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**

по дисциплине «Обработка больших данных»

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

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

Краснодар 2023

Исходный код: <https://github.com/SnowLukin/BigData/tree/main>

**Набор данных**

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

Краткая 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).

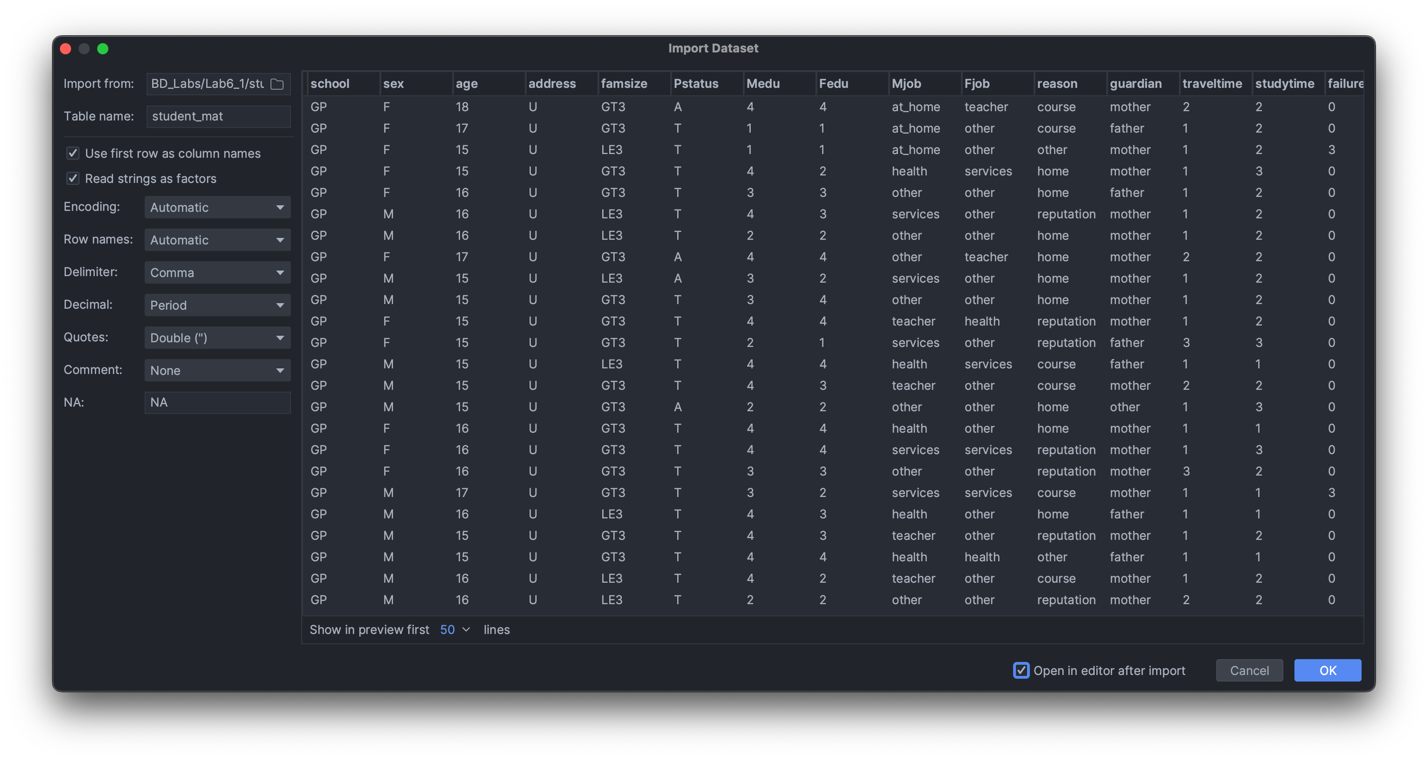
****

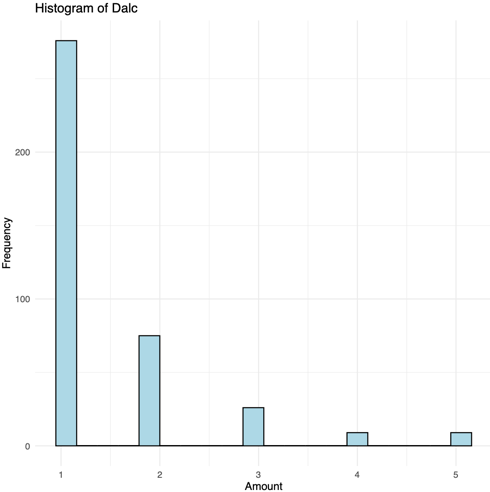
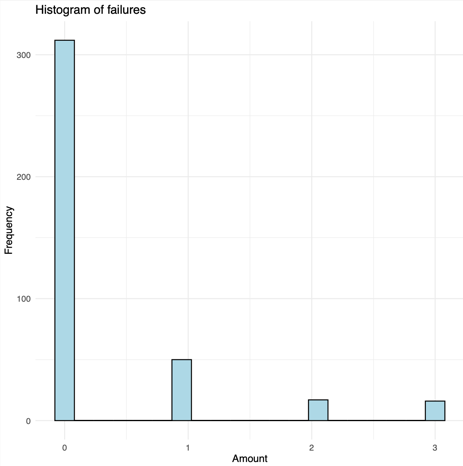
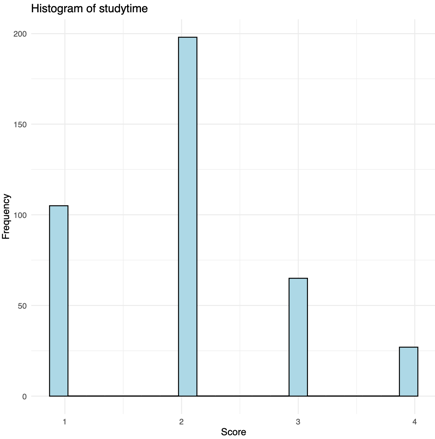
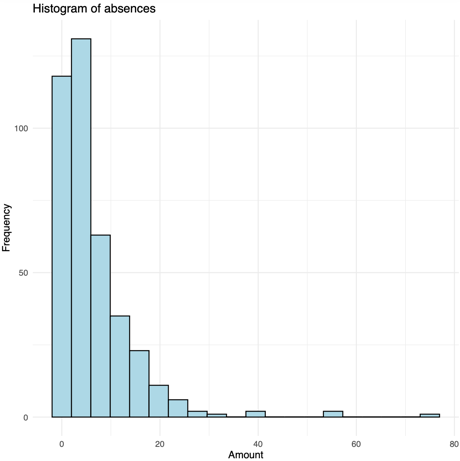
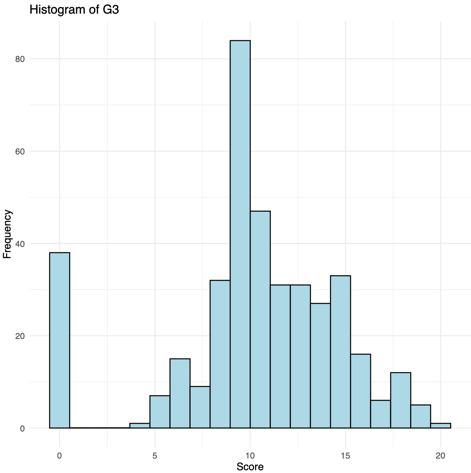
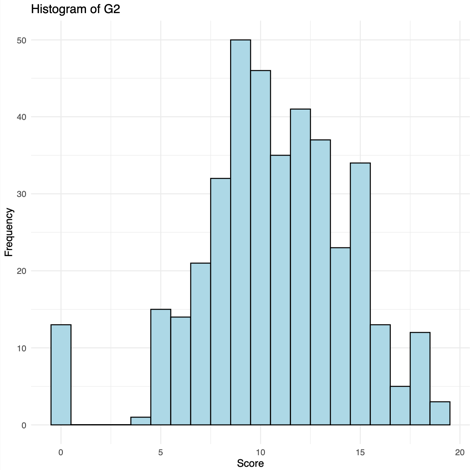
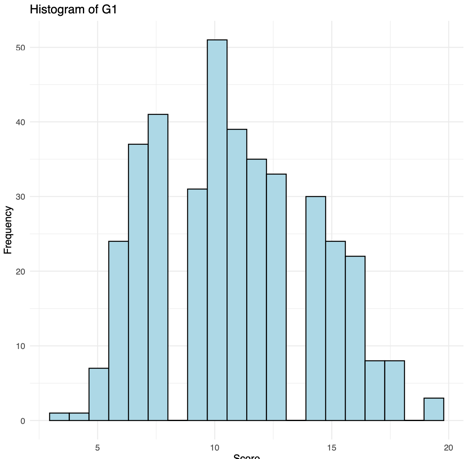
Рисунок 1 – Используемый набор данных

**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 выделить 5 основных кластеров.

Diagram

Description automatically generated

Рисунок 2 – Результат построения дендограммы с выделением выбранных кластеров

Chart, line chart

Description automatically generated

Рисунок 3 – Метод Локтя

Диаграмма метода Локтя подтверждает правильность выбора количества кластеров для рассмотрения.

**Характеристики кластеров**

Chart, bar chart

Description automatically generated

Рисунок 4 – Графическое представление средних характеристик кластеров

Данный график хорошо показывает отличающиеся характеристики кластеров. Каждый кластер характеризуется разными значениями возраста, уровня образования родителей (Medu и Fedu), времени на дорогу до школы, времени учебы, количества неудач, отношений в семье, свободного времени, прогулок с друзьями, употребления алкоголя в будние дни (Dalc). и выходные (Walc), состояние здоровья, количество пропусков и оценки (G1, G2 и G3).

**Диаграмма "Каменная осыпь"**

Chart, line chart

Description automatically generated

Рисунок 5 – Результат построения диаграммы «Каменная осыпь»

График показывает резкое падение дисперсии, объясненное после первых нескольких групп роста, причем первые несколько групп объясняют большую часть дисперсии. Это указывает на то, что в наборе данных несколько доминирующих факторов или основных компонентов, которые объясняют большую часть изменчивости данных.

**Построение scatterplot**

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Рисунок 6 – Графики scatterplot

Chart, scatter chart

Description automatically generated

Рисунок 7 – 3D График scatterplot для оценок G1, G2, G3

**Вывод по первой части**

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

Кластер 1: в этой группе средний возраст 16,6 лет и относительно высокий уровень образования родителей. У них хорошие семейные отношения, они проводят меньше времени с друзьями и имеют низкий уровень потребления алкоголя. Их оценки выше по сравнению с другими группами, со средним баллом около 8,5 в G1 и G2 и 7,5 в G3.

Кластер 2: эта группа имеет более низкий уровень образования родителей, большее количество неудач и более высокое потребление алкоголя, чем кластер 1. У них более низкий средний балл по сравнению с кластером 1, со средним баллом 7 в G1 и G2 и 5 в G3.

Кластер 3: Эта группа имеет самый высокий уровень образования родителей и проводит больше всего времени за учебой. У них низкий уровень потребления алкоголя и мало прогулов, что приводит к высоким оценкам со средним баллом около 14,5 в G1, G2 и G3.

Кластер 4: Эта группа имеет относительно более низкий уровень образования родителей, тратит меньше времени на учебу и имеет более высокий уровень потребления алкоголя. У них более низкие оценки по сравнению с кластером 3 и кластером 1, со средним баллом около 11 в G1, G2 и G3.

Кластер 5: Эта группа имеет высокий уровень потребления алкоголя, проводит больше времени с друзьями и чаще всего отсутствует. У них самый низкий средний балл по сравнению с другими кластерами со средним баллом около 10 в G1, G2 и G3.

**Наивного Байесовского классификатора**

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

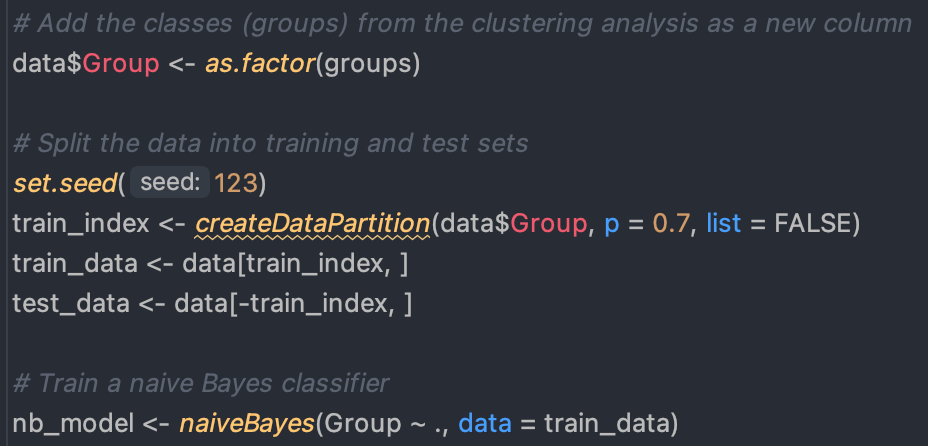


Рисунок 7 – Реализация наивного байесовского классификатора

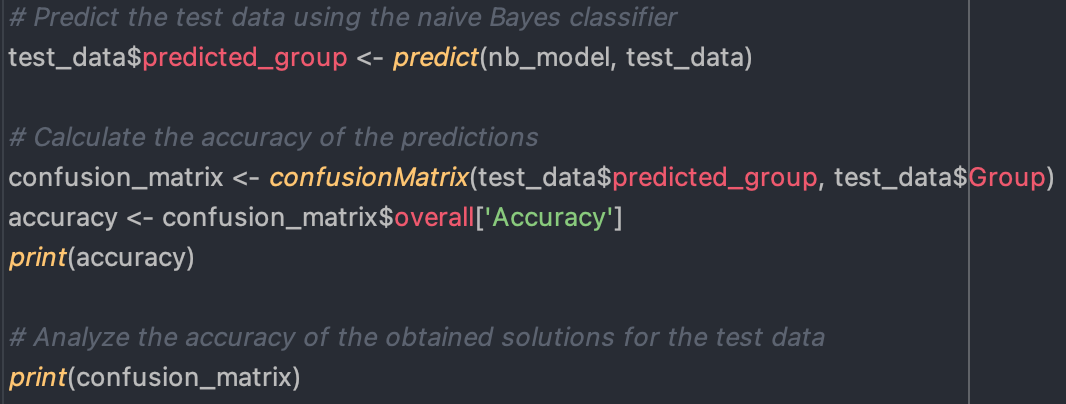


Рисунок 8 – Реализация определения кластера и вывод результатов

A picture containing calendar

Description automatically generated

Рисунок 9 – Результаты определения кластера и матрица ошибок

В целом, наивный байесовский классификатор достиг 75%-ной точности. Это означает, что из всех экземпляров в наборе данных классификатор правильно предсказал метку кластера для 75% из них.

Глядя на матрицу путаницы, мы можем видеть, что классификатор работал хорошо для некоторых кластеров, но не так хорошо для других. Например, он правильно классифицировал 28 экземпляров как принадлежащие кластеру 1, но также неправильно классифицировал 9 экземпляров из других кластеров как принадлежащие кластеру 1. С другой стороны, он правильно классифицировал 27 экземпляров как принадлежащие кластеру 3, и только неправильно классифицировал 4 экземпляра из других кластеров как принадлежащие кластеру 3.

Мы также можем видеть, что некоторые кластеры легче различимы, чем другие. Например, кластер 2 очень легко отличить от других, поскольку в нем не было неправильных классификаций по сравнению с другими кластерами. С другой стороны, кластер 1 имел наибольшее количество неправильных классификаций по сравнению с другими кластерами, что указывает на то, что он может быть более похож на другие кластеры с точки зрения признаков, используемых классификатором.

**Метод деревьев решений**

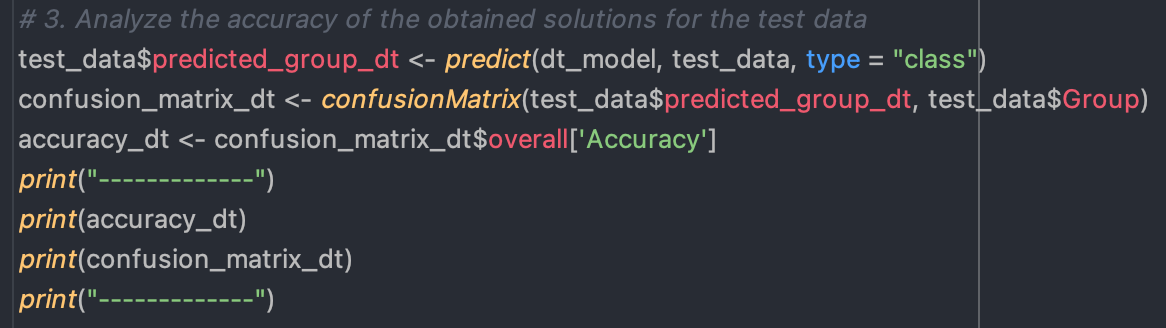


Рисунок 10 – Реализация метода деревьев решений

Calendar

Description automatically generated with medium confidence

Рисунок 11 – Результаты метода деревьев решений и матрица ошибок

Diagram

Description automatically generated

Рисунок 14 – Графическое представление дерева решений

Общая точность модели составляет 63,79%. Хотя это лучше, чем случайное угадывание (которое составило бы около 20% для задачи классификации по 5 классам), это не очень высокая точность.

Из матрицы путаницы мы можем наблюдать следующее:

Класс 1: 26 образцов классифицированы правильно, но имеются значительные ошибки в классификации: 10 образцов ошибочно отнесены к классу 4, а 8 образцов - к классу 5.

Класс 2: 7 образцов классифицированы правильно, и только 1 образец ошибочно отнесен к классу 1.

Класс 3: 28 образцов классифицированы правильно, но есть некоторые ошибки в классификации, при этом 5 образцов ошибочно отнесены к классу 1, 3 - к классу 4 и 3 - к классу 5.

Класс 4: 4 образца классифицированы правильно, но имеются ошибочные классификации, при этом 3 образца ошибочно отнесены к классу 1, а 3 - к классу 3.

Класс 5: 9 образцов классифицированы правильно, при этом 3 образца ошибочно отнесены к классу 4.

Из матрицы путаницы очевидно, что модели сложнее точно классифицировать выборки по классам 1, 4 и 5. Сосредоточение внимания на улучшении производительности для этих классов может привести к общему повышению точности. Этого можно достичь путем более глубокого анализа особенностей, влияющих на эти классы, или изучения других методов классификации, которые могли бы лучше обрабатывать эти классы.

Text

Description automatically generated

Рисунок 12 – Реализация Random Forest

Calendar

Description automatically generated with low confidence

Рисунок 13 – Результат Random Forest и матрица ошибок

Chart, bar chart

Description automatically generated

Рисунок 15 – Графическое представление важности параметров при классификации в модели Random Forest

Точность:

Общая точность модели случайного леса составляет 68,10%, что является улучшением по сравнению с классификатором дерева решений (63,79%). Несмотря на достигнутый прогресс, все еще есть возможности для улучшения с помощью настройки гиперпараметров, разработки функциональных возможностей или даже изучения других алгоритмов классификации.

Из матрицы путаницы мы можем наблюдать следующее:

Класс 1: 31 образец классифицирован правильно, но имеются значительные ошибки в классификации: 9 образцов ошибочно отнесены к классу 4, а 9 образцов - к классу 5.

Класс 2: 6 образцов классифицированы правильно, без каких-либо ошибок.

Класс 3: 24 образца классифицированы правильно, но есть некоторые ошибки в классификации, при этом 2 образца ошибочно отнесены к классу 1, 1 - к классу 4 и 1 - к классу 5.

Класс 4: 8 образцов классифицированы правильно, но имеются ошибочные классификации, при этом 2 образца ошибочно отнесены к классу 1, 1 - к классу 2 и 2 - к классу 3.

Класс 5: 10 образцов классифицированы правильно, при этом 1 образец ошибочно отнесен к классу 3 и 2 - к классу 4.

Из матрицы путаницы видно, что модели randomForest по-прежнему сложнее точно классифицировать выборки в классах 1, 4 и 5. Однако производительность улучшилась по сравнению с классификатором дерева решений. Дальнейшая работа по улучшению характеристик для этих классов может привести к общему повышению точности.

**Вывод по второй части**

Основываясь на точности, наивный байесовский алгоритм работает лучше всего (75,00%), за ним следует случайный лес (68,10%), а затем дерево решений (63,79%). Матрицы путаницы показывают, что у каждого классификатора есть некоторые проблемы с неправильной классификацией. Однако наивный байесовский классификатор, по-видимому, имеет лучший баланс с точки зрения правильной классификации по всем классам. Классификатор случайного леса показывает улучшения в некоторых классах по сравнению с деревом решений, но он по-прежнему испытывает трудности с классами 1, 4 и 5.