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**

**Отчёт лабораторная №5.2**

**Дисциплина: Обработка больших данных**

Работу выполнил \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Логвина А.В.

Направление подготовки 02.03.02 Фундаментальная информатика и

информационные технологии

Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Яхонтов А.А.

Краснодар

2025

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

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

**Задание:**

Вам необходимо добавить найденные 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. Важно: понадобится преобразовать вектор groups в фактор.

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.

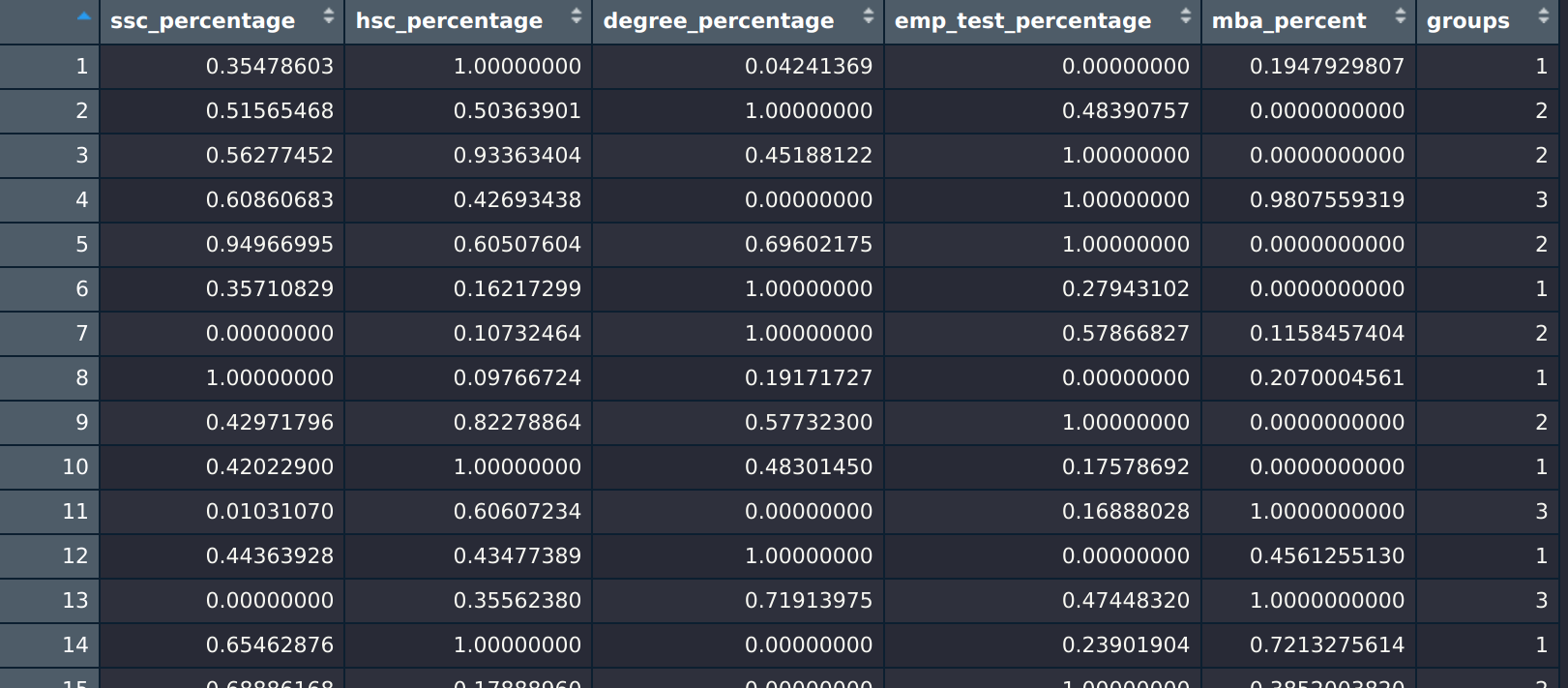
3. Проанализируйте точность полученных решений для тестовых данных (с известным значением переменной отклика), сравните результаты с ранее полученными.

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

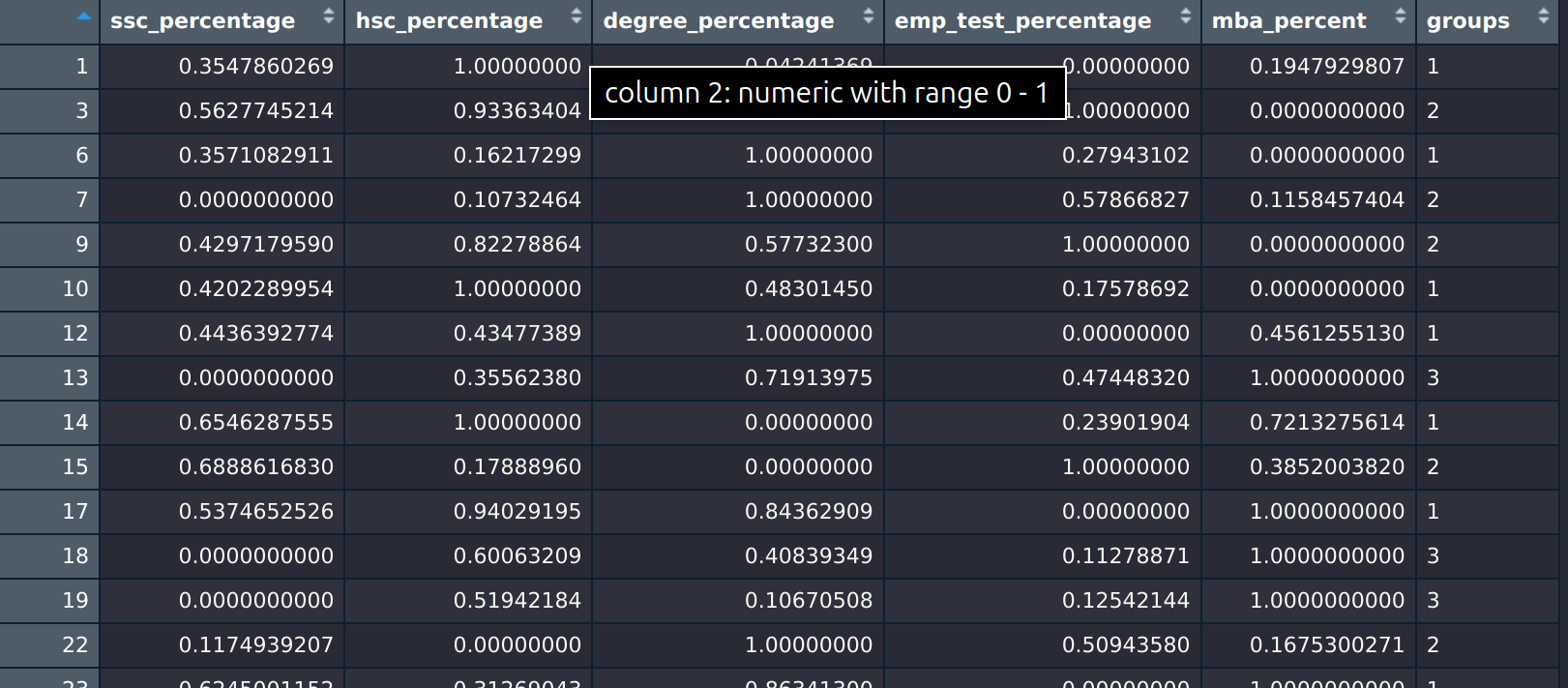
5. Сопоставьте результаты с результатами Байесовского классификатора.

**Ход работы:**

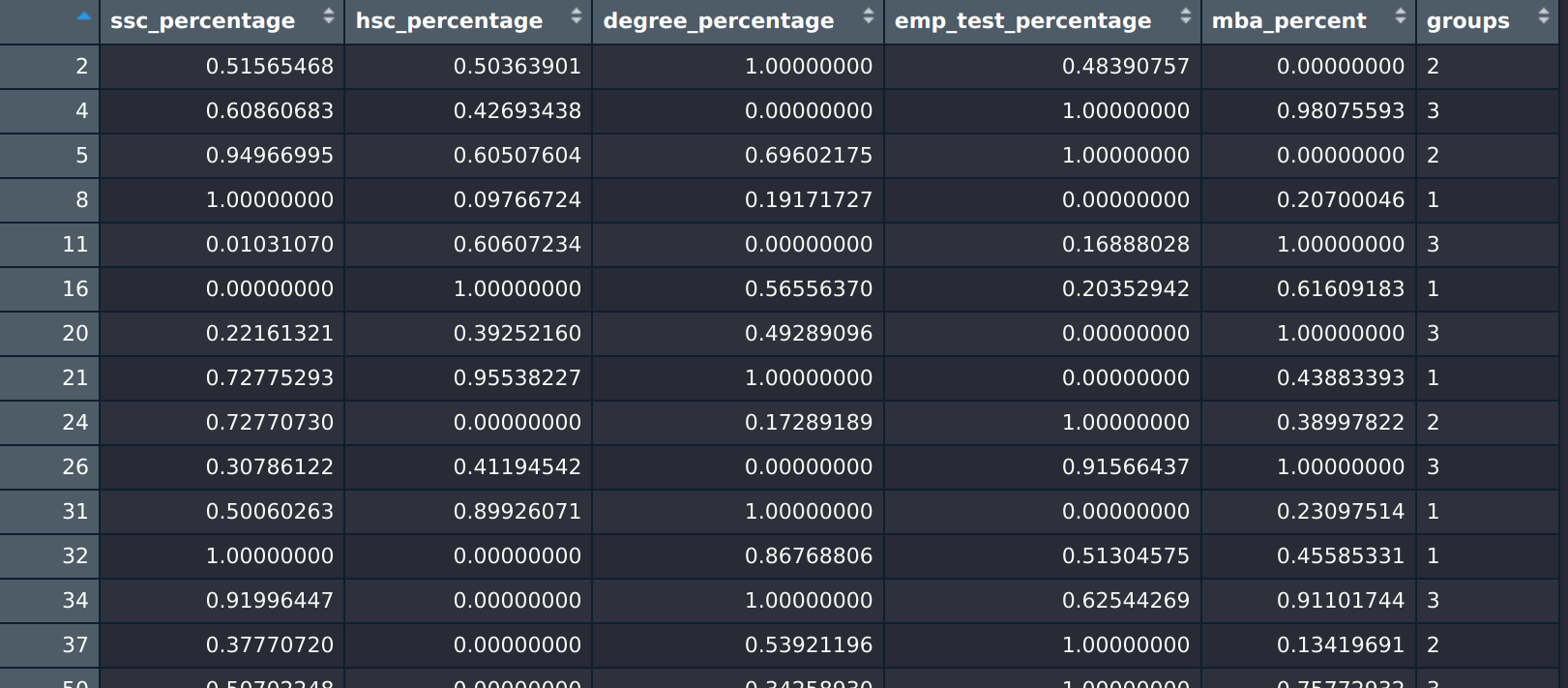
Добавим найденные в результате кластерного анализа классы (groups), как вектор-столбец в предыдущий DATASET, разделить его на две части (обучающую и тестовую) обучить на обучающей выборке классификатор



**Данные для обучения 70% от всех данных:**

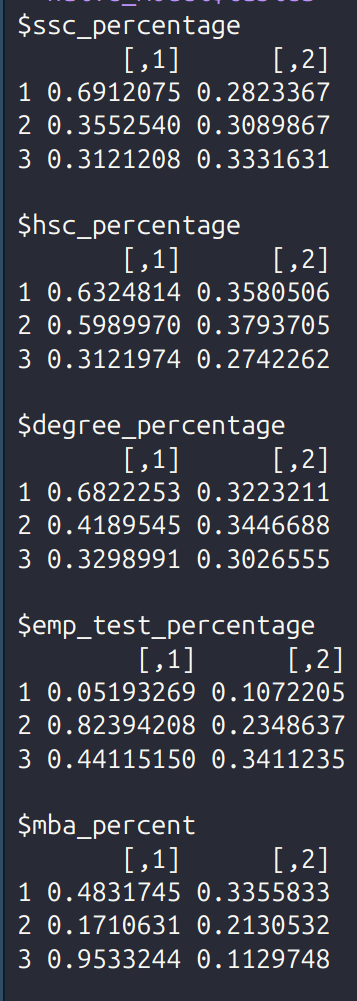


**Данные для тестирования 30% от всех данных:**

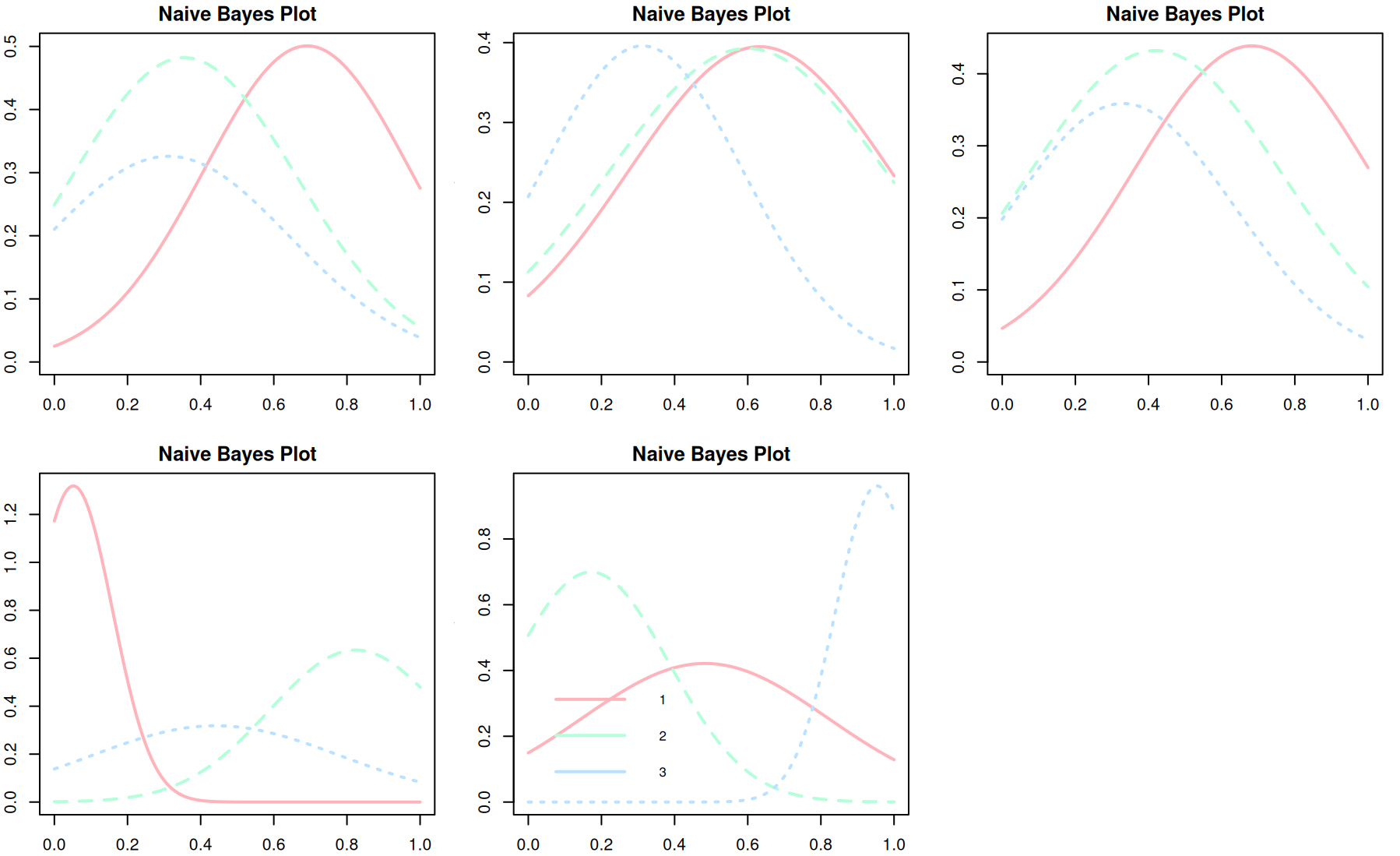


1. Решите задачу с помощью наивного Байесовского классификатора;

Таблицы вероятностей для каждого критерия. Эти таблицы позволяют оценить, насколько вероятно принадлежность нового объекта к каждому классу и выбирать наиболее вероятный.

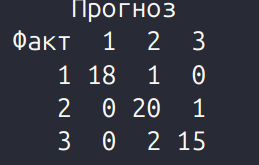


Графики по каждому критерию с классами



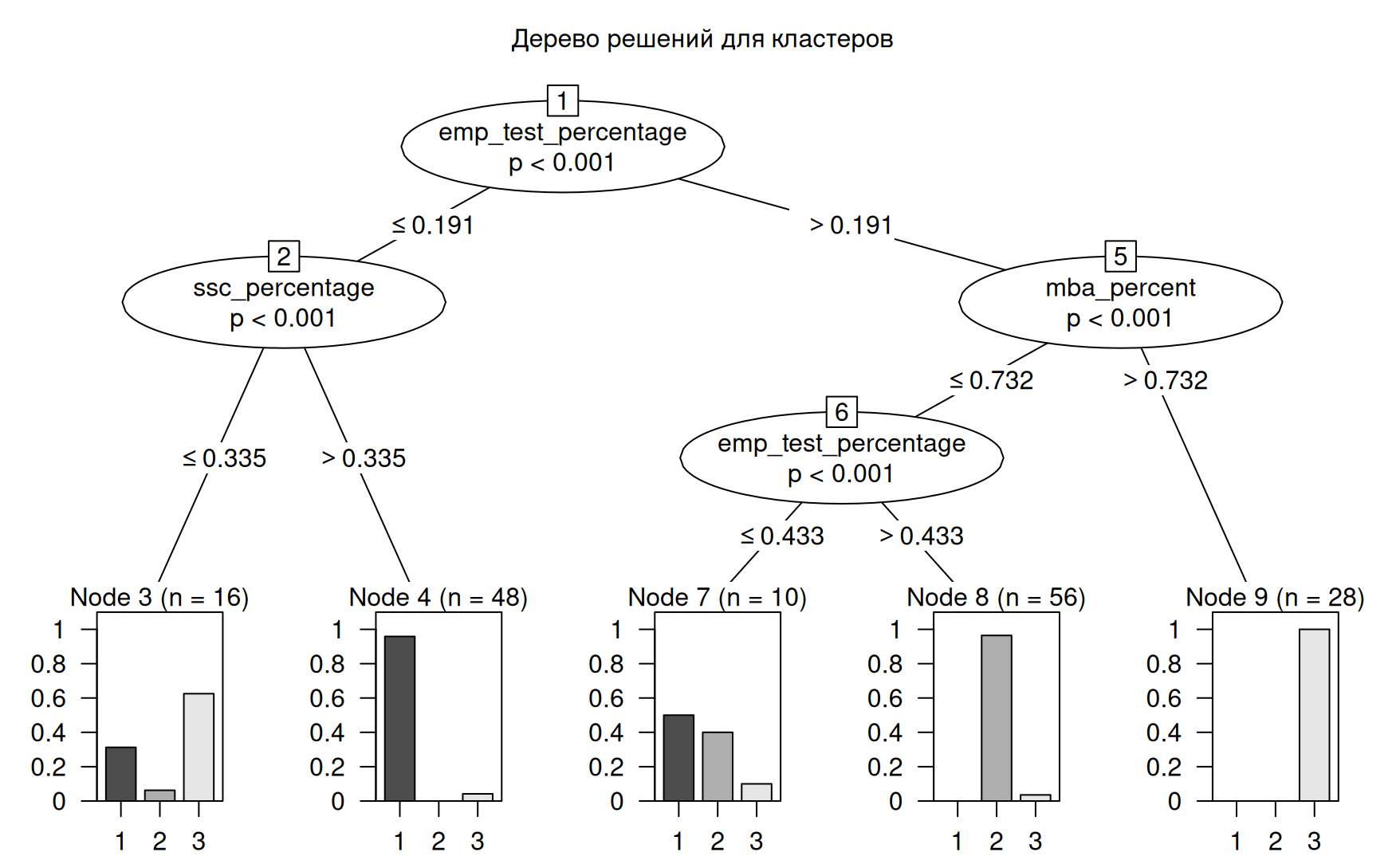
Каждая кривая – это распределение значений признака внутри одного кластера. Чем выше кривая в какой-то области, тем чаще там встречаются объекты из соответствующего кластера.

1. Проанализируйте точность полученных решений для тестовых данных

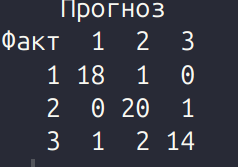


Матрица ошибок сравнивает фактические значения классов, то есть, к какому кластеру действительно принадлежит объект, и предсказанные моделью значения – те, к какому кластеру модель отнесла объект.

1. Решения деревьев.



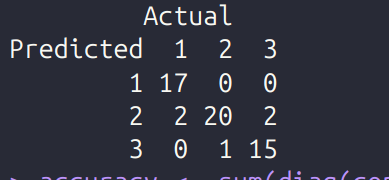
Дерево решений – модель, которая делит пространство признаков на несколько областей, основываясь на условиях для каждого признака.



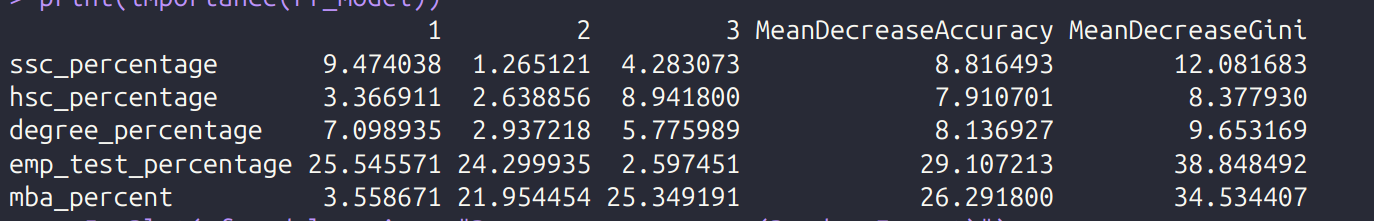
Точность дерева решений: 0.912

Точность наивного байеса: 0.93

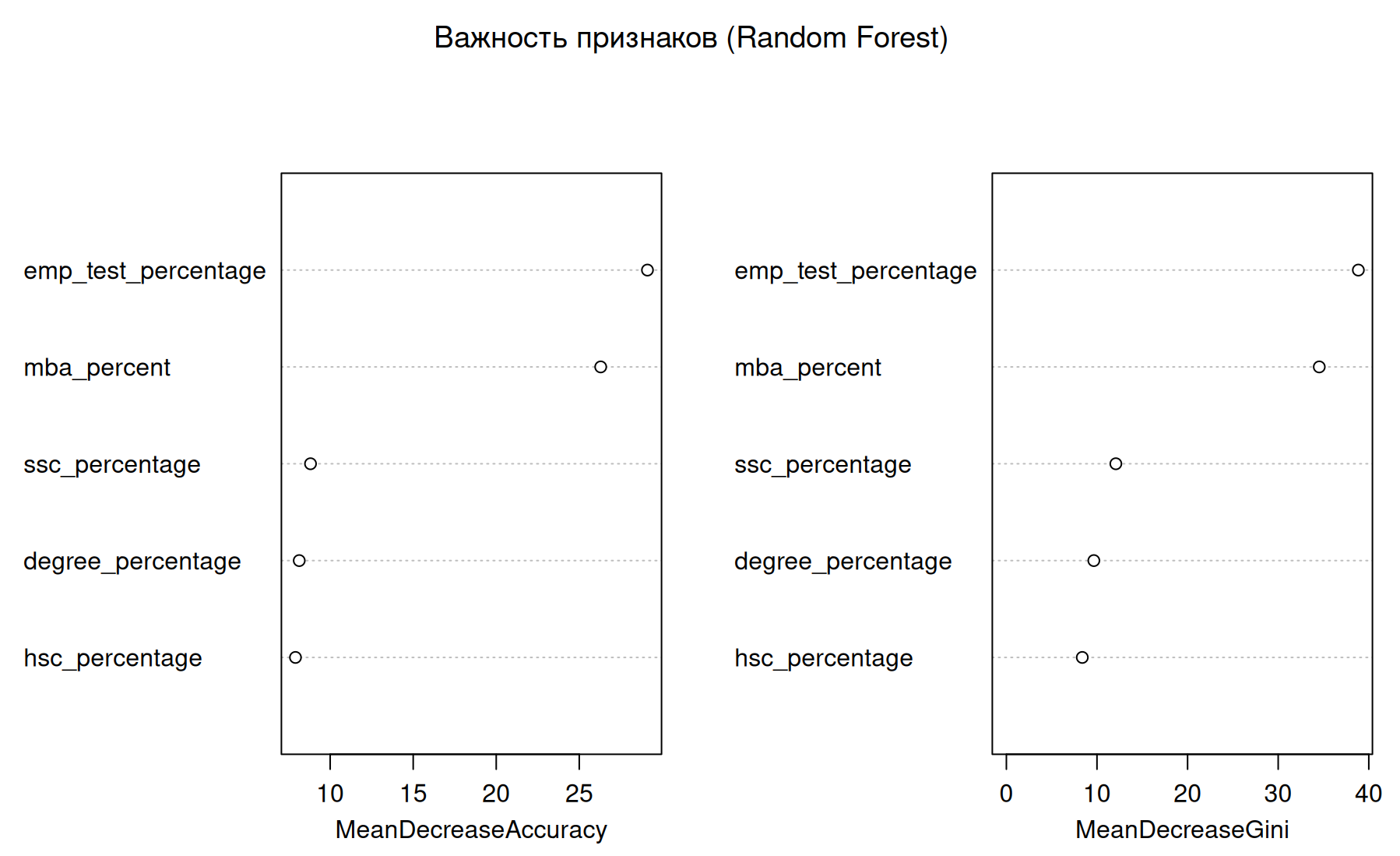
Выполнить классификацию с помощью случайного леса, сопоставить результат с результатом дерева решения, прокомментировать результат сравнения.



Точность модели случайного леса: 0.912



Здесь выводится важность каждого признака для классификации в модели случайного леса. Есть 2 основные метрики MeanDecreaseAccuracy (показывает, насколько бы уменьшилась точность модели, если бы этот признак был случайным образом перемешан) и MeanDecreaseGini (как часто признак использовался для разделения данных в деревьях).



**Вывод: я** научилась выполнять классификацию на основе формулы Байеса и деревьев решений**.**