

Lab 2 NumPy+Pandas+Sklearn

Анализ цен на жилье в Бостоне

сдается до 18 октября!

Требования к заданию:

- Отчёт сдаётся в виде jupyter notebook + выслать тетрадку в формате html/pdf. Ноутбук должен быть аккуратным и читаемым (а не черновиком, где ячейки запущены в произвольном порядке, а результаты не воспроизводятся)

Исходные данные:

Используется классический датасет Boston Housing (506 строк, 13 признаков + целевая переменная MEDV — медианная стоимость жилья).

Доступ: <https://gist.github.com/nnbphuong/def91b5553736764e8e08f6255390f37>

Python

```
from sklearn.datasets import load_boston
import pandas as pd
import numpy as np
boston = load_boston()
df = pd.DataFrame(boston.data, columns=boston.feature_names)
df['MEDV'] = boston.target
```

Задание 1: Первичный анализ

- Выведите размерность и типы данных
- Проверьте наличие пропущенных значений
- Постройте распределение MEDV (гистограмма + boxplot)
- Найдите признаки с наибольшей корреляцией с MEDV

Задание 2: Предобработка

- Нормализуйте числовые признаки с помощью StandardScaler
- Создайте бинарный признак: `is_high_value = MEDV > 30`
- Удалите выбросы по Z-score (> 3 стандартных отклонения)

Задание 3: Визуализация

- Постройте тепловую карту корреляций (seaborn.heatmap)
- Нарисуйте scatter plot RM vs MEDV (кол-во комнат vs цена)
- Постройте график зависимости MEDV от LSTAT (доля бедного населения)

Задание 4: Линейная регрессия

- Постройте модель линейной регрессии $MEDV \sim RM + LSTAT + PTRATIO$
- Выведите коэффициенты и интерпретируйте их
- Оцените качество модели: R^2 , MAE, RMSE

Задание 5: NumPy-анализ вручную

- Реализуйте линейную регрессию вручную через матрицы: $\beta = (X^T X)^{-1} X^T y$.
- Сравните коэффициенты с результатами sklearn

Задание 6: Кластеризация (дополнительно)

- Примените KMeans для кластеризации районов по признакам CRIM, NOX, DIS
- Визуализируйте кластеры на scatter plot
- Проанализируйте среднюю цену жилья в каждом кластере

В отчете должно быть:

- Всё описанное в пунктах выше
- Выводы по корреляциям и модели
- Возможные улучшения (например, полиномиальные признаки, регуляризация)
- На доп баллы модель посложнее и краткое обоснование, чем выбранная вами модель лучше

