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
Ракецкий П. П.
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 4

Задание 1. Загрузите данные и выведите информацию о типах столбцов.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler

# Загрузка встроенного датасета iris из seaborn
df = sns.load_dataset('iris')

# Переименуем колонки для соответствия вашему коду
df = df.rename(columns={
    'sepal_length': 'sepal.length',
    'sepal_width': 'sepal.width',
    'petal_length': 'petal.length',
    'petal_width': 'petal.width',
    'species': 'variety'
})

print("Первые 5 строк данных:")
print(df.head())
print("\n" + "="*50)

# 1. Проверка пропущенных значений
print("Пропущенных значений:")
print(df.isnull().sum())
print("\n" + "="*50)
```

```
Первые 5 строк данных:
   sepal.length  sepal.width  petal.length  petal.width variety
0           5.1          3.5           1.4          0.2  setosa
1           4.9          3.0           1.4          0.2  setosa
2           4.7          3.2           1.3          0.2  setosa
3           4.6          3.1           1.5          0.2  setosa
4           5.0          3.6           1.4          0.2  setosa

==================================================
Пропущенных значений:
sepal.length    0
sepal.width     0
petal.length    0
petal.width     0
variety         0
dtype: int64
```

Задание 2. Преобразуйте целевую переменную quality в категориальную: "плохое" (<=4), "среднее" (5-6), "хорошее" (>=7).

```
# 2. Количество образцов каждого вида
print("Количество образцов каждого вида:")
counts = {}  # пустой словарь для подсчёта

for item in df['variety']:
```

```
    if item in counts:
        counts[item] += 1
    else:
        counts[item] = 1
for key, value in counts.items():
    print(f"{key}: {value}")
print("\n" + "="*50)
```

```
==================================================
Количество образцов каждого вида:
setosa: 50
versicolor: 50
virginica: 50

==================================================
```

Задание 3. Обработать пропущенные значения (например, заполнить средним значением или удалить строки/столбцы).

```
# 3. Парные диаграммы рассеяния
sns.pairplot(df, hue='variety')
plt.suptitle("Парные диаграммы рассеяния для Iris dataset", y=1.02)
plt.show()
```

Задание 4. Проверьте корреляцию между fixed acidity и pH. Визуализируйте эту зависимость на диаграмме рассеяния.

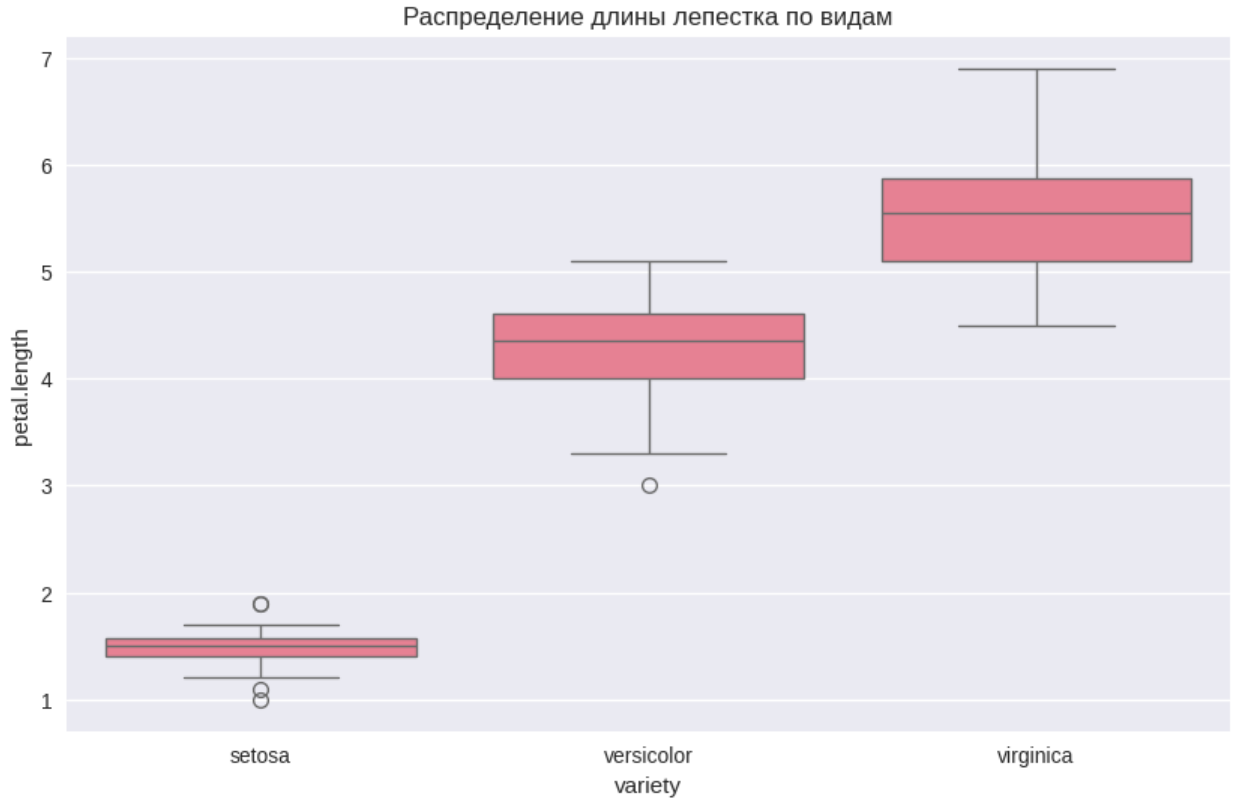
```
# 4. Средние значения
mean = df.groupby('variety').mean()
print("Средние значения по видам:")
print(mean)
print("\n" + "="*50)
```

```
Средние значения по видам:
            sepal.length  sepal.width  petal.length  petal.width
variety
setosa             5.006        3.428         1.462        0.246
versicolor         5.936        2.770         4.260        1.326
virginica          6.588        2.974         5.552        2.026

==================================================
```

Задание 5. Найдите признак с наибольшим количеством выбросов, используя "ящик с усами" (box plot).

```
# 5. Ящик с усами
plt.figure(figsize=(10, 6))
sns.boxplot(x='variety', y='petal.length', data=df)
plt.title("Распределение длины лепестка по видам")
plt.show()
```



Задание 6. Выполните стандартизацию всех числовых признаков.

```
# 6. Стандартизация данных
features = ['sepal.length', 'sepal.width', 'petal.length', 'petal.width']

scaler = StandardScaler()
df_scaled = df.copy()
df_scaled[features] = scaler.fit_transform(df[features])

print("Данные после стандартизации:")
print(df_scaled.head())
```

Данные после стандартизации:

	sepal.length	sepal.width	petal.length	petal.width	variety
0	-0.900681	1.019004	-1.340227	-1.315444	setosa
1	-1.143017	-0.131979	-1.340227	-1.315444	setosa
2	-1.385353	0.328414	-1.397064	-1.315444	setosa
3	-1.506521	0.098217	-1.283389	-1.315444	setosa
4	-1.021849	1.249201	-1.340227	-1.315444	setosa

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.