Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»


Лабораторная работа №1
по дисциплине
# «Методы машинного обучения»
на тему

## «Истории о данных»

Выполнил:
студент группы ИУ5и-24М
Аунг Пьио Нанда

Москва — 2024 г.

## 1.Цель лабораторной работы

Изучение различных методов визуализация данных и создание истории на основе данных.

## 2.Задание

Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:

1. История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
2. На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
3. Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
4. Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
5. История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

• Сформировать отчет и разместить его в своем репозитории на GitHub

# 3. Ход выполнения работы

**Шаг 1:** Загрузка и первичный анализ данных

На этом шаге мы загрузим данные и посмотрим на первые строки датасета.

[28]:
```
import pandas as pd
data = pd.read_csv("D:\\ИУ-5 2сем\\ММО\\Iris.csv")
data.head()
```

[28]:

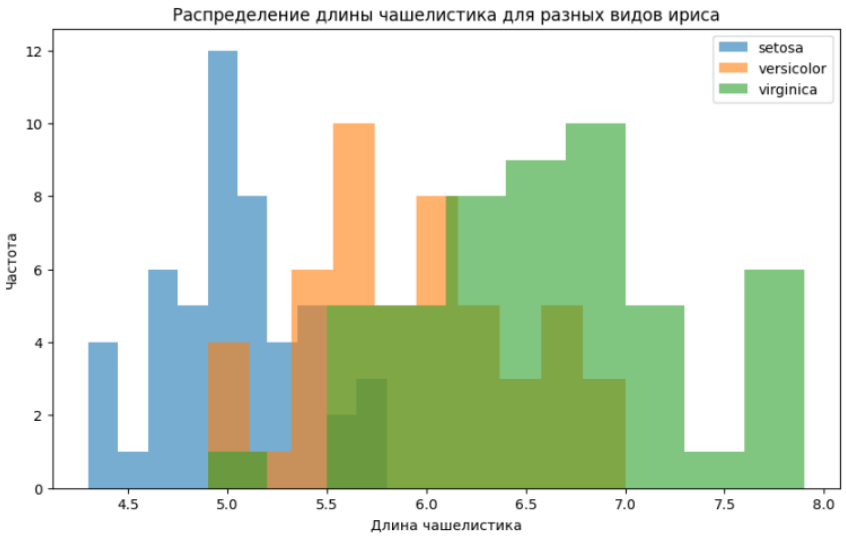
| | sepal_length | sepal_width | petal_length | petal_width | species |
|---|---|---|---|---|---|
| **0** | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| **1** | 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| **2** | 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| **3** | 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| **4** | 5.0 | 3.6 | 1.4 | 0.2 | setosa |

**Шаг 2:** Построение гистограммы распределения длины чашелистика для каждого вида ириса.

На этом шаге мы построим гистограмму, чтобы увидеть распределение длины чашелистика для каждого вида ириса.

[29]:
```
import matplotlib.pyplot as plt
plt.figure(figsize=(10, 6))
for species in data['species'].unique():
    subset = data[data['species'] == species]
    plt.hist(subset['sepal_length'], alpha=0.6, label=species)

plt.legend()
plt.xlabel('Длина чашелистика')
plt.ylabel('Частота')
plt.title('Распределение длины чашелистика для разных видов ириса')
plt.show()
```

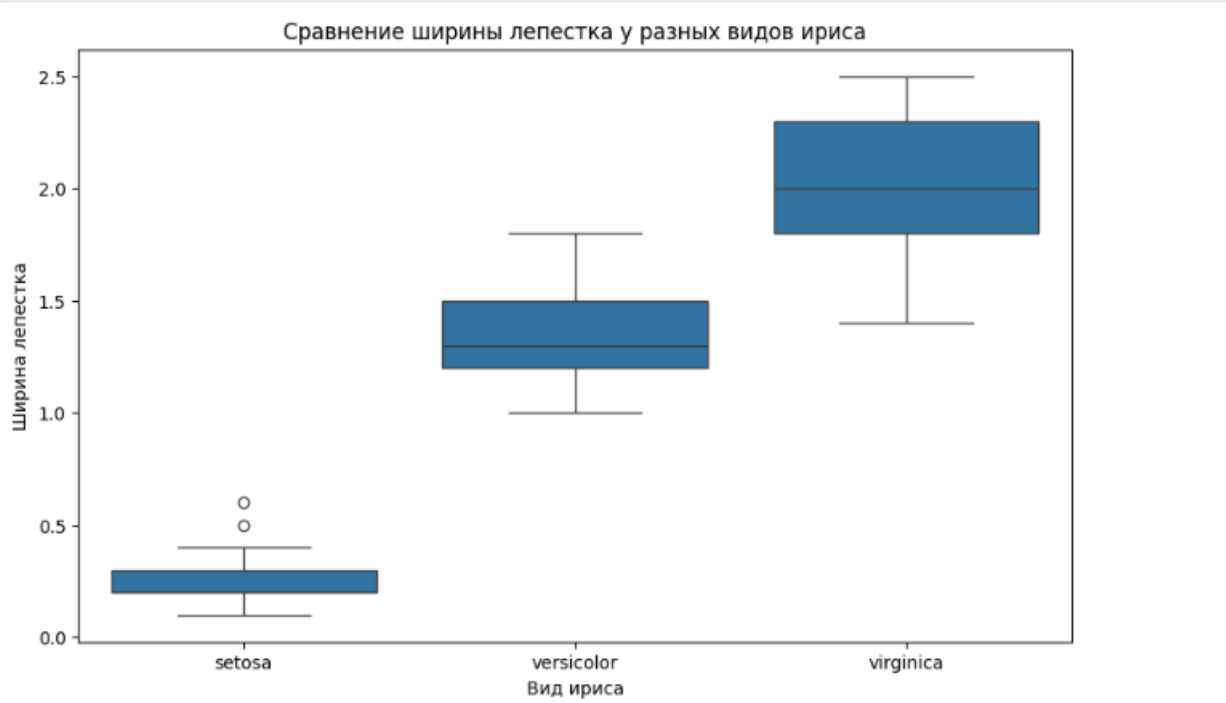


На гистограмме видно, что Iris-setosa имеет меньшую длину чашелистика по сравнению с другими видами ириса.

**Шаг 3:** Построение ящика с усами для сравнения ширины лепестка у разных видов ириса

На этом шаге мы построим ящик с усами для сравнения ширины лепестка у разных видов ириса.

```
[30]: import seaborn as sns
      # Ящик с усами для сравнения ширины лепестка
      plt.figure(figsize=(10, 6))
      sns.boxplot(x='species', y='petal_width', data=data)
      plt.xlabel('Вид ириса')
      plt.ylabel('Ширина лепестка')
      plt.title('Сравнение ширины лепестка у разных видов ириса')
      plt.show()
```

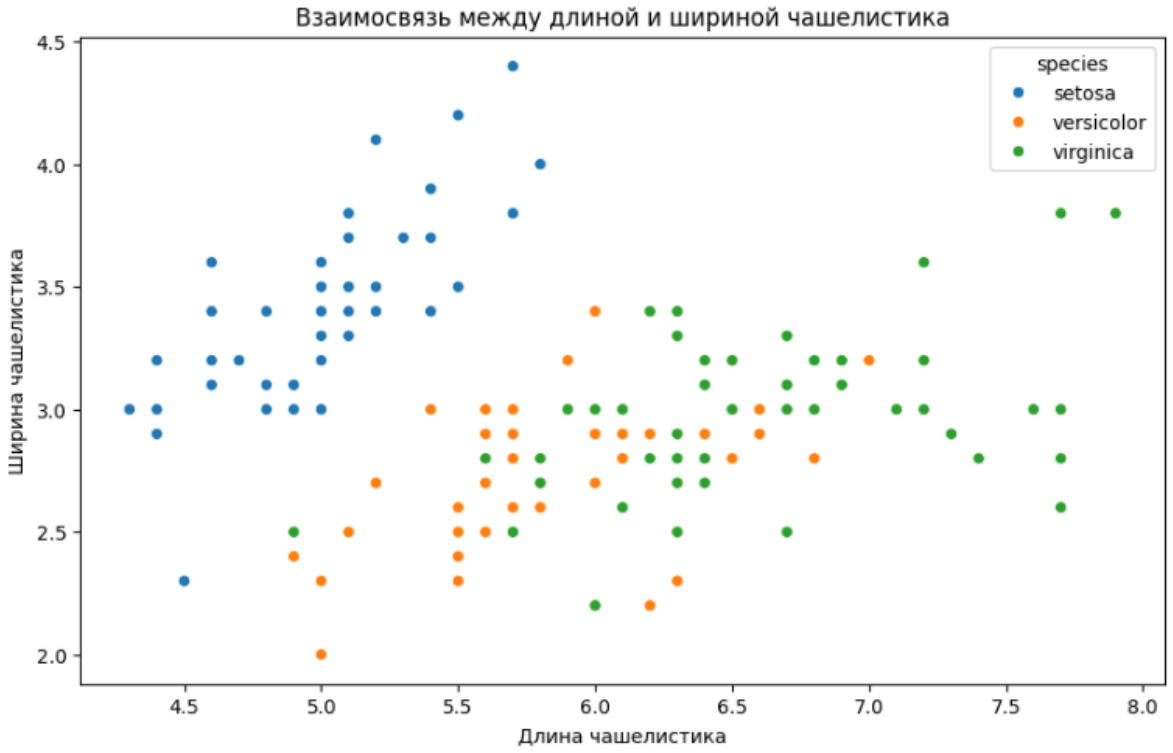



По ящику с усами видно, что Iris-versicolor имеет наибольшую ширину лепестка среди всех видов ириса.

**Шаг 4:** Построение точечной диаграммы для визуализации взаимосвязи между длиной чашелистика и шириной чашелистика

На этом шаге мы построим точечную диаграмму для визуализации взаимосвязи между длиной и шириной чашелистика.

```
[31]: # Точечная диаграмма для взаимосвязи между длиной и шириной чашелистика
plt.figure(figsize=(10, 6))
sns.scatterplot(x='sepal_length', y='sepal_width', data=data, hue='species')
plt.xlabel('Длина чашелистика')
plt.ylabel('Ширина чашелистика')
plt.title('Взаимосвязь между длиной и шириной чашелистика')
plt.show()
```

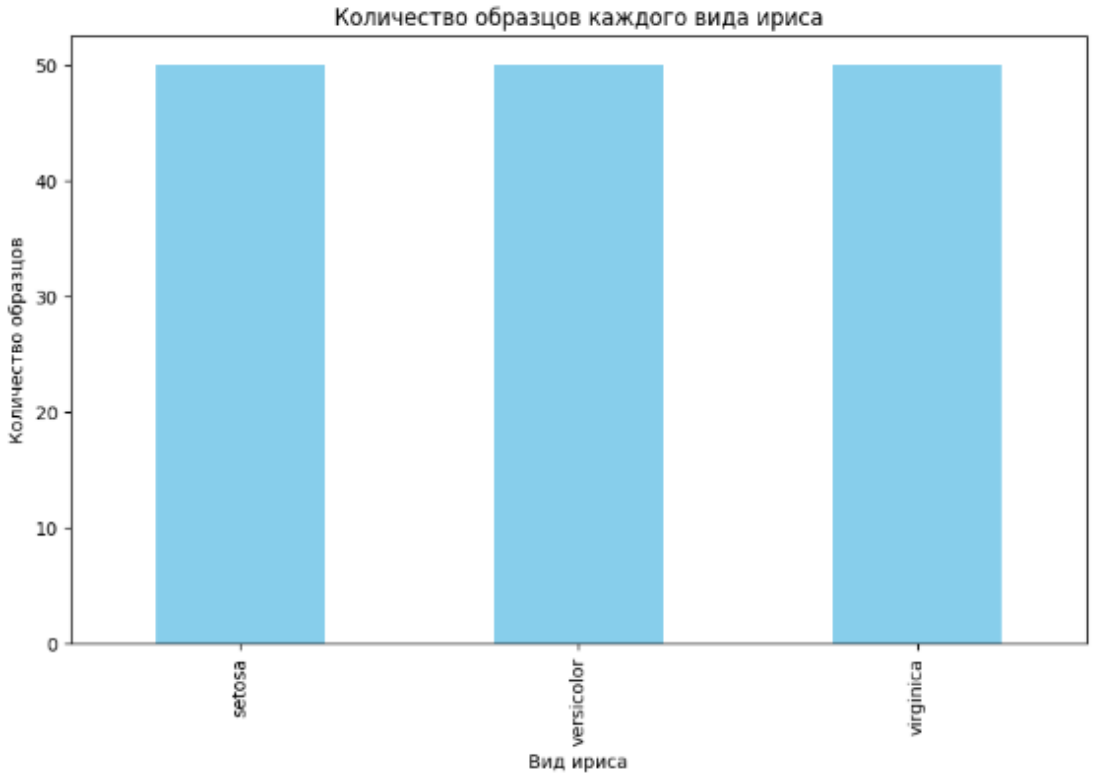



По точечной диаграмме видно, что есть определенная зависимость между длиной и шириной чашелистика для каждого вида ириса.

**Шаг 5:** Построение столбчатой диаграммы для подсчета количества образцов каждого вида ириса

На этом шаге мы построим столбчатую диаграмму для подсчета количества образцов каждого вида ириса.

[32]:
```
# Столбчатая диаграмма для подсчета количества образцов каждого вида ириса
plt.figure(figsize=(10, 6))
data['species'].value_counts().plot(kind='bar', color='skyblue')
plt.xlabel('Вид ириса')
plt.ylabel('Количество образцов')
plt.title('Количество образцов каждого вида ириса')
plt.show()
```




По столбчатой диаграмме видно, что каждый вид ириса представлен примерно одинаковым количеством образцов в наборе данных.

**выводы:**

1. Iris-setosa имеет наименьшую длину чашелистика среди всех видов ириса.

2. Iris-versicolor имеет наибольшую ширину лепестка среди всех видов ириса.

3. Существует определенная зависимость между длиной и шириной чашелистика для каждого вида ириса.

4. Каждый вид ириса представлен примерно одинаковым количеством образцов в наборе данных.

# Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Разведочный анализ данных. Исследование и визуализация данных» [Электронный ресурс] // GitHub. — 2024. — Режим доступа: https://github.com/ugapanyuk/courses_current/wiki/LAB_MMO__DATA_STORY.