

# **ЗАДАЧА КЛАССИФИКАЦИИ В МАШИННОМ ОБУЧЕНИИ**

Леу И.А. М80-309Б-23

Тарасов Е.Д. М80-309Б-23

Кривошапкин Е.Б М80-309Б-23

# ЦЕЛЬ И ЗАДАЧИ РАБОТЫ

- **Цель:** построить модель, которая предсказывает итоговый экзаменационный балл студента на основе его привычек и условий обучения.
- **Задачи:**
  - подготовить и проанализировать данные;
  - обучить модель «Дерево решений»;
  - оценить метрики и визуализировать результаты.



# ИСПОЛЬЗУЕМАЯ МОДЕЛЬ

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*.

Преимущества:

- Простая интерпретация (можно визуализировать дерево)
- Не требует масштабирования данных

Недостатки:

- Склонно к **переобучению**, особенно при большой глубине дерева
- Может быть нестабильным (небольшие изменения данных -> другое дерево)

# ДАТАСЕТ

**Название:** student\_habits\_performance.csv

**Размер:** 1000 наблюдений, 16 признаков

**Целевая переменная:** exam\_score

**Основные признаки:**

- количество часов учёбы, сна, соцсетей и Netflix
- процент посещаемости
- питание, физическая активность, наличие работы
- оценка психического здоровья

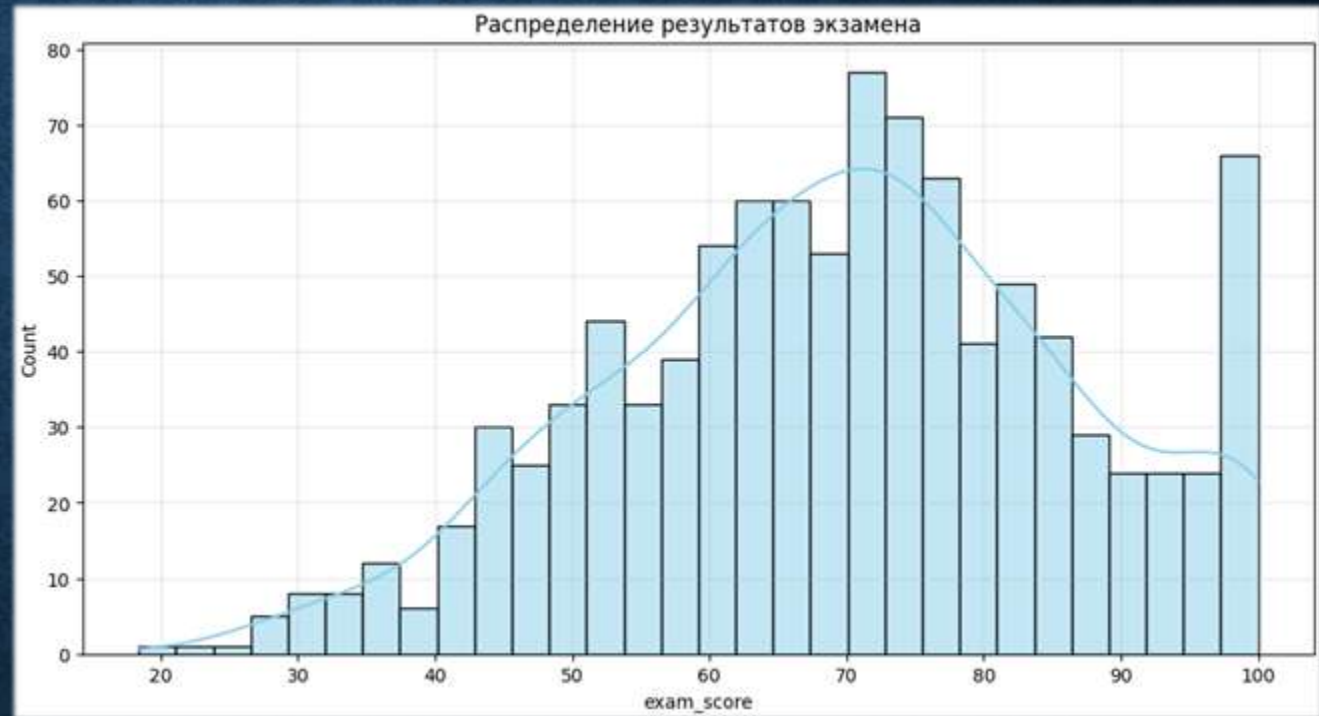


# ПОДГОТОВКА ДАННЫХ

- Удален идентификатор `student_id`
- Разделены признаки и целевая переменная
- Категориальные признаки преобразованы в числовые
- Разделение данных: 80% для тренировки, 20% для тестирования
- Без масштабирования — дерево не чувствительно к шкале

# РАСПРЕДЕЛЕНИЕ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

- Оценки распределены близко к нормальному закону
- Средний балл  $\approx 69$ , но имеется большое количество отличников
- Наблюдаются студенты как с низкой, так и с высокой успеваемостью





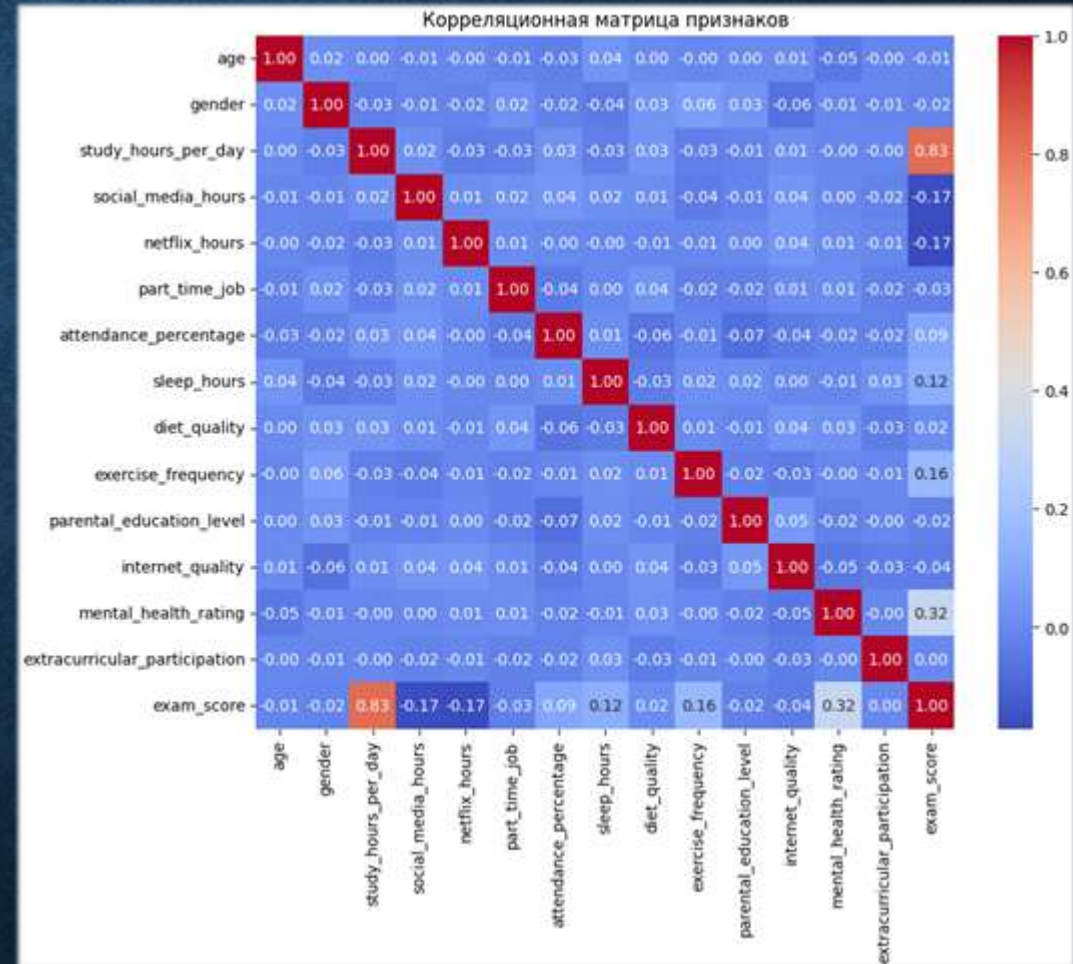
# КОРРЕЛЯЦИИ ПРИЗНАКОВ

Наибольшая положительная корреляция:

- `study_hours_per_day`: 0,83

Отрицательная корреляция:

- `social_media_hours`: -0,17
- `netflix_hours`: -0,17



# НАСТРОЙКА МОДЕЛИ

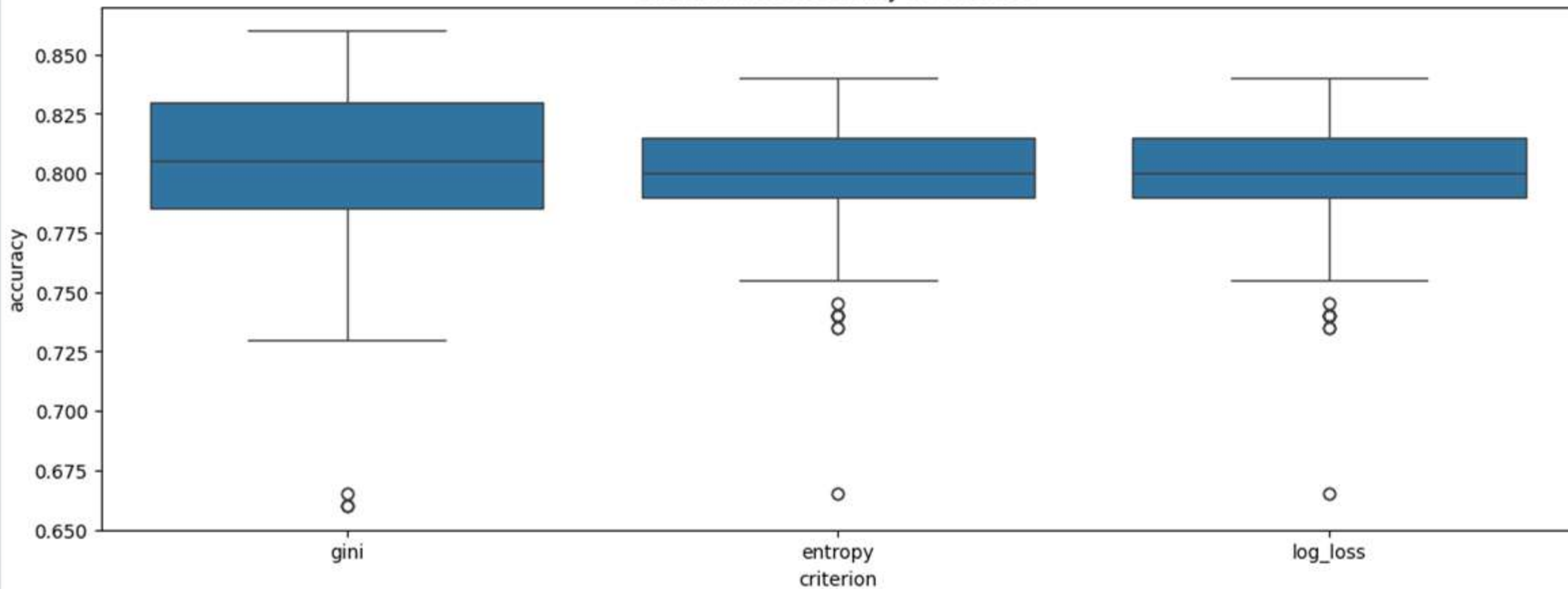
Основные параметры (подчеркнутые – по умолчанию) :

- Критерий качества разбиения: индекс Джини, энтропия, логарифмическая потеря
- Выбор признака для разбиения: наилучшее разбиение, случайный признак
- Максимальная глубина дерева: без ограничения
- Минимальное число образцов для разбиения узла: 2
- Минимальное количество образцов в листе: 1
- Веса классов для компенсации дисбаланса: None, balanced



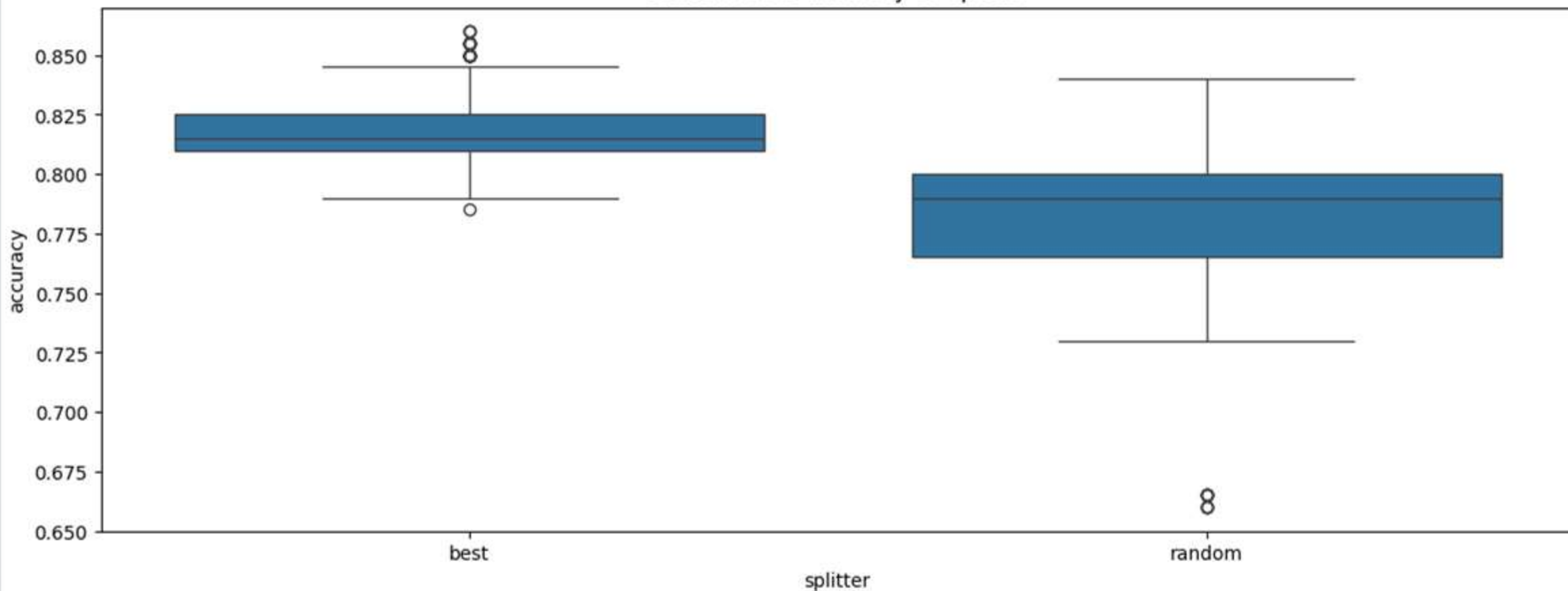
# ЗАВИСИМОСТЬ МЕТРИК ОТ НАСТРОЕК МОДЕЛИ

Зависимость accuracy от criterion



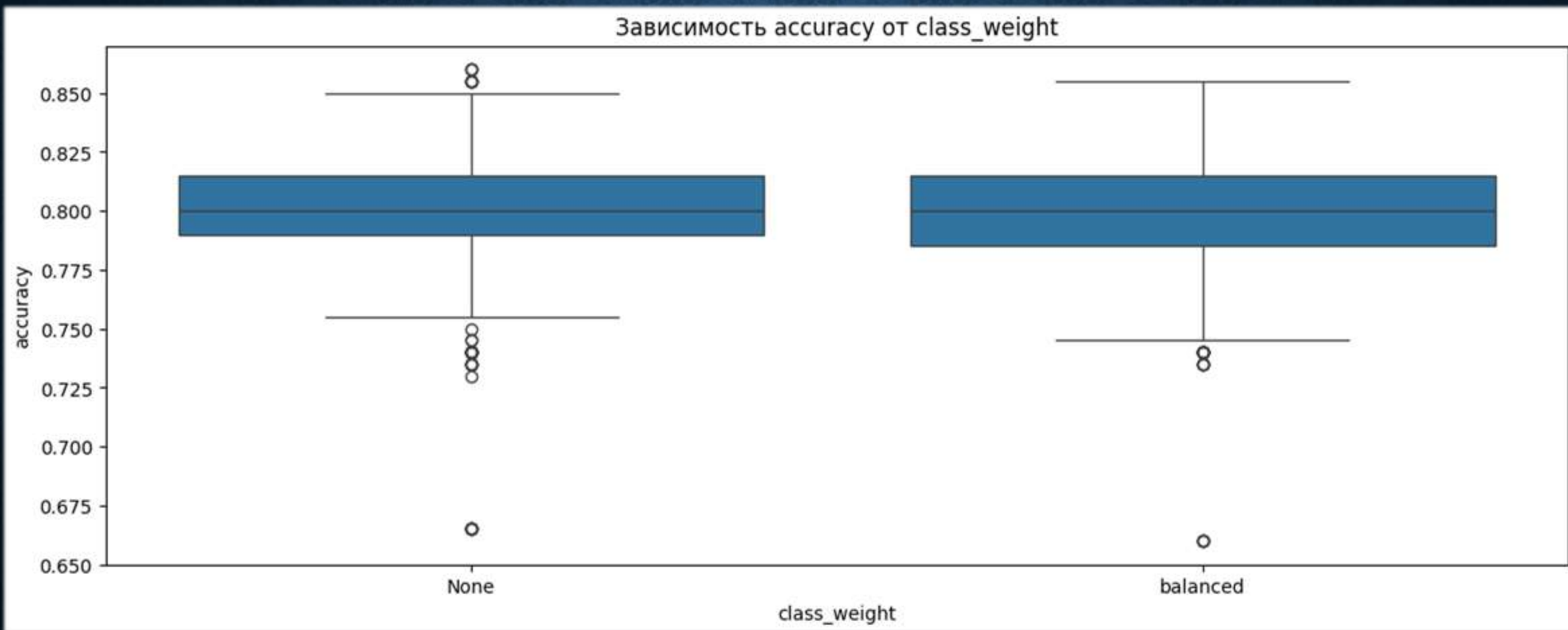
# ЗАВИСИМОСТЬ МЕТРИК ОТ НАСТРОЕК МОДЕЛИ

Зависимость accuracy от splitter

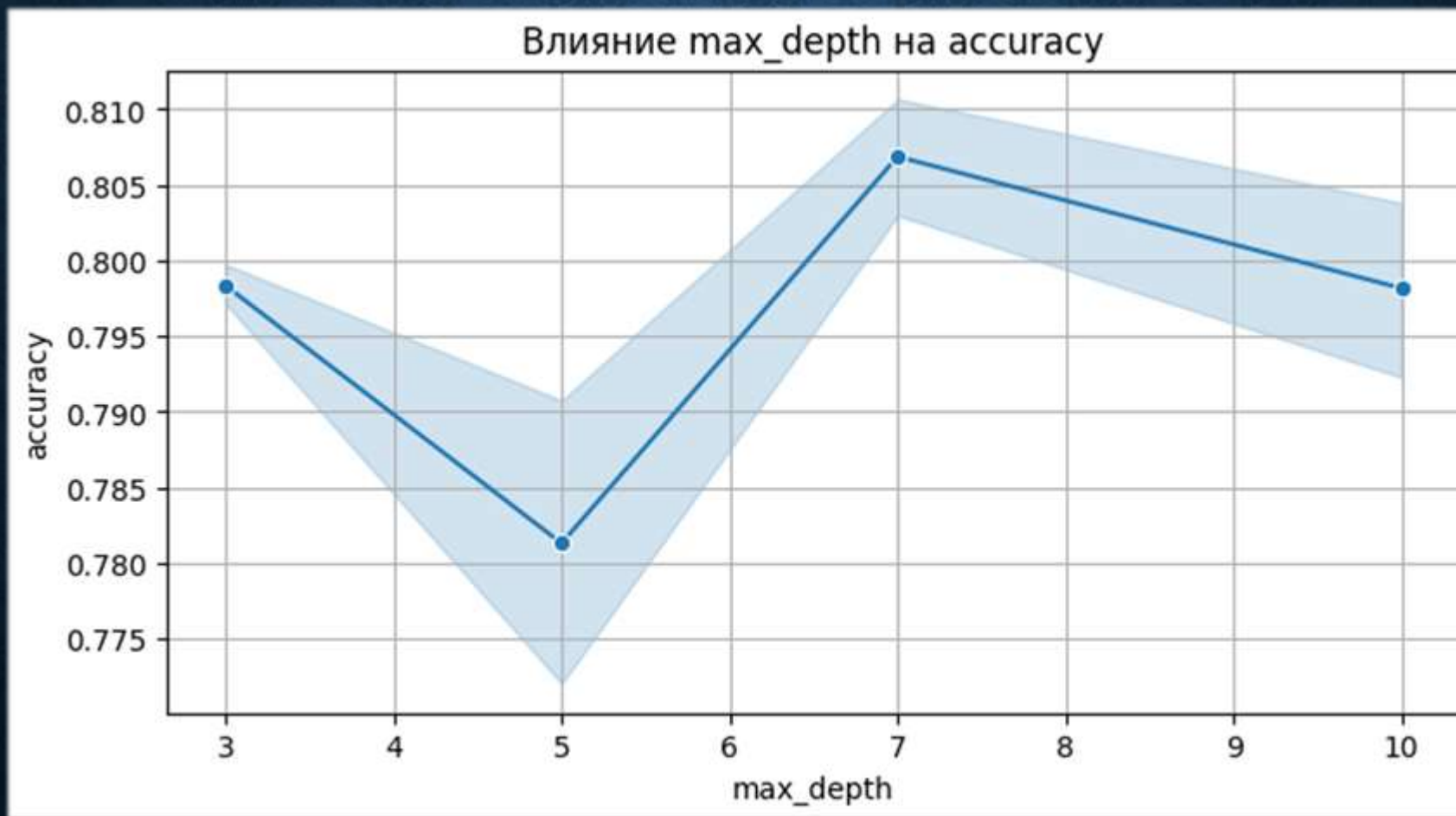




# ЗАВИСИМОСТЬ МЕТРИК ОТ НАСТРОЕК МОДЕЛИ

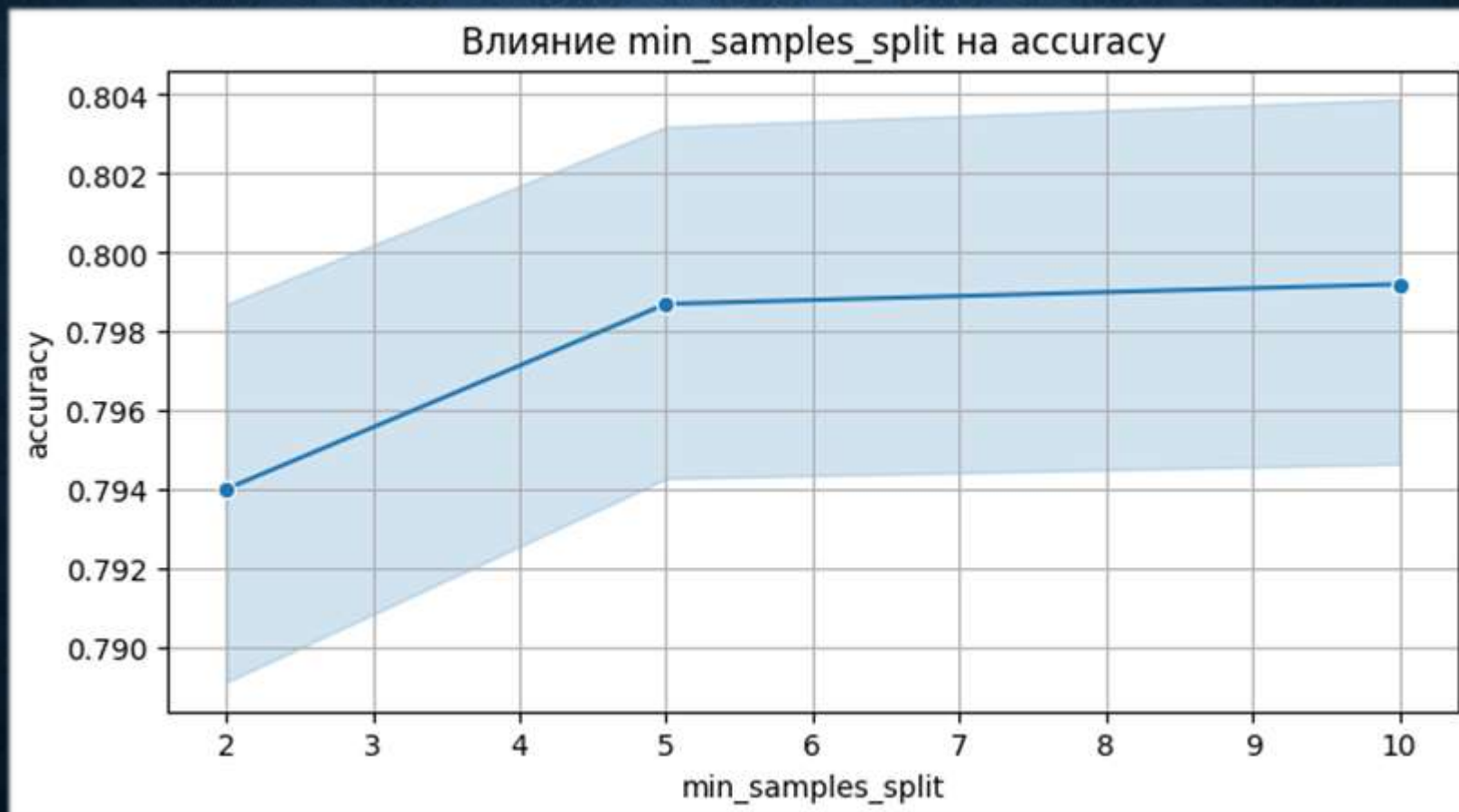


# ЗАВИСИМОСТЬ МЕТРИК ОТ НАСТРОЕК МОДЕЛИ

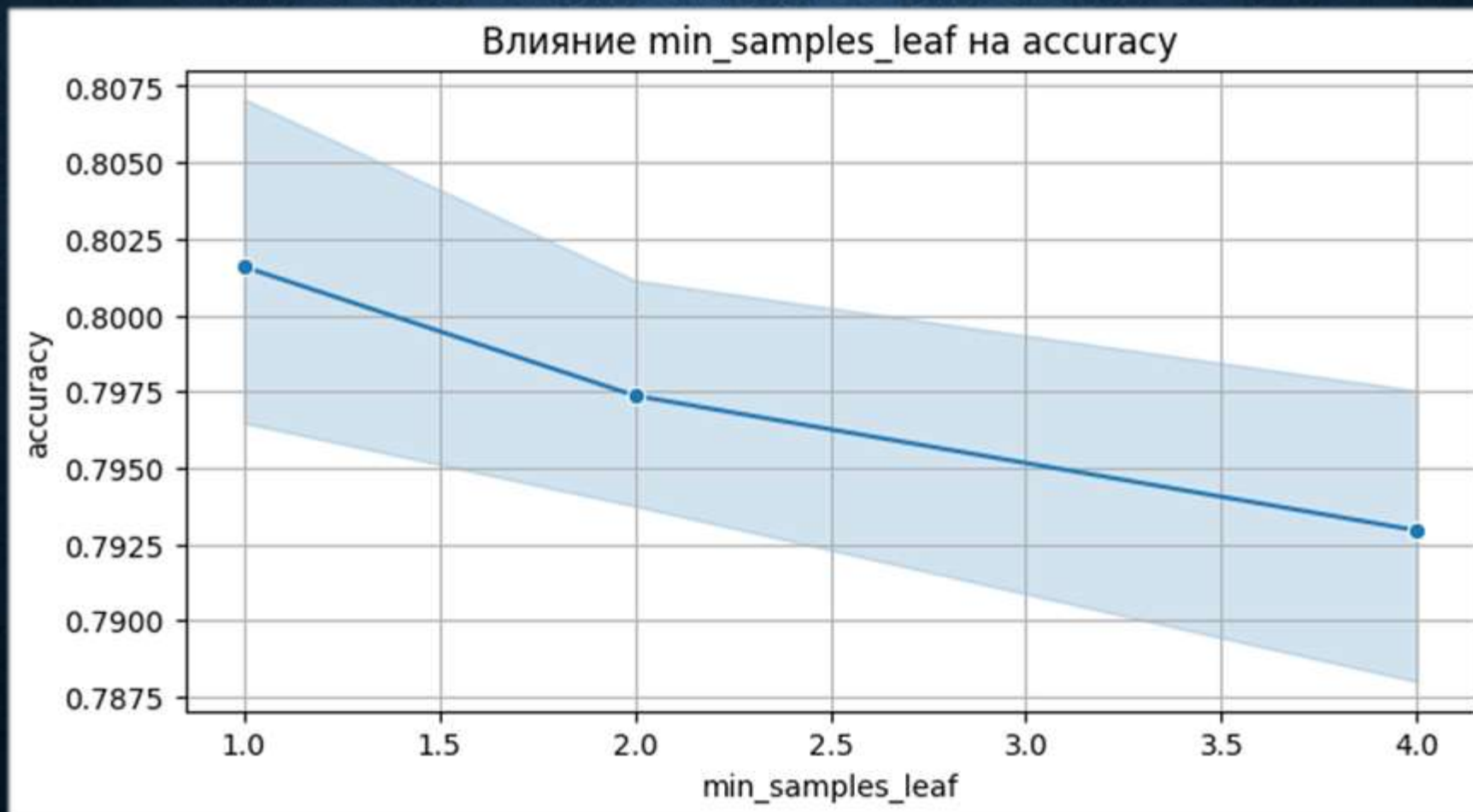




# ЗАВИСИМОСТЬ МЕТРИК ОТ НАСТРОЕК МОДЕЛИ



# ЗАВИСИМОСТЬ МЕТРИК ОТ НАСТРОЕК МОДЕЛИ





**Спасибо за внимание!**

