

**Московский государственный технический  
университет им. Н. Э. Баумана**

Курс «Технологии машинного обучения»

Отчёт по лабораторной работе №1

Выполнил:  
Фролов М. К.  
группа ИУ5-64Б

Проверил:  
Гапанюк Ю.Е.

Дата: 07.04.25

Дата:

Подпись:

Подпись:

Москва, 2025 г.

**Цель лабораторной работы:** изучение различных методов визуализация данных.

**Краткое описание.** Построение основных графиков, входящих в этап разведочного анализа данных.

Рекомендуемые инструментальные средства можно посмотреть [здесь](#).

**Задание:**

- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов [здесь](#).
- Для первой лабораторной работы рекомендуется использовать датасет без пропусков в данных, например из [Scikit-learn](#).
- Пример преобразования датасетов Scikit-learn в Pandas Dataframe можно посмотреть [здесь](#).

Для лабораторных работ не рекомендуется выбирать датасеты большого размера.

- Создать ноутбук, который содержит следующие разделы:
  1. Текстовое описание выбранного Вами набора данных.
  2. Основные характеристики датасета.
  3. Визуальное исследование датасета.
  4. Информация о корреляции признаков.
- Сформировать отчет и разместить его в своем репозитории на github.

**Ход выполнения:**

```
import pandas as pd
from sklearn.datasets import load_diabetes
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

diabetes = load_diabetes()
```

[1] 13.4s

```
df = pd.DataFrame(data=diabetes.data, columns=diabetes.feature_names)
df['target'] = diabetes.target
df.head()
```

[ ]

```
...
   age    sex    bmi    bp    s1    s2    s3    s4    s5    s6  target
0  0.038076  0.050680  0.061696  0.021872 -0.044223 -0.034821 -0.043401 -0.002592  0.019907 -0.017646  151.0
1 -0.001882 -0.044642 -0.051474 -0.026328 -0.008449 -0.019163  0.074412 -0.039493 -0.068332 -0.092204   75.0
2  0.085299  0.050680  0.044451 -0.005670 -0.045599 -0.034194 -0.032356 -0.002592  0.002861 -0.025930  141.0
3 -0.089063 -0.044642 -0.011595 -0.036656  0.012191  0.024991 -0.036038  0.034309  0.022688 -0.009362  206.0
4  0.005383 -0.044642 -0.036385  0.021872  0.003935  0.015596  0.008142 -0.002592 -0.031988 -0.046641  135.0
```



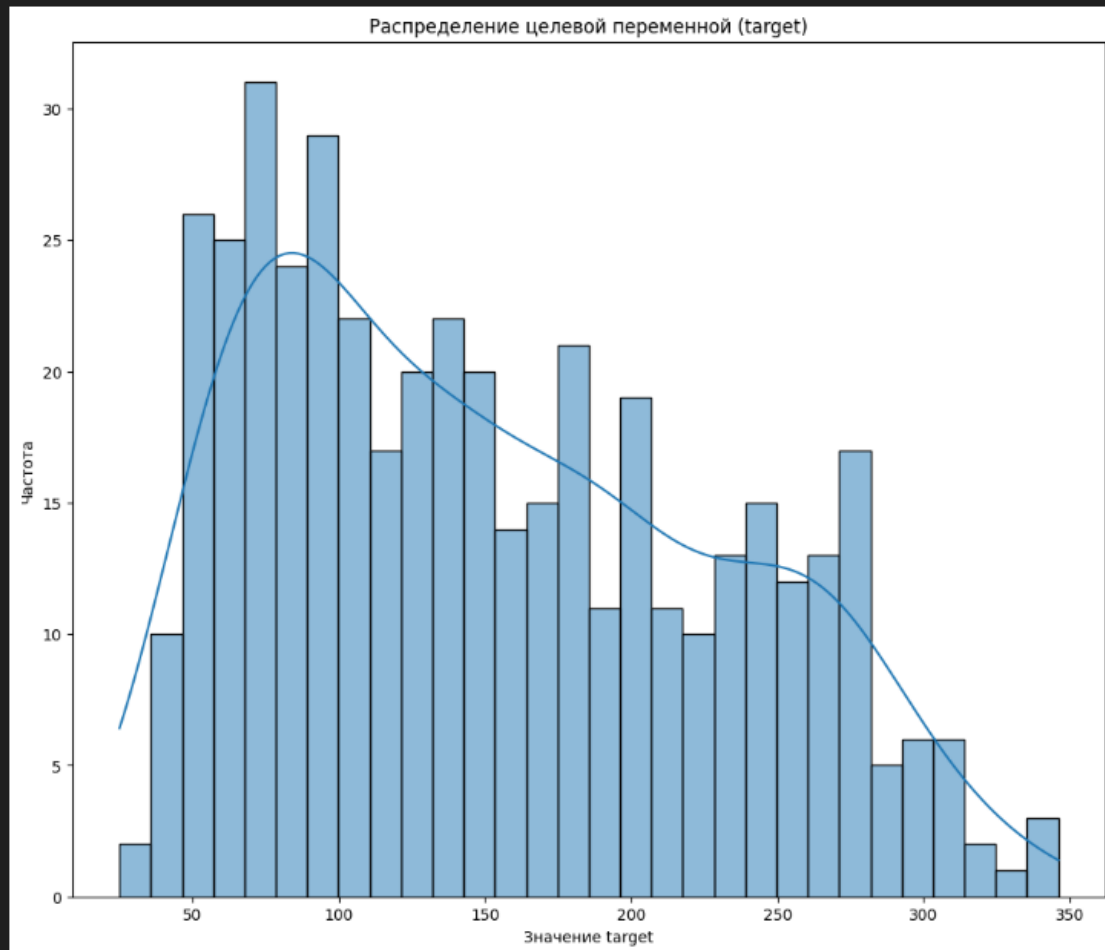
```
df.info()
```

[3] 0.0s

```
...
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 442 entries, 0 to 441
Data columns (total 11 columns):
 #   Column  Non-Null Count  Dtype  
---  -
 0   age     442 non-null    float64
 1   sex     442 non-null    float64
 2   bmi     442 non-null    float64
 3   bp      442 non-null    float64
 4   s1      442 non-null    float64
 5   s2      442 non-null    float64
 6   s3      442 non-null    float64
 7   s4      442 non-null    float64
 8   s5      442 non-null    float64
 9   s6      442 non-null    float64
10  target  442 non-null    float64
dtypes: float64(11)
memory usage: 38.1 KB
```

```
plt.figure(figsize=(12, 10))
sns.histplot(df['target'], kde=True, bins=30)
plt.title('Распределение целевой переменной (target)')
plt.xlabel('Значение target')
plt.ylabel('Частота')
plt.show()
```

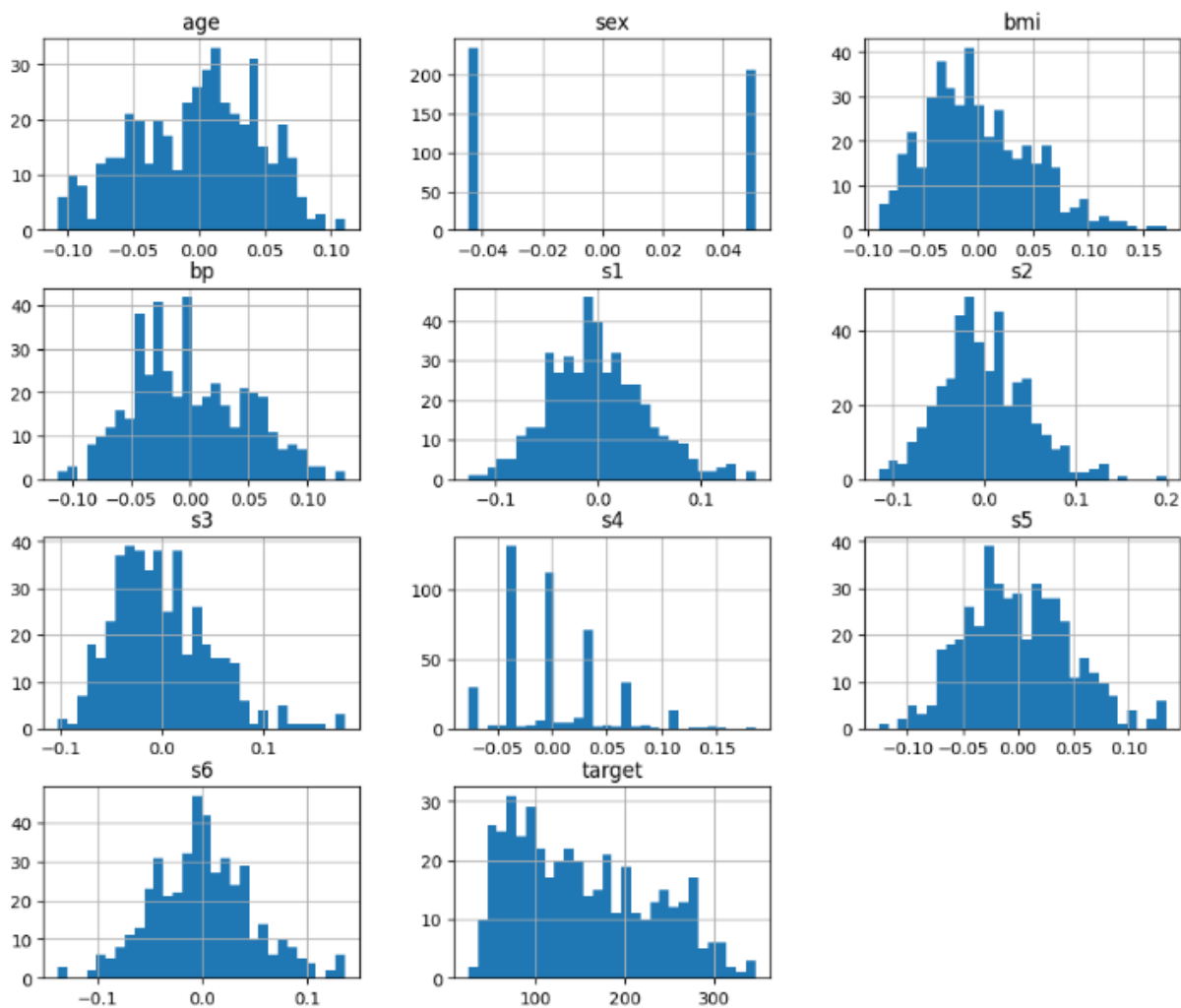
✓ 0.3s



```
df.hist(figsize=(12, 10), bins=30)
plt.suptitle('Распределение признаков')
plt.show()
```

✓ 1.5s

Распределение признаков



## ИНФОРМАЦИЯ О КОРРЕЛЯЦИИ ПРИЗНАКОВ

```
target_corr = df.corr()['target'].sort_values(ascending=False)
print(target_corr)
```

✓ 0.0s

```
target    1.000000
bmi       0.586450
s5        0.565883
bp        0.441482
s4        0.430453
s6        0.382483
s1        0.212022
age       0.187889
s2        0.174054
sex       0.043062
s3       -0.394789
Name: target, dtype: float64
```

```
sns.pairplot(df)
plt.show()
```

✓ 23.7s

