**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 ИУ5 «Системы обработки информации и управления»**

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №1

«Технологии разведочного анализа и обработки данных»

Вариант №1

Выполнила:

студентка группы ИУ5-62Б

Андреева А.А.

Преподаватель:

Гапанюк Ю. Е.

2023 г.

## Выполнение работы

Для выполнения задачи проведения корреляционного анализа данных был представлен набор данных sklearn iris dataset

```
import numpy as np
import pandas as pd
import itertools
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn
```

```
from sklearn.datasets import load_iris
iris = load_iris()
```

```
iris.data.shape
```

```
(150, 4)
```

```
iris.feature_names
```

```
['sepal length (cm)',
 'sepal width (cm)',
 'petal length (cm)',
 'petal width (cm)']
```

Был создан датафрейм

```
iris_df = pd.DataFrame(iris.data, columns=iris.feature_names)
iris_df.head()
```

| | sepal length (cm) | sepal width (cm) | petal length (cm) | petal width (cm) |
|---|---|---|---|---|
| **0** | 5.1 | 3.5 | 1.4 | 0.2 |
| **1** | 4.9 | 3.0 | 1.4 | 0.2 |
| **2** | 4.7 | 3.2 | 1.3 | 0.2 |
| **3** | 4.6 | 3.1 | 1.5 | 0.2 |
| **4** | 5.0 | 3.6 | 1.4 | 0.2 |

Типы данных всех полей являются числовыми

```
iris_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 4 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   sepal length (cm)  150 non-null    float64
 1   sepal width (cm)   150 non-null    float64
 2   petal length (cm)  150 non-null    float64
 3   petal width (cm)   150 non-null    float64
dtypes: float64(4)
memory usage: 4.8 KB
```

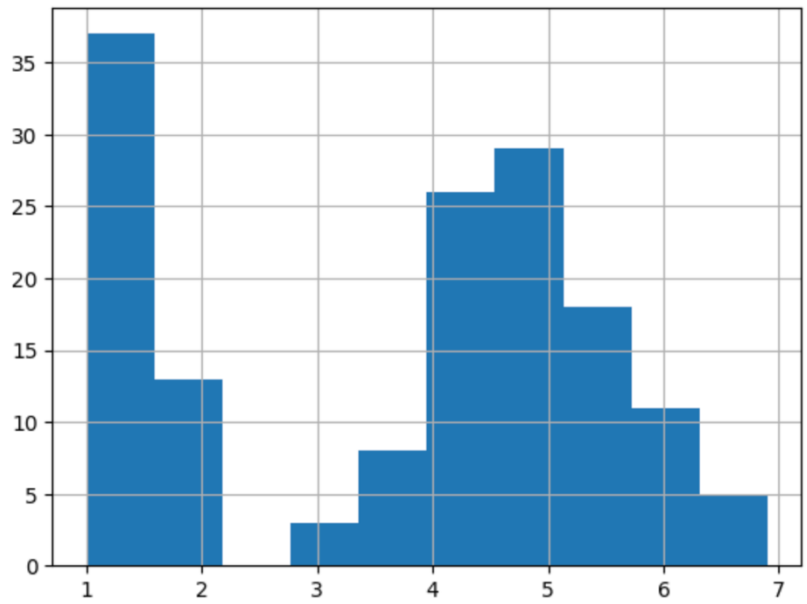
В наборе данных отсутствуют пропуски

```
iris_df.isna().sum()
```

```
sepal length (cm)    0
sepal width (cm)     0
petal length (cm)    0
petal width (cm)     0
dtype: int64
```

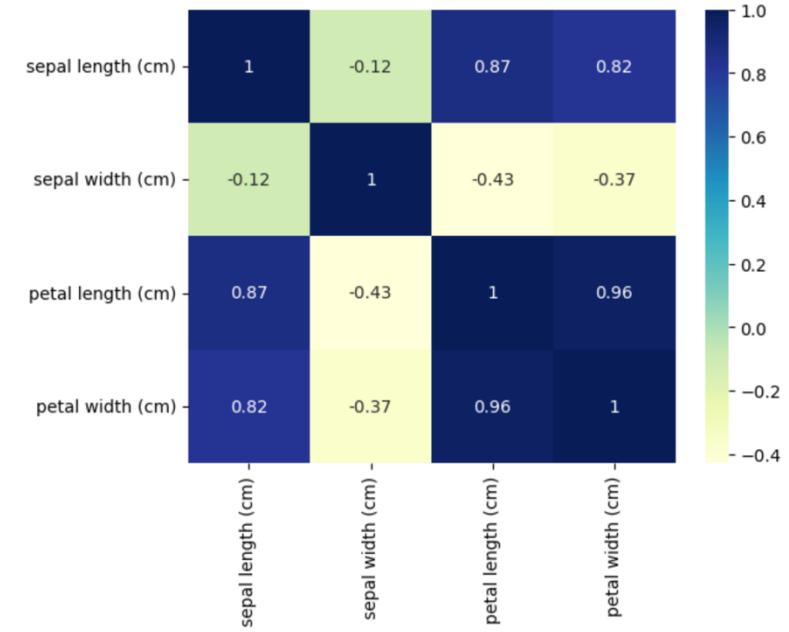
Для колонки «длина лепестка» была построена гистограмма

```
: iris_df['petal length (cm)'].hist();
```



Для визуализации корреляционной матрицы была использована «тепловая карта»

```
dataplot = sns.heatmap(iris_df.corr(), cmap="YlGnBu", annot=True)
```




Наиболее сильную зависимость можно заметить между переменными "petal length (cm)" и "petal width (cm)" , а также между "sepal length (cm)" и "petal length (cm)" и "sepal length (cm)" и "petal width (cm)". Эти признаки будут наиболее информативными при построении моделей машинного обучения.

Таким образом, на основе признаков "petal length (cm)", "petal width (cm)", "sepal length (cm)" могут быть построены модели машинного обучения.

Так как в представленном датасете не были обнаружены пропуски, возьмем дополнительный датасет, содержащий пропуски.

Для этого был взят набор данных, содержащий данные о фильмах на платформе Disney+.

```
df = pd.read_csv("disney_plus_shows.csv")
```

```
df.head()
```

| | imdb_id | title | plot | type | rated | year | released_at | added_at | runtime | genre | director | writer | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | tt0147800 | 10 Things I Hate About You | A pretty, popular teenager can't go out on a d... | movie | PG-13 | 1999 | 31 Mar 1999 | November 12, 2019 | 97 min | Comedy, Drama, Romance | Gil Junger | Karen McCullah, Kirsten Smith | Ju |
| 1 | tt7019028 | 101 Dalmatian Street | This series follows the lives of Delilah and D... | series | NaN | 2018– | 25 Mar 2019 | February 28, 2020 | NaN | Animation, Comedy, Family | NaN | NaN | D |
| | | 101 | An evil high-fashion | | | | 27 Nov | November | | Adventure, Comedy | Stephen | Dodie Smith (novel) | C |

Посчитаем количество пропусков

```
df.isna().sum()
```

```
imdb_id          98
title            98
plot            126
type             98
rated           250
year             98
released_at     118
added_at          0
runtime         154
genre           107
director        303
writer          249
actors          122
language        136
country         123
awards          436
metascore       700
imdb_rating     113
imdb_votes      113
dtype: int64
```

Удалим пропущенные значения

```
df = df.dropna()
```

```
df.isna().sum()
```

```
imdb_id         0
title           0
plot            0
type            0
rated           0
year            0
released_at     0
added_at        0
runtime         0
genre           0
director        0
writer          0
actors          0
language        0
country         0
awards          0
metascore       0
imdb_rating     0
imdb_votes      0
dtype: int64
```