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 №6**

**Дисциплина: Обработка больших данных**

**Тема: «Задачи классификации и кластеризации»**

Работу выполнил: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Курбатский В.А.

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Направленность (профиль): Компьютерные технологии и прикладная математика.

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Приходько Т. А.

Краснодар

2023

Вариант 17

**Постановка задачи**

1. Выполнить дескриптивный анализ данных (здесь приветствуются дополнительные исследования).
2. Выполнить иерархическую кластеризацию Вашего набора данных, построив дендрограмму. Подробно обосновать Ваш выбор числа групп.
3. Построить диаграмму "Каменная осыпь"
4. Выполнить построение scatterplot (рис. 6.10) с помощью ggplot2, это действие эквивалентно кластеризации k-means.
5. В целом: выполнить шаги 1-6 анализа для своего набора данных (если какие-то из шагов нерелевантны вашему набору данных, объяснить почему).

**Информация о наборе:**

Набор данных взят из репозитория машинного обучения UCL. Ссылка на набор данных

Источник:

Пауло Кортес, университет Минью, Гимарайнш, Португалия,

П. Кортес и А. Сильва. Использование интеллектуального анализа данных для прогнозирования успеваемости учащихся средней школы. В ред. А. Брито и Ж. Тейшейры. Труды 5-го будущего бизнес-технологии конференция (FUBUTEC 2008), стр. 5-12, Порту, Португалия, апрель, 2008, EUROSIS, номер ISBN 978-9077381-39-7.

веб-ссылка http://www3.dsi.uminho.pt/pcortez

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). В работе [Cortez and Silva, 2008] эти два набора данных были смоделированы в рамках задач двоичной/пятиуровневой классификации и регрессии. Важное примечание: целевой атрибут G3 имеет сильную корреляцию с атрибутами G2 и G1. Это происходит потому, что G3-это итоговая оценка года (выданная в 3-м периоде), в то время как G1 и G2 соответствуют 1-му и 2-му классам периода. Предсказать G3 без G2 и G1 сложнее, но такое предсказание гораздо полезнее (Подробнее см. Источник статьи).

**Атрибутивная информация:**

школа - школьник (двоичный: 'ГП' - Габриель Перейра или 'MS' - Mousinho да Силвейра)

секс студент секс (двоичное: 'Ф' - женщина или " М " - мужчина)

возраст-возраст студента (числовой: от 15 до 22 лет)

адрес-тип домашнего адреса студента (двоичный: 'U' - городской или 'R' - сельский)

famsize-размер семейства (двоичный: 'LE3' - меньше или равно 3 или 'GT3' - больше 3)

Pstatus - родительского сожительства состояния (двоичное: 'Т' - жить вместе или " А " - отдельно)

Medu-образование матери (числовое: 0-Нет, 1-начальное образование (4-й класс), 2-5-9-й класс, 3-среднее образование или 4-высшее образование)

Феду-образование отца (числовое: 0-Нет, 1-начальное образование (4 класс), 2-5-9 класс, 3-среднее образование или 4-высшее образование)

Mjob - матери работа (номинальное: "учитель", "здоровье" медицинских услуг, связанных, гражданской "услуги" (например, административные или полицейские), 'at\_home' или 'другое')

Fjob-работа отца (номинально: "учитель", "здравоохранение", связанные с ним гражданские "услуги" (например, административные или полицейские), "at\_home" или "другое")

причина-причина выбрать эту школу (номинально: близко к "дому", школе "репутация", "курс" предпочтения или "другое")

опекун-опекун студента (номинально: "мать", "отец" или "другой")

время в пути - время в пути от дома до школы (числовое: 1-1 час)

studytime-еженедельное время обучения (числовое: 1-10 часов)

отказы-количество прошлых отказов класса (числовое: n, если 1<=n<3, иначе 4)

schoolsup-дополнительная образовательная поддержка (двоичный код: да или нет)

famsup-семейная образовательная поддержка (двоичный код: да или нет)

платные-дополнительные платные занятия в рамках предмета курса (математика или португальский язык) (двоичный код: да или нет)

мероприятия-внеклассные мероприятия (бинарные: да или нет)

детский сад-посещаемый детский сад (двоичный код: да или нет)

высшее - хочет получить высшее образование (бинарное: да или нет)

интернет - доступ в интернет дома (двоичный код: да или нет)

романтика - с романтическими отношениями (бинарно: да или нет)

фамрел-качество семейных отношений (числовое: от 1-очень плохо до 5-отлично)

freetime-свободное время после школы (числовое значение: от 1-очень низкий до 5-очень высокий)

goout-выход с друзьями (числовое значение: от 1-очень низкий до 5-очень высокий)

Dalc-потребление алкоголя в течение рабочего дня (числовое значение: от 1-очень низкое до 5-очень высокое)

Walc-потребление алкоголя в выходные дни (числовое значение: от 1-очень низкое до 5-очень высокое)

здоровье-текущее состояние здоровья (числовое значение: от 1-очень плохо до 5-очень хорошо)

прогулы - количество школьных прогулов (числовое значение: от 0 до 93)

эти оценки связаны с предметом курса, математикой или португальским языком:

G1-класс первого периода (числовой: от 0 до 20)

G2-класс второго периода (числовой: от 0 до 20)

G3-итоговая оценка (числовая: от 0 до 20, выходная цель)

**Ход работы:**

Проведем дескриптивный анализ.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунки 1–2 – Значение функции summary для столбцов с числовыми значениями.

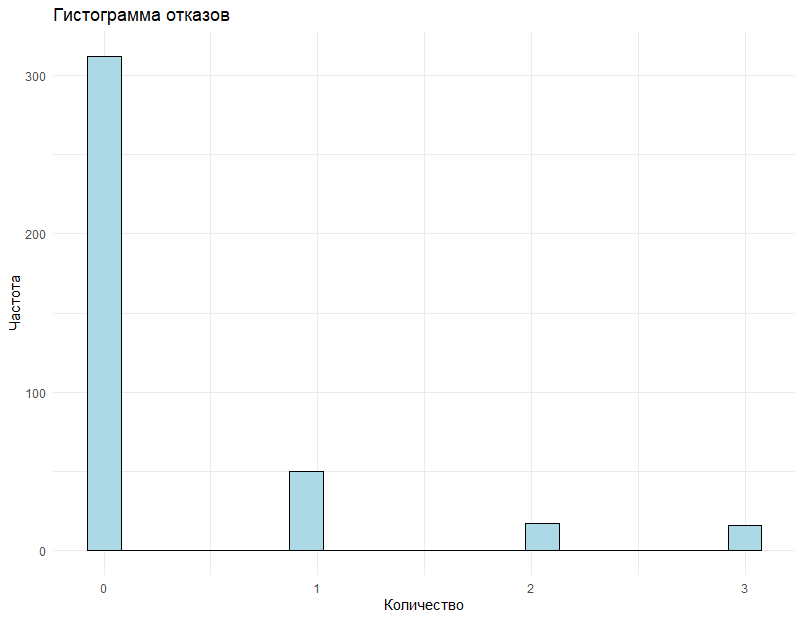
Представим наиболее значимые для школьной успеваемости параметры в виде графиков.

Изображение выглядит как диаграмма, текст, снимок экрана, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, диаграмма, линия

Автоматически созданное описание



Рисунки 3–5 – Гистограммы некоторых параметров

Процесс кластеризации

Изображение выглядит как диаграмма, зарисовка, Технический чертеж, текст

Автоматически созданное описание

Рисунок 6 – Дендрограмма

При построении дендрограммы учитывались следующие показатели:

* Промежуточная аттестация
* Возраст
* Время пути до учебного заведения
* Время уходящие на обучение
* Коэффициент свободного времени
* Количество прогулов
* Образование матери
* Образование отца
* Работа отца
* Работа матери
* Состояние здоровья
* Потребление алкоголя в рабочие дни
* Опекун
* Потребление алкоголя в выходные дни

Изображение выглядит как текст, линия, диаграмма, График

Автоматически созданное описание

Рисунок 7 – Метод «Локтя» для помощи в определении кластеров.

Отсюда видим, что у нас 4–5 кластера. Выберем 5.

Изображение выглядит как текст, диаграмма, снимок экрана, График

Автоматически созданное описание

Рисунок 8 – Дендрограмма с выделенными кластерами.

Изображение выглядит как текст, снимок экрана, График, Красочность

Автоматически созданное описание

Рисунок 9 – Группы учеников

Изображение выглядит как диаграмма, линия, График, текст

Автоматически созданное описание

Рисунок 10 – Диаграмма 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.

Изображение выглядит как текст, снимок экрана, диаграмма, линия

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, диаграмма, линия

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, число, диаграмма

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, число, линия

Автоматически созданное описаниеИзображение выглядит как текст, число, снимок экрана, График

Автоматически созданное описаниеИзображение выглядит как текст, снимок экрана, диаграмма, линия

Автоматически созданное описание Изображение выглядит как текст, снимок экрана, число, линия

Автоматически созданное описание

Рисунки 11–17 – Диаграммы рассеивания

Изображение выглядит как текст, диаграмма, линия, снимок экрана

Автоматически созданное описание

Рисунок 18 – 3Д диаграмма рассеивания

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.

**Часть 2**

**Постановка задачи:**

Данные берутся из предыдущего набора (ЛР 6 часть 1). 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.

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.

6. Проанализируйте полученные результаты, сделайте вывод относительно результатов кластерного анализа (ЛР 6.1), оформите общий отчёт по ЛР 6.1 и 6.2.

**Ход работы:**

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

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок 19 – Результаты определения кластера

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.

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.

Реализуем метод деревьев решений.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок 20 – Результаты метода деревьев решений и матрица ошибок

Изображение выглядит как текст, диаграмма, План, снимок экрана

Автоматически созданное описание

Рисунок 21 – Дерево 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.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рисунок 22 – Реализация метода случайного леса

Изображение выглядит как текст, диаграмма, линия, График

Автоматически созданное описание

Рисунок 23 – Графическое представление важности параметров при классификации в модели случайного леса.

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:**

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.