

Министерство образования Республики Беларусь
Учреждение образования
«Брестский Государственный технический университет»
Кафедра ИИТ

Лабораторная работа №1

По дисциплине: «Основы машинного обучения»

Тема: «Знакомство с анализом данных: предварительная обработка и визуализация»

Выполнила:

Студент 3 курса
Группы АС-66
Гончерёнок К.А.

Проверил:

Крощенко А. А.

Брест 2025

Цель работы: получить практические навыки работы с данными с использованием библиотек Pandas для манипуляции и Matplotlib для визуализации. Научиться выполнять основные шаги предварительной обработки данных, такие как очистка, нормализация и работа с различными типами признаков.

Вариант 2

Выборка Boston Housing. Содержит информацию о жилье в разных районах Бостона, включая уровень преступности, количество комнат и медианную стоимость.

Задачи:

1. Загрузите данные и выведите их основные статистические характеристики (.describe()).
2. Постройте матрицу корреляции и визуализируйте ее с помощью тепловой карты (heatmap).
3. Найдите признак, наиболее сильно коррелирующий с целевой переменной MEDV (медианная стоимость дома).
4. Постройте диаграмму рассеяния (scatter plot) для этого признака и MEDV.
5. Нормализуйте все числовые признаки, приведя их к диапазону от 0 до 1.
6. Визуализируйте распределение уровня преступности (CRIM) с помощью гистограммы.

Код программы:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

data = pd.read_csv("BostonHousing.csv")

corr_matrix = data.corr()

plt.figure(figsize=(12, 8))
plt.imshow(corr_matrix, cmap="coolwarm", interpolation="nearest")
plt.colorbar(label="Корреляция")
plt.xticks(range(len(corr_matrix.columns)), corr_matrix.columns, rotation=90)
plt.yticks(range(len(corr_matrix.columns)), corr_matrix.columns)
plt.title("Матрица корреляции признаков", fontsize=16)

for i in range(len(corr_matrix.columns)):
    for j in range(len(corr_matrix.columns)):
        plt.text(j, i, f"{corr_matrix.iloc[i, j]:.2f}",
                 ha="center", va="center", color="black", fontsize=8)

plt.tight_layout()
plt.savefig("corr_matrix.png", dpi=300)
```

```

plt.close()

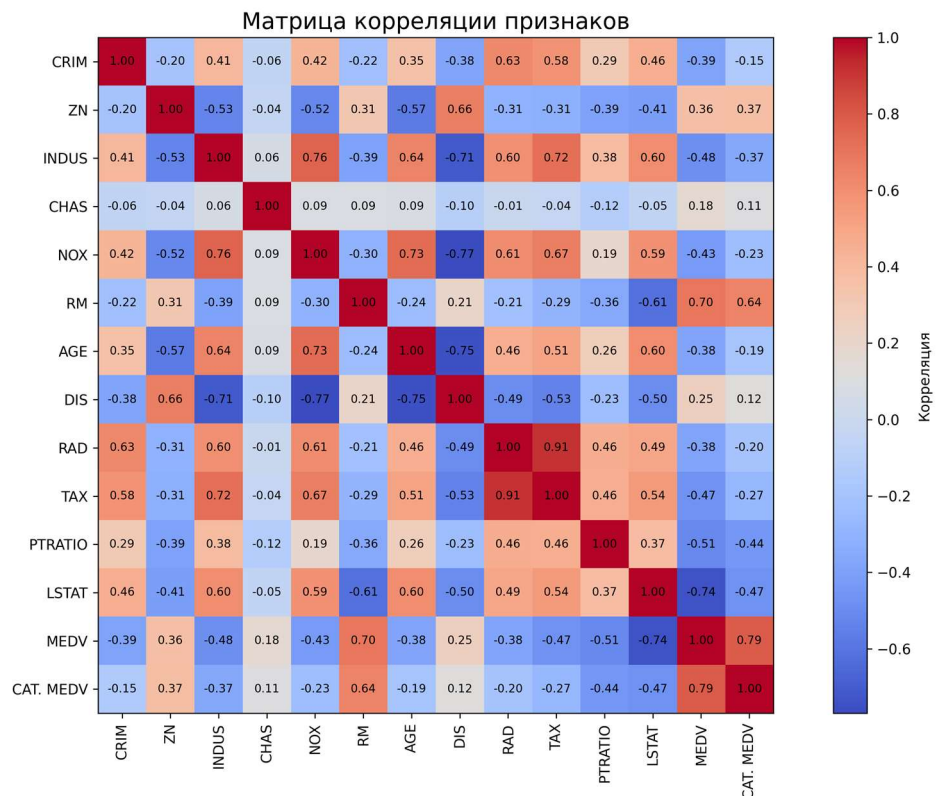
x = data["MEDV"]
y = data["LSTAT"]

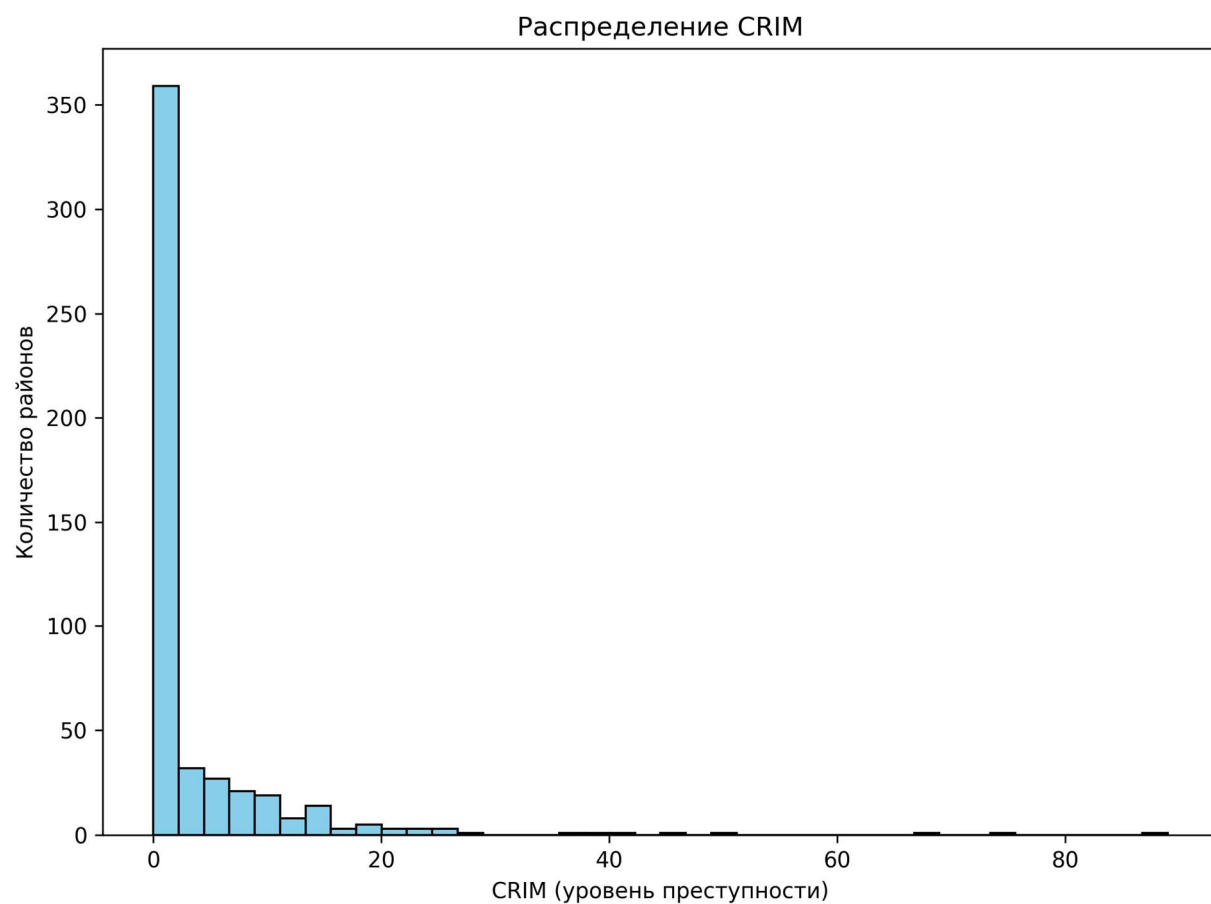
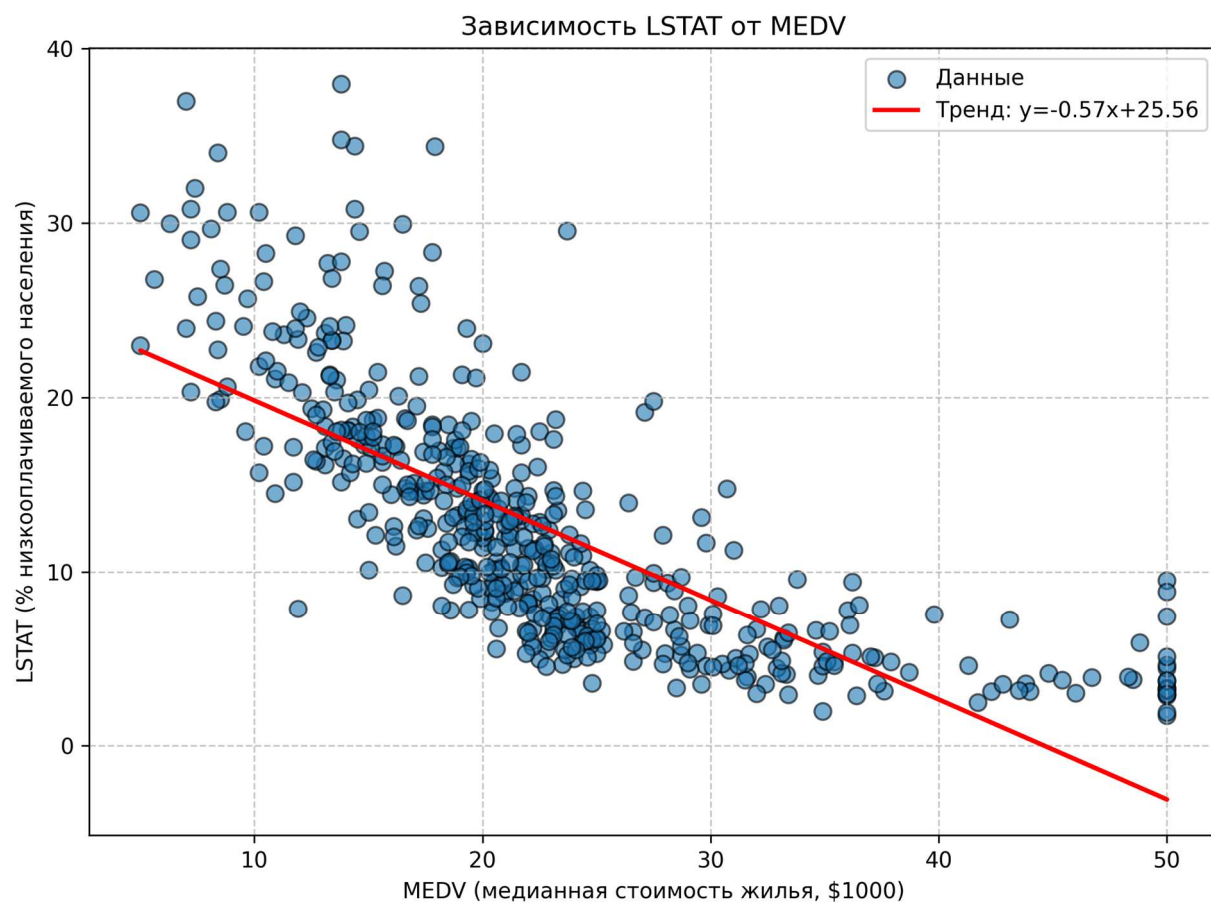
plt.figure(figsize=(8, 6))
plt.scatter(x, y, alpha=0.6, edgecolors="k", s=60, label="Данные")
a, b = np.polyfit(x, y, 1)
plt.plot(x, a*x + b, color="red", linewidth=2, label=f"Тренд: y={a:.2f}x+{b:.2f}")
plt.xlabel("MEDV (медианная стоимость жилья, $1000)")
plt.ylabel("LSTAT (% низкооплачиваемого населения)")
plt.title("Зависимость LSTAT от MEDV")
plt.legend()
plt.grid(True, linestyle="--", alpha=0.7)
plt.tight_layout()
plt.savefig("scatter_MEDV_LSTAT.png", dpi=300)
plt.close()

plt.figure(figsize=(8, 6))
plt.hist(data["CRIM"], bins=40, color="skyblue", edgecolor="black")
plt.xlabel("CRIM (уровень преступности)")
plt.ylabel("Количество районов")
plt.title("Распределение CRIM")
plt.tight_layout()
plt.savefig("hist_CRIM.png", dpi=300)
plt.close()

```

Результат работы программы:





Вывод: получил практические навыки работы с данными с использованием библиотек Pandas для манипуляции и Matplotlib для визуализации. Научился выполнять основные шаги предварительной обработки данных, такие как очистка, нормализация и работа с различными типами признаков.