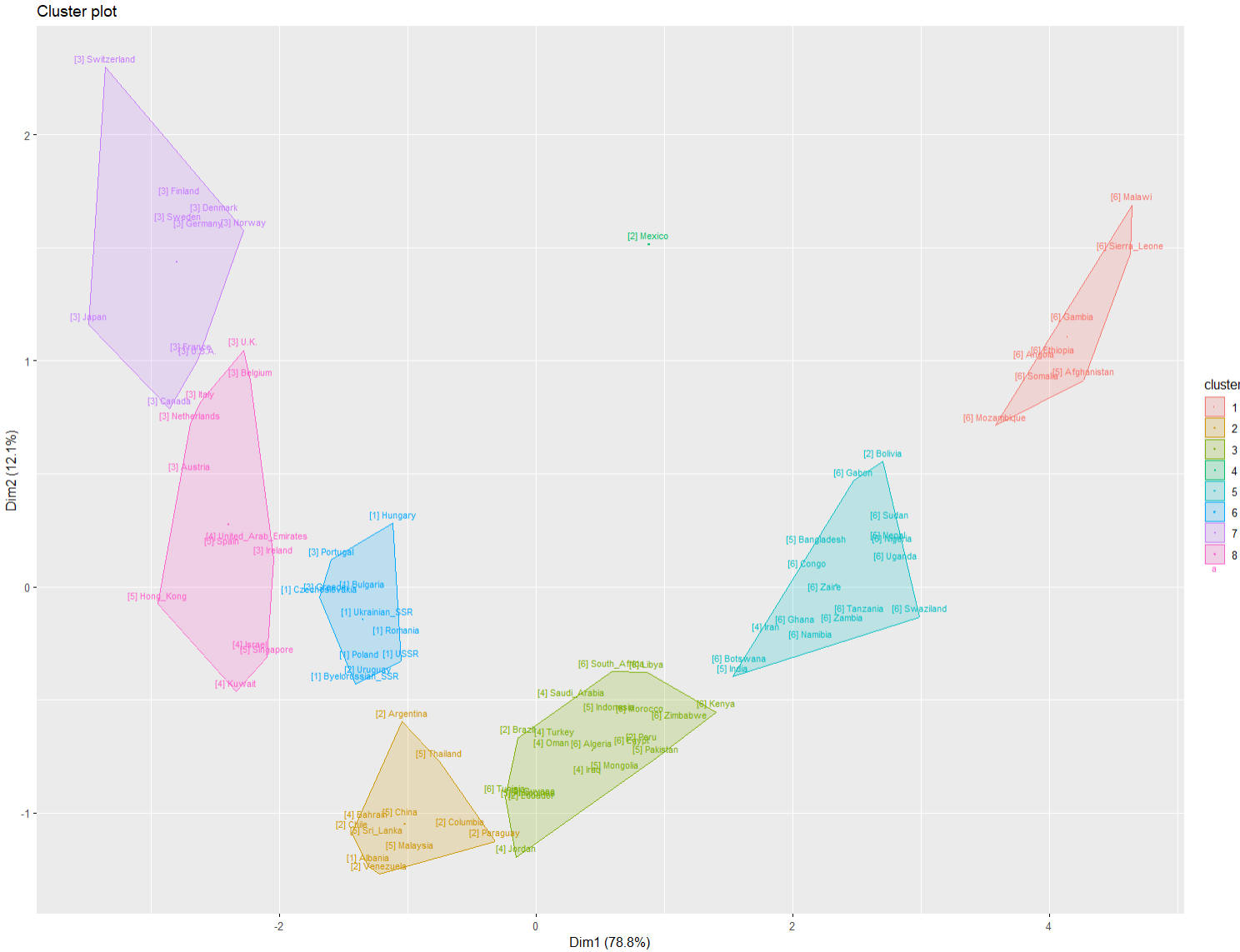


Рисунок 11. Кластеризация методом k-means

Данный метод кластеризации во многом повторяет кластеры, полученные в дендрограмме. Здесь также выделяются наиболее бедные страны, страны Восточной Европы, страны развитой и наиболее развитой экономики, а также хорошо демонстрируется отдалённость Мексики от остальных стран, что доказывает необходимость выделения её в отдельный кластер.

7. Построим scatterplot для кластеров, полученных методом К-средних.

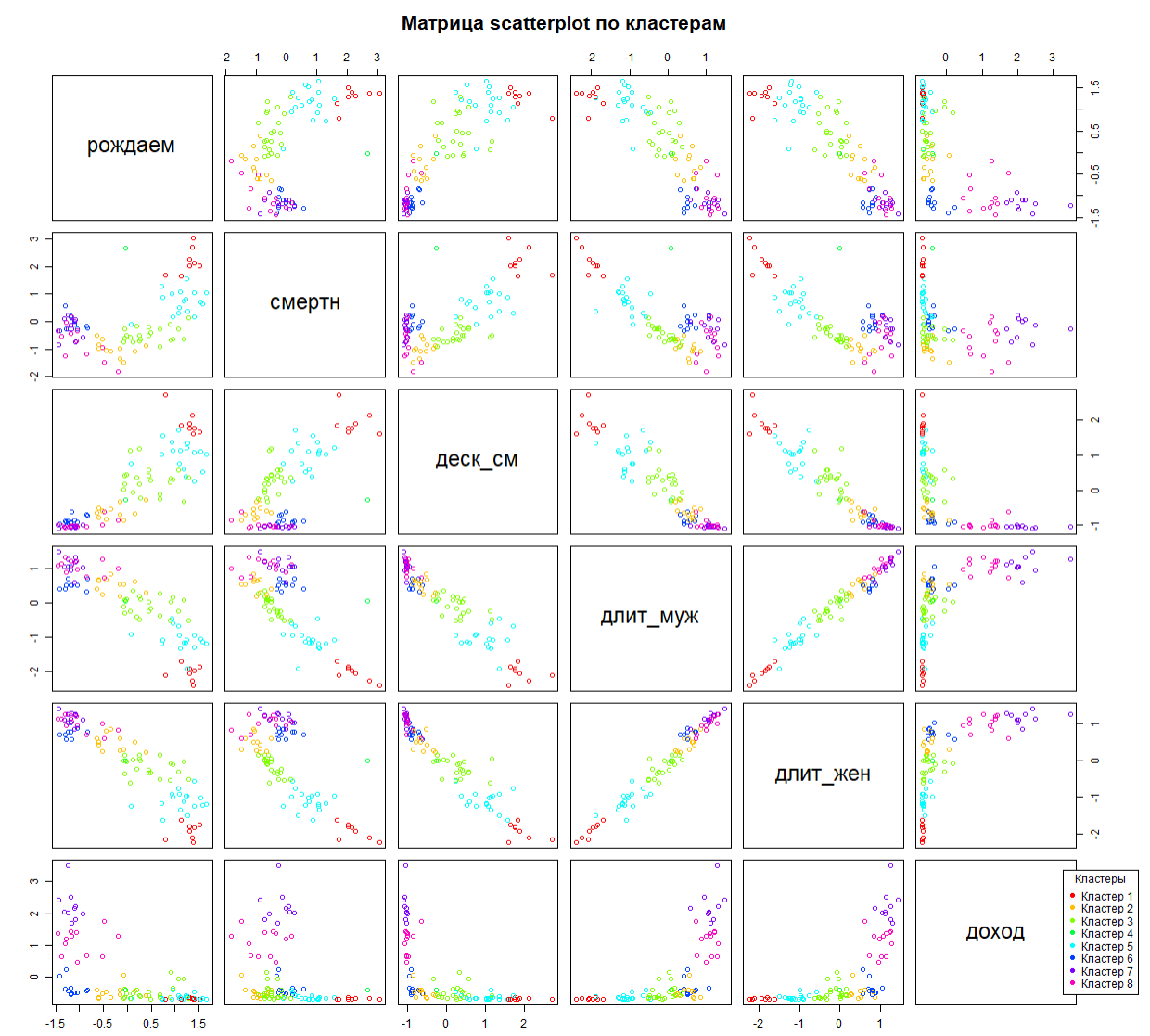


Рисунок 12. Диаграмма рассеяния для кластеров

На графике не наблюдается значительных выбросов и пересечений кластеров, что свидетельствует о корректном их выборе. Заметно значительное отклонение в значениях смертности у Мексики (кластер 4), а также достаточно строгая зависимость между уровнем дохода, детской смертностью и длительностью жизни.

8. Модифицируем начальный датасет, добавив в него вектор с полученными в результате кластеризации кластерами, и разделим его на обучающие и тестовые данные в приблизительной пропорции 7 к 3, которые будем в дальнейшем использовать при классификации.

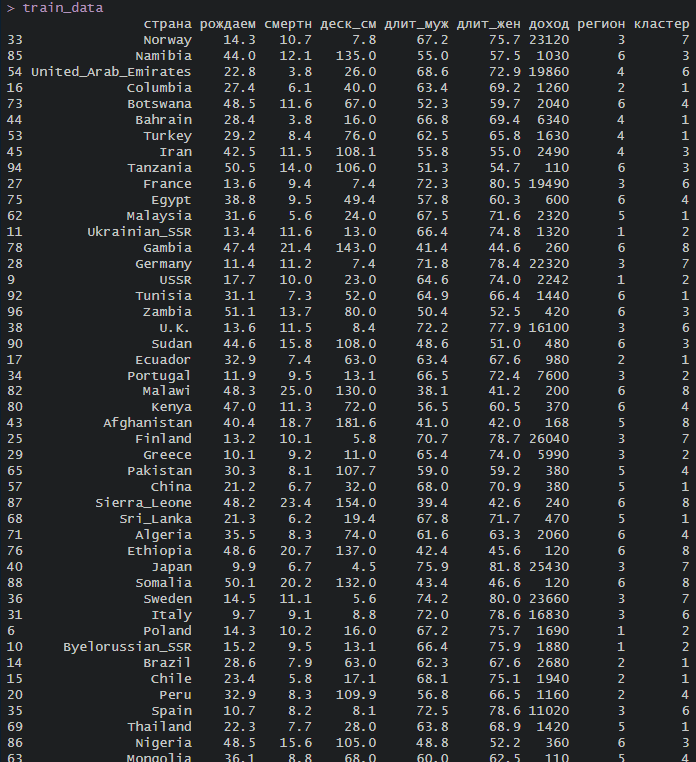


Рисунок 13. Фрагмент обучающих данных датасета

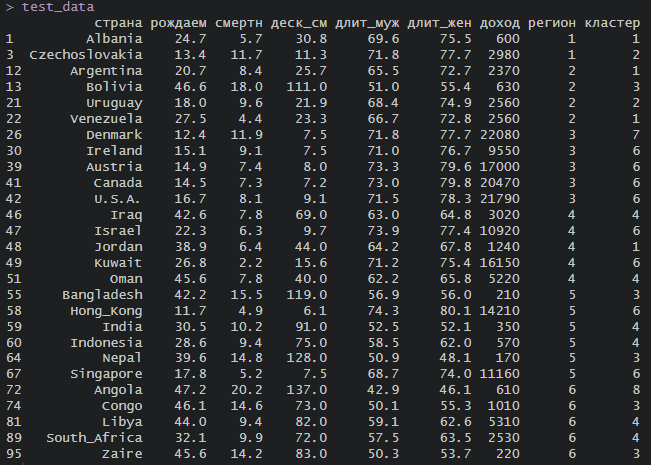


Рисунок 14. Тестовые данные датасета

9. Обучим Байесовский классификатор на обучающих данных. Далее выведем одну из полученных таблиц, например, по параметру *длит\_муж* (средняя длительность жизни среди мужчин).

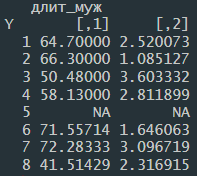


Рисунок 15. Таблица Байесовского классификатора для параметра *длит\_муж*

Первый столбец в таблице отвечает за среднее значение параметра, второй – за его стандартное отклонение. Можно заметить, что классификатор не может вычислить значения для кластера 5, в который входит только Мексика, поскольку значения всех параметров в нём фактически имеют нулевую дисперсию. Самое низкое среднее значение имеет кластер беднейших стран, а самое высокое – кластер стран с наиболее развитой экономикой.

10. Проведём прогноз тестовых данных обученным классификатором и оценим полученный результат.

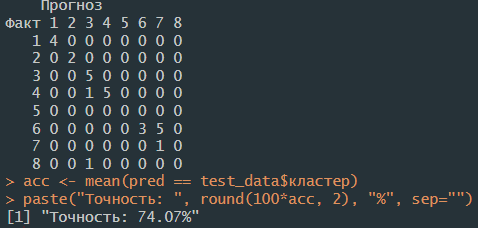


Рисунок 16. Результаты прогноза Байесовского классификатора

Классификатор показывает приемлемый результат с точностью в 74% для относительно маленького набора данных, состоящего из около 30 записей. Наибольшие трудности у него возникают с распредлением между 6 и 7 кластерами (страны с развитой экономикой), а также кластерами 3 и 8 (бедные страны).

11. Выполним обучение классификатора на основе дерева решений и построим график полученного дерева.

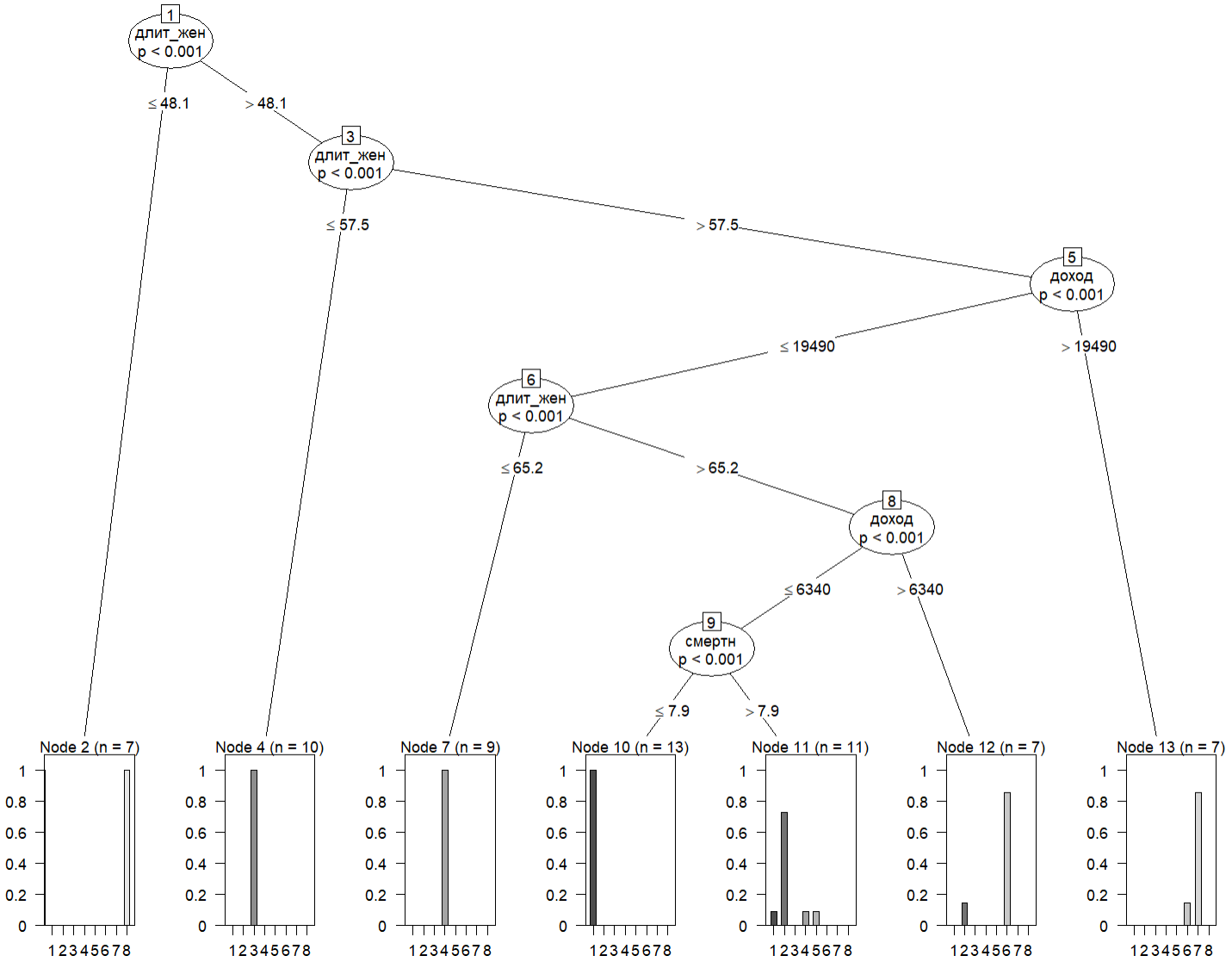


Рисунок 17. График дерева решений

Проанализировав данный график, можно заметить, что дерево решений классифицирует данные в 3, 8 и 4 кластер по низкой средней продолжительности жизни женщин, в 6 и 7 – по высокому уровню доходов, а между 1 и 2 кластером выбирает по уровню смертности.

12. Cделаем прогноз на основе классификатора дерева решений и проверим его точность.

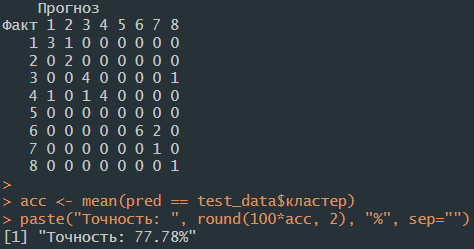


Рисунок 18. Результаты прогноза дерева решений

Дерево решений показывает более хорошие результаты с точностью ~78%. Вместо большого выброса на кластере 6, который присутствовал в прошлом методе, здесь наблюдаются редкие выбросы, но в большем числе кластеров.

13. Выполним обучение классификатора случайного леса для 100 деревьев и проведём на его основе классификацию тестовых данных. Далее проанализируем полученные результаты и сравним с результатами предыдущих методов.

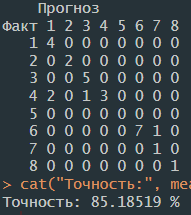


Рисунок 19. Результаты прогноза случайного леса

Сопоставим все полученные результаты классификаторов в таблице:

|  |  |  |  |
| --- | --- | --- | --- |
| Классификатор | Байесовский | Дерево решений | Случайный лес |
| Точность | 74,07% | 77,78% | 85,19% |

Метод случайного леса показывает наилучшие результаты с точностью более 85%. Хуже всего он справляется с кластером 4, но остальные кластеры распределяет с очень высокой точностью.

**Вывод**: в процессе выполнения данной лабораторной работы я изучил методы кластеризации и подбора оптимального числа кластеров, методы классификации и способы их оценки, а также освоил инструменты языка R для кластеризации и классификации и применил их для анализа набора данных с информацией о параметрах, связанных с уровнем жизни различных стран, выполнив в результате кластеризацию исходного датасета, обучение классификаторов и классификацию данных.