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:
Студентка 3 курса
Группы АС-65
Рапин Е. Ю.
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 11

Выборка Titanic. Содержит информацию о пассажирах лайнера, включая их возраст, пол, класс каюты и факт выживания.

Задание 1. Загрузите данные и выведите первые 5 строк, а также общую информацию о столбцах (.info()).

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
import numpy as np

data = pd.read_csv("Titanic-Dataset.csv")

print("Первые 5 строк:")
print(data.head())

print("\nИнформация о данных:")
print(data.info())
```

```
Первые 5 строк:
   PassengerId  Survived  Pclass  ...     Fare Cabin  Embarked
0            1         0       3  ...   7.2500   NaN         S
1            2         1       1  ...  71.2833   C85         C
2            3         1       3  ...   7.9250   NaN         S
3            4         1       1  ...  53.1000  C123         S
4            5         0       3  ...   8.0500   NaN         S

[5 rows x 12 columns]
```

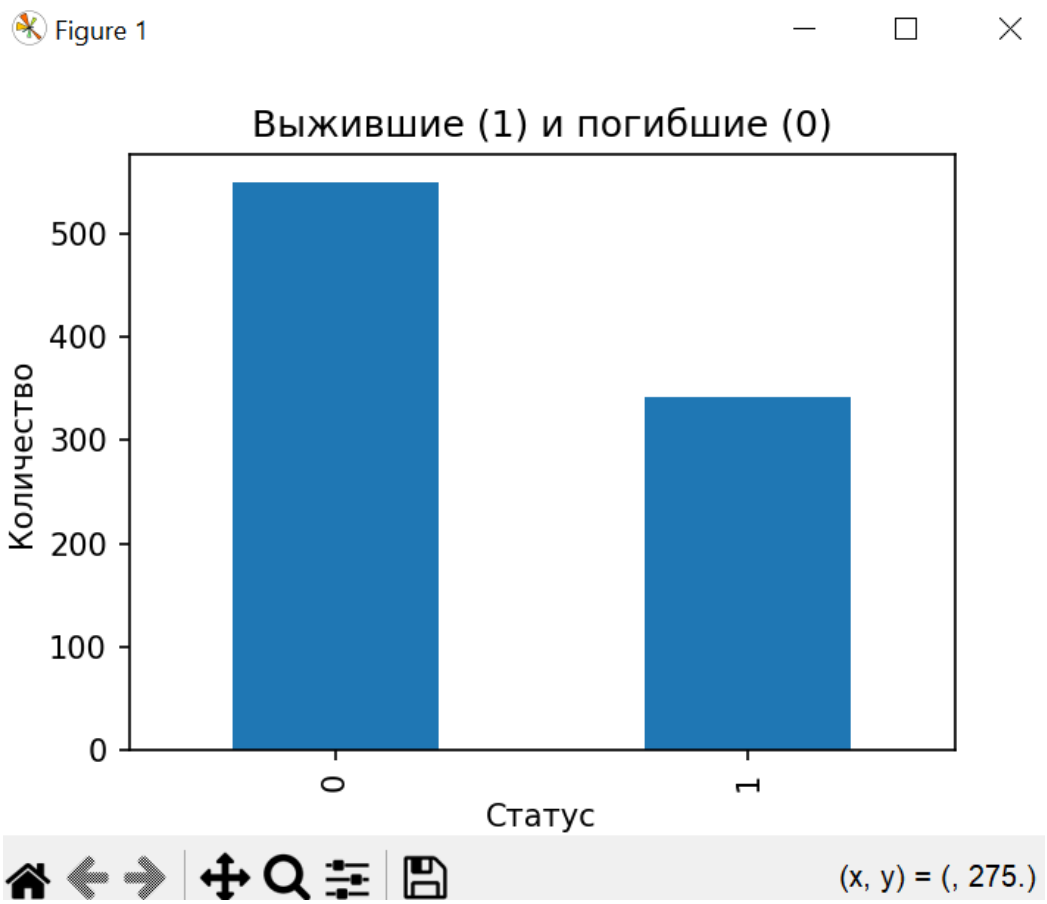
```
Информация о данных:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
None
```

Задание 2. Найдите и визуализируйте количество выживших и погибших пассажиров с помощью столбчатой диаграммы.

```
print("\nКоличество выживших и погибших:")
print(data['Survived'].value_counts())
```

```
Количество выживших и погибших:
Survived
0    549
1    342
Name: count, dtype: int64
```

```
# Столбчатая диаграмма выживаемости
data['Survived'].value_counts().plot(kind='bar')
plt.title("Выжившие (1) и погибшие (0)")
plt.xlabel("Статус")
plt.ylabel("Количество")
plt.show()
```



Задание 3. Обработайте пропуски в столбце Age, заполнив их медианным значением.

```
print("\nПропуски в Age до обработки:", data['Age'].isnull().sum())

median_age = data['Age'].median()
data['Age'] = data['Age'].fillna(median_age)

print("Пропуски в Age после обработки:", data['Age'].isnull().sum())
```

```
Пропуски в Age до обработки: 177
Пропуски в Age после обработки: 0
```

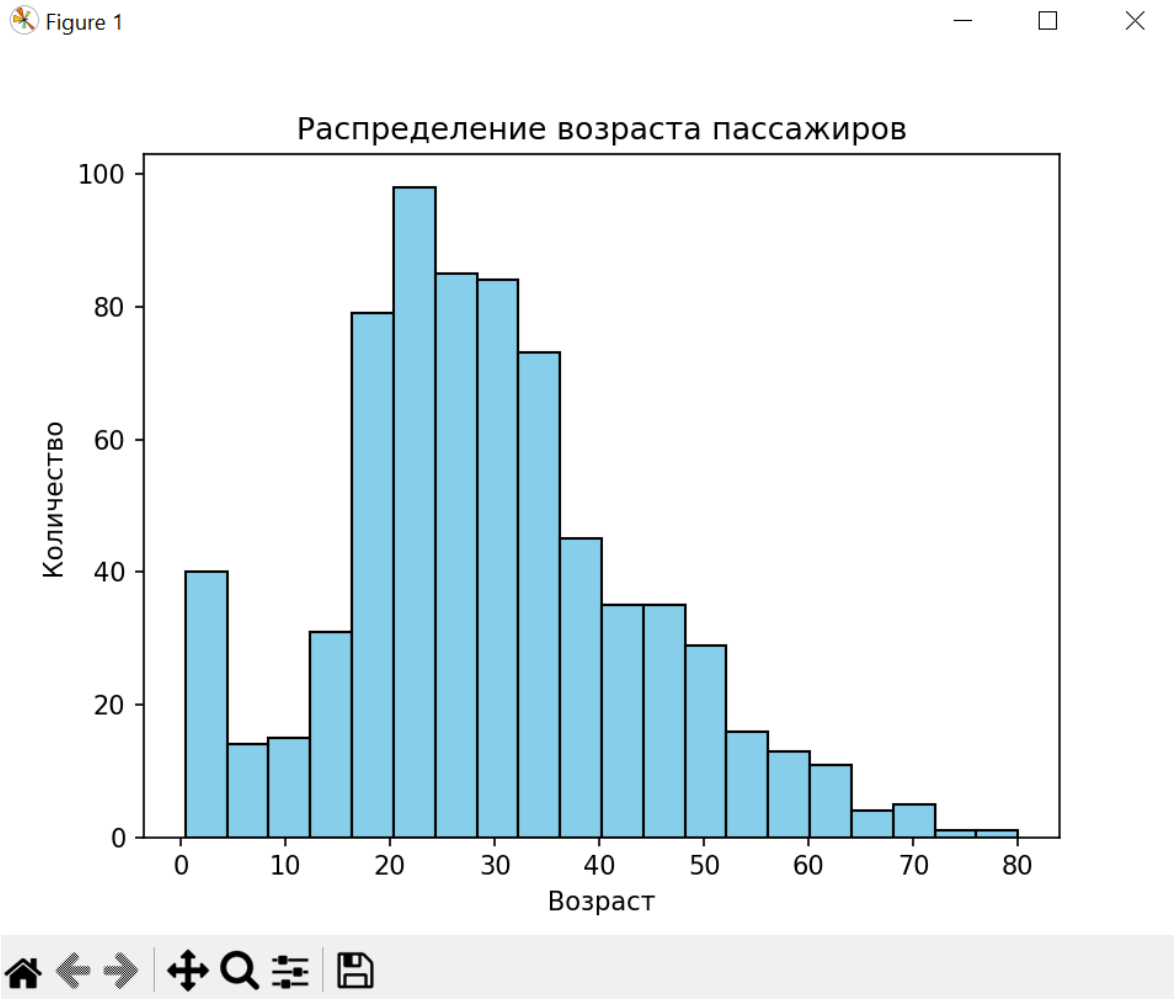
Задание 4. Преобразуйте категориальные признаки Sex и Embarked в числовые с помощью One-Hot Encoding.

```
data_encoded = pd.get_dummies(data, columns=['Sex', 'Embarked'],
drop_first=True)

print("\nДанные после One-Hot Encoding:")
print(data_encoded.head())
```

```
Данные после One-Hot Encoding:
   PassengerId  Survived  Pclass  ... Sex_male  Embarked_Q  Embarked_S
0            1         0       3  ...     True       False        True
1            2         1       1  ...    False       False       False
2            3         1       3  ...    False       False        True
3            4         1       1  ...    False       False        True
4            5         0       3  ...     True       False        True

[5 rows x 13 columns]
```

Задание 5. Постройте гистограмму распределения возрастов пассажиров.

```
plt.hist(data['Age'], bins=20, color='skyblue', edgecolor='black')
plt.title("Распределение возраста пассажиров")
plt.xlabel("Возраст")
plt.ylabel("Количество")
plt.show()
```



Задание 6. Создайте новый признак FamilySize путем сложения значений из столбцов SibSp и Parch.

```
data_encoded['FamilySize'] = data_encoded['SibSp'] + data_encoded['Parch']

print("\nПервые строки с новым признаком FamilySize:")
print(data_encoded[['SibSp', 'Parch', 'FamilySize']].head())
```

```
Первые строки с новым признаком FamilySize:
   SibSp  Parch  FamilySize
0      1      0           1
1      1      0           1
2      0      0           0
3      1      0           1
4      0      0           0
```

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