Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ Информатика и системы управления

КАФЕДРА Системы обработки информации и управления

Отчет по лабораторной работе № 1

«Разведочный анализ данных. Исследование и визуализация данных»

по дисциплине «Технологии машинного обучения»

Студент ИУ5-65Б _____ Д.А. Шиленок
(Группа) (Подпись, дата) (И.О.Фамилия)

Преподаватель _____ Ю.Е. Гапанюк
(Подпись, дата) (И.О.Фамилия)

Москва

2025

Цель лабораторной работы: изучение различных методов визуализация данных.

Задание

- Выбрать набор данных (датасет).
- Создать ноутбук, который содержит следующие разделы:
    - Текстовое описание выбранного Вами набора данных.
    - Основные характеристики датасета.
    - Визуальное исследование датасета.
    - Информация о корреляции признаков.

Текст программы

Лабораторная работа №1

Разведочный анализ данных. Исследование и визуализация данных

Шиленок Даниил ИУ5-65Б

Описание выбранного набора данных

Набор данных Iris (Ирисы Фишера) — это классический датасет в машинном обучении, используемый для задачи классификации. Он содержит измерения длины и ширины чашелистиков и лепестков трёх видов ирисов:
Setosa
Versicolor
Virginica

Основные характеристики датасета

Количество объектов: 150 (50 от каждого класса)

Количество признаков: 4

Признаки:

- sepal length (длина чашелистика)
- sepal width (ширина чашелистика)
- petal length (длина лепестка)
- petal width (ширина лепестка)

Целевая переменная:

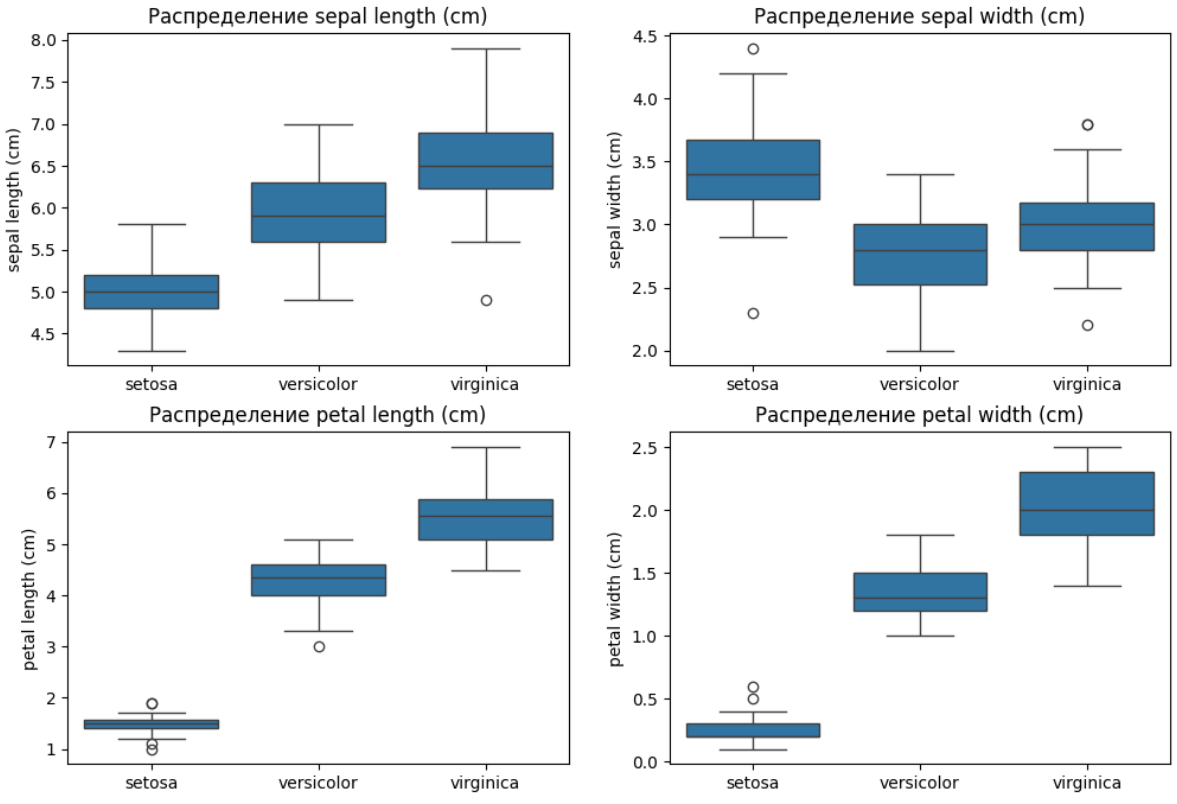
- Вид ириса (target)

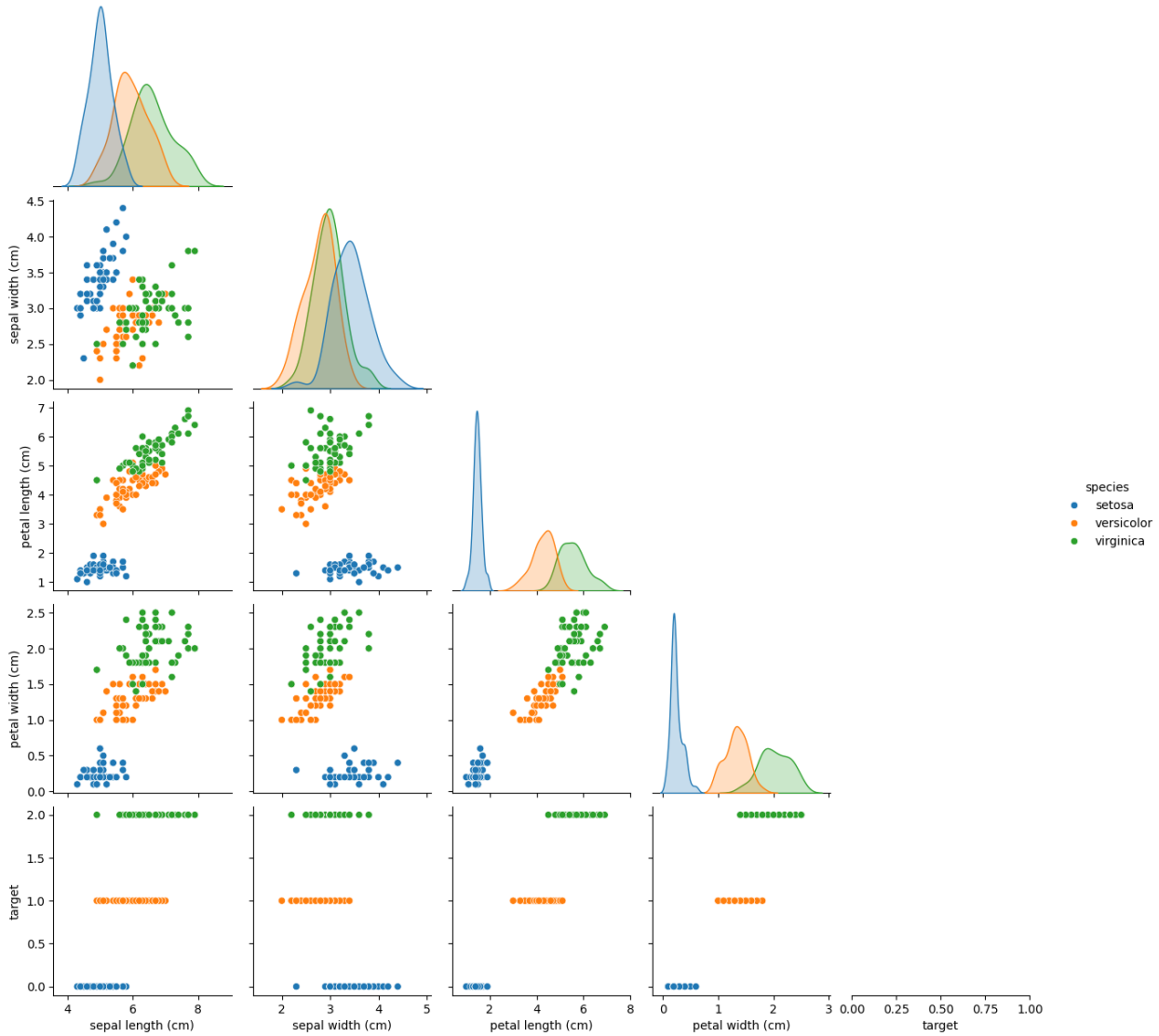
```
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
import pandas as pd

# Загрузка датасета
iris = load_iris()
df = pd.DataFrame(data=iris.data, columns=iris.feature_names)
df['target'] = iris.target
df['species'] = df['target'].map({i: species for i, species in enumerate(iris.target_names)})

# Вывод основных характеристик
print(iris.DESCR)
```

```
.. _iris_dataset:

Iris plants dataset
--------------------

**Data Set Characteristics:**

:Number of Instances: 150 (50 in each of three classes)
:Number of Attributes: 4 numeric, predictive attributes and the class
:Attribute Information:
    - sepal length in cm
    - sepal width in cm
    - petal length in cm
    - petal width in cm
    - class:
            - Iris-Setosa
            - Iris-Versicolour
            - Iris-Virginica
```

Визуальное исследование датасета

```
plt.figure(figsize=(12, 8))
for i, column in enumerate(df.columns[:4]):
    plt.subplot(2, 2, i+1)
    sns.boxplot(data=df, x='species', y=column)
    plt.title(f'Распределение {column}')
    plt.xlabel('')
plt.show()

sns.pairplot(df, hue='species', corner=True)
plt.show()
```




species
setosa
versicolor
virginica
sepal length (cm)
sepal width (cm)
petal length (cm)
petal width (cm)
target

Информация о корреляции признаков

```
import numpy as np

# Матрица корреляций
correlation_matrix = df.iloc[:, :-2].corr()

# Визуализация матрицы
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title("Корреляционная матрица признаков")
plt.show()
```