ML. Лабораторная работа №2.

Задача классификации.

Студент: Велиев Рауф Рамиз оглы

Группа: М8О-309Б-23

Цель и задачи

Цель:

- предсказать exam_score по поведенческим и демографическим признакам.

Задачи:

- подготовить датасет
- исследовать распределения
- обучить Decision Tree
- визуализировать результаты

Датасет и признаки

Файл: student_habits_performance.csv
Объём: 1000 наблюдений, 16 признаков
Цель: exam_score (регрессия)

Фичи: study_hours_per_day, social_media_hours, sleep_hours, attendance_percentage, exercise_frequency, mental_health_rating, netflix_hours, др.

Предобработка: очистка NaN/inf, простые новые признаки, train/test split 80/20.

Дереву масштабирование не нужно — поэтому шаг стандартизации не обязателен.

Модель Decision Tree

Decision Tree Classifier (дерево решений) — это алгоритм классификации и регрессии, который принимает решения, разбивая данные по признакам в виде дерева с узлами и ветвями.
Каждый узел дерева соответствует условию на значение признака (например, «возраст > 30?»), а листья — это предсказанные классы.

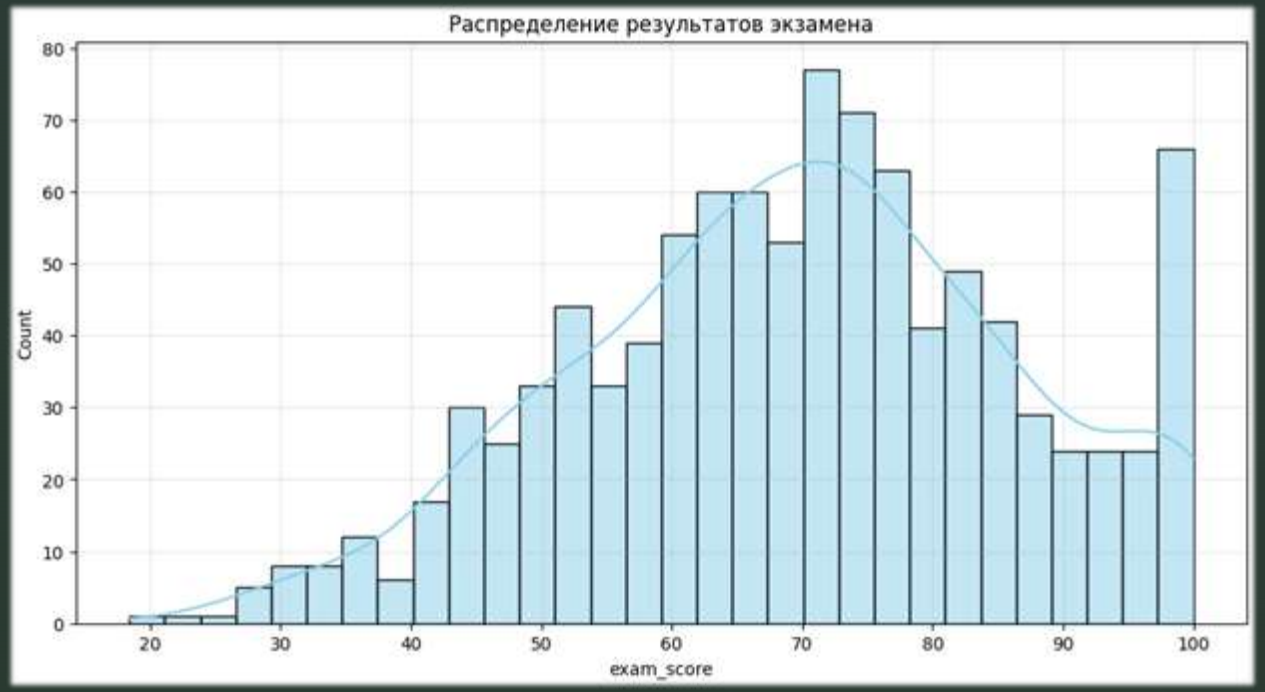
Идея метода:
Алгоритм рекурсивно делит выборку на подмножества так, чтобы в каждом из них объекты как можно больше принадлежали одному классу.
Критерием “хорошего разбиения” служат меры чистоты узла — например, Gini или энтропия.

Модель Decision Tree

- Основные шаги:
  1. Выбирается признак и порог, которые лучше всего разделяют данные по целевой переменной.
  2. Создаются новые ветви для каждого значения или диапазона признака.
  3. Процесс повторяется до тех пор, пока не достигнуты условия остановки (например, максимальная глубина).
  4. Листовые узлы содержат итоговые классы.

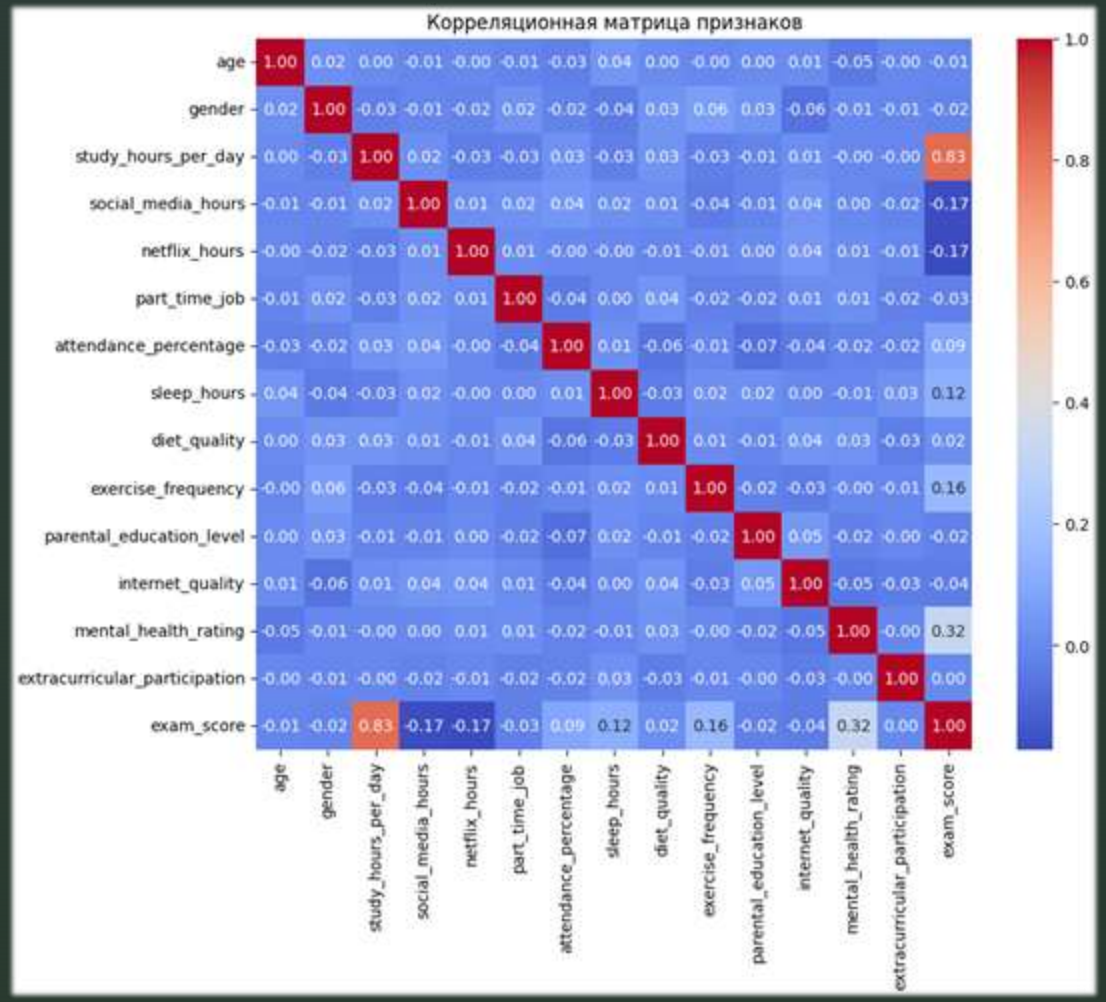
- Преимущества: простая интерпретация, не требует масштабирования данных.
- Недостатки: склонно к переобучению, особенно при большой глубине дерева; может быть нестабильным.

Целевая переменная. Распределение



Целевая переменная имеет среднее примерно 70 и значительную дисперсию. Это значит, что модель должна уметь отличать как слабых, так и сильных студентов. Гистограмма показывает небольшую правую асимметрию — есть отличники с близкими к 100 баллам.

Матрица корреляций. Влияние признаков



- Heatmap корреляций (ключевые корреляции):

  1. study_hours_per_day — сильная положительная корреляция (~0.8)

  2. social_media_hours, netflix_hours — умеренно отрицательные

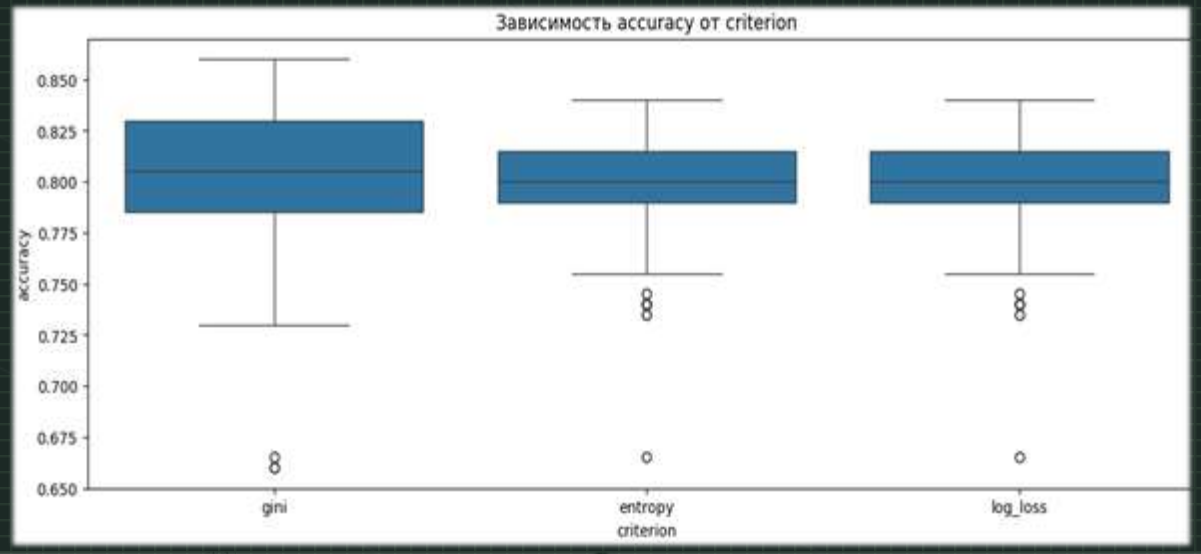
Модель Decision Tree

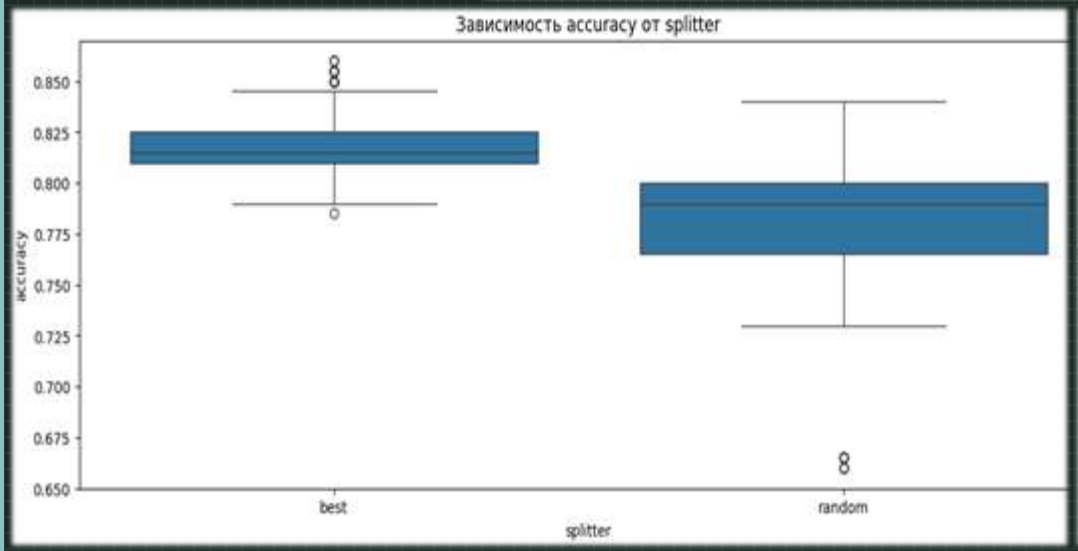
Параметры:

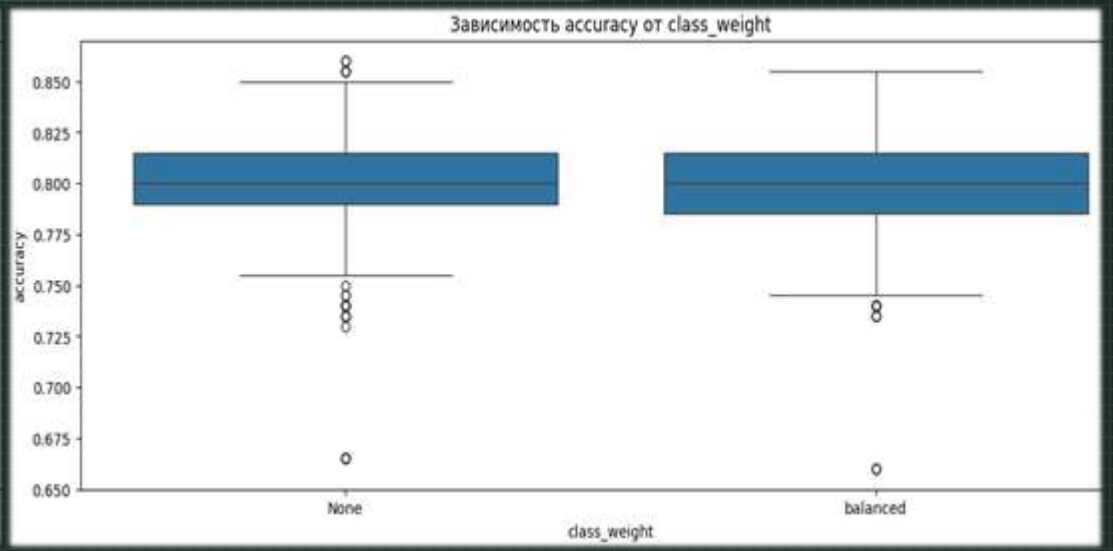
1. Критерий качества разбиения: индекс Джини (gini), энтропия (entropy), логарифмическая потеря (log_loss)
2. Стратегия выбора признака для разбиения: наилучшее разбиение (best), случайный признак (random)
3. Максимальная глубина дерева: без ограничения (None)
4. Минимальное число образцов для разбиения узла - 2
5. Минимальное количество образцов в листе - 1

Почему дерево: интерпретируемость, визуализация, не требуется масштабирование

Зависимость метрик от настроек используемой модели







Зависимость метрик от настроек используемой модели