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»**

Выполнил:
Студент 3 курса
Группы АС-65
Ярмак К. А.
Проверил:
Крощенко А. А.

Брест 2025

Цель: 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 12

Задание 1. Загрузите данные и выведите их основные статистические характеристики (.describe()).

```
From df import df

print(df.describe())
```

df.py:

```
import pandas as pd

fileName = "BostonHousing.csv"

df = pd.read_csv(fileName)

corr_matrix = df.corr()
targetCorr = corr_matrix['MEDV'].drop(["MEDV", "CAT. MEDV"])
mostCorrelated = targetCorr.idxmax()
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	...	RAD	TAX	PTRATIO	LSTAT	MEDV	CAT. MEDV
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	...	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	...	9.549407	408.237154	18.455534	12.653063	22.532806	0.166008
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	...	8.707259	168.537116	2.164946	7.141062	9.197104	0.372456
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	...	1.000000	187.000000	12.600000	1.730000	5.000000	0.000000
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	...	4.000000	279.000000	17.400000	6.950000	17.025000	0.000000
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	...	5.000000	330.000000	19.050000	11.360000	21.200000	0.000000
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	...	24.000000	666.000000	20.200000	16.955000	25.000000	0.000000
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	...	24.000000	711.000000	22.000000	37.970000	50.000000	1.000000

[8 rows x 14 columns]

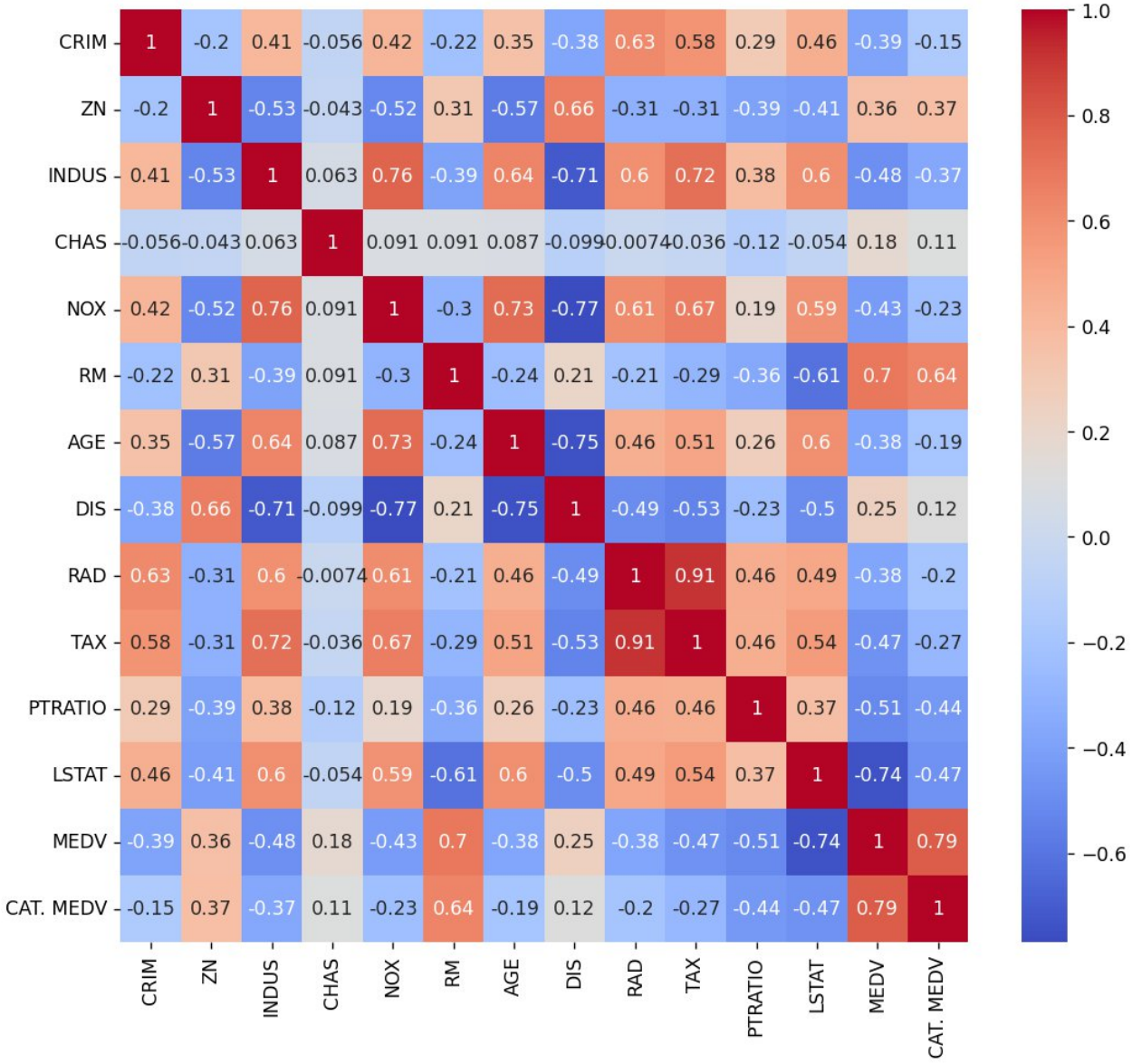
Задание 2. Постройте матрицу корреляции и визуализируйте ее с помощью тепловой карты (heatmap).

```
import seaborn as sns
import matplotlib.pyplot as plt

from df import df


corr_matrix = df.corr()

plt.figure(figsize=(10, 10))
sns.heatmap(corr_matrix, annot=True, cmap="coolwarm")
plt.show()
```



Задание 3. Найдите признак, наиболее сильно коррелирующий с целевой переменной MEDV (медианная стоимость дома).

```
from df import df

corr_matrix = df.corr()


targetCorr = corr_matrix["MEDV"].drop(["MEDV", "CAT. MEDV"])

mostCorrelated = targetCorr.idxmax()
print(f"Most corr with MEDV: {mostCorrelated}")
```

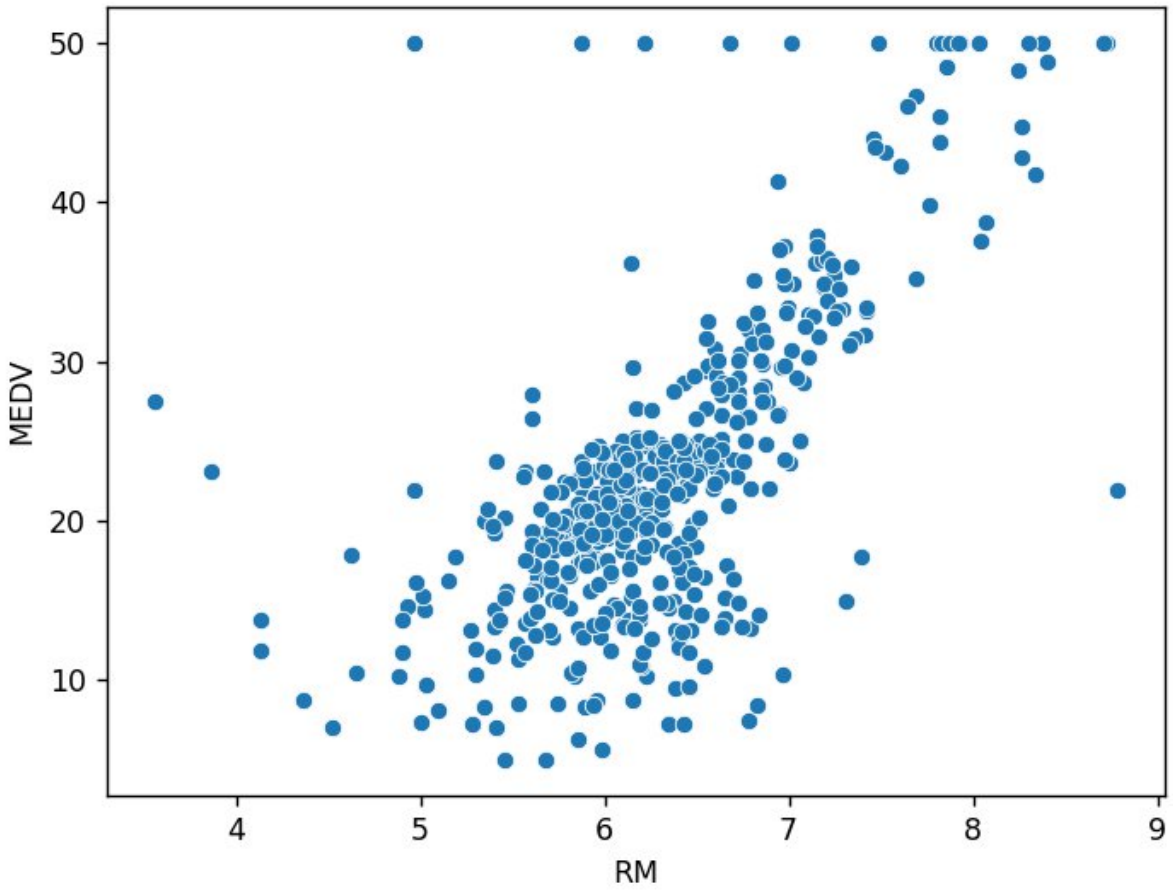
```
Most corr with MEDV: RM
```

Задание 4. Постройте диаграмму рассеяния (scatter plot) для этого признака и MEDV.

```
import seaborn as sns
import matplotlib.pyplot as plt

from df import df, mostCorrelated


sns.scatterplot(x=df[mostCorrelated], y=df["MEDV"])
plt.xlabel(mostCorrelated)
plt.ylabel("MEDV")
plt.show()
```



Задание 5. Нормализуйте все числовые признаки, приведя их к диапазону от 0 до 1.

```
from sklearn.preprocessing import MinMaxScaler
from df import pd, df

scaler = MinMaxScaler()
df_normalized = pd.DataFrame(scaler.fit_transform(df), columns=df.columns)

print(df_normalized.describe())
```

	CRIM	ZN	INDUS	CHAS	NOX	...	TAX	PTRATIO	LSTAT	MEDV	CAT. MEDV
count	506.000000	506.000000	506.000000	506.000000	506.000000	...	506.000000	506.000000	506.000000	506.000000	506.000000
mean	0.040544	0.113636	0.391378	0.069170	0.349167	...	0.422208	0.622929	0.301409	0.389618	0.166008
std	0.096679	0.233225	0.251479	0.253994	0.238431	...	0.321636	0.230313	0.197049	0.204380	0.372456
min	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000851	0.000000	0.173387	0.000000	0.131687	...	0.175573	0.510638	0.144040	0.267222	0.000000
50%	0.002812	0.000000	0.338343	0.000000	0.314815	...	0.272901	0.686170	0.265728	0.360000	0.000000
75%	0.041258	0.125000	0.646628	0.000000	0.491770	...	0.914122	0.808511	0.420116	0.444444	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	1.000000	1.000000	1.000000	1.000000

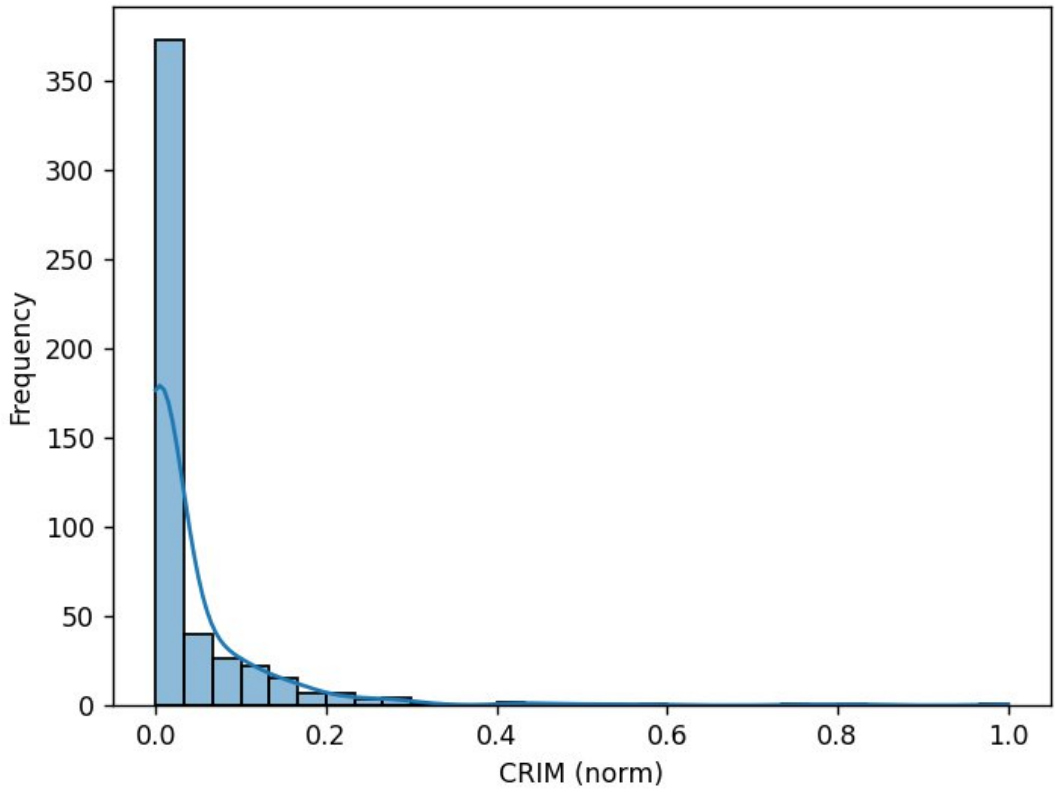
[8 rows x 14 columns]

Задание 6. Визуализируйте распределение уровня преступности (CRIM) с помощью гистограммы.

```
import matplotlib.pyplot as plt
import seaborn as sns

from t5 import df_normalized

sns.histplot(df_normalized["CRIM"], bins=30, kde=True)
plt.xlabel("CRIM (norm)")
plt.ylabel("Frequency")
plt.show()
```



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