РК1

In [1]:

```
import numpy as np
import pandas as pd
from sklearn.datasets import *
```

In [2]:

```
iris = load_iris()
```

In [3]:

```
iris = pd.DataFrame(data=np.c_[iris['data'], iris['target']], columns=iris['feature_names']+['target'])
```

In [4]:

```
iris
```

Out[4]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	0.0
1	4.9	3.0	1.4	0.2	0.0
2	4.7	3.2	1.3	0.2	0.0
3	4.6	3.1	1.5	0.2	0.0
4	5.0	3.6	1.4	0.2	0.0
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	2.0
146	6.3	2.5	5.0	1.9	2.0
147	6.5	3.0	5.2	2.0	2.0
148	6.2	3.4	5.4	2.3	2.0
149	5.9	3.0	5.1	1.8	2.0

150 rows × 5 columns

Задание:

- Провести корреляционный анализ.
- В случае наличия пропусков в данных удалить строки или колонки, содержащие пропуски.
- Сделать выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

Кореляционный анализ

Таблица корреляции

In [5]:

```
iris.corr()
```

Out[5]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
sepal length (cm)	1.000000	-0.117570	0.871754	0.817941	0.782561
sepal width (cm)	-0.117570	1.000000	-0.428440	-0.366126	-0.426658
petal length (cm)	0.871754	-0.428440	1.000000	0.962865	0.949035
petal width (cm)	0.817941	-0.366126	0.962865	1.000000	0.956547
target	0.782561	-0.426658	0.949035	0.956547	1.000000

Тепловая карта

In [6]:

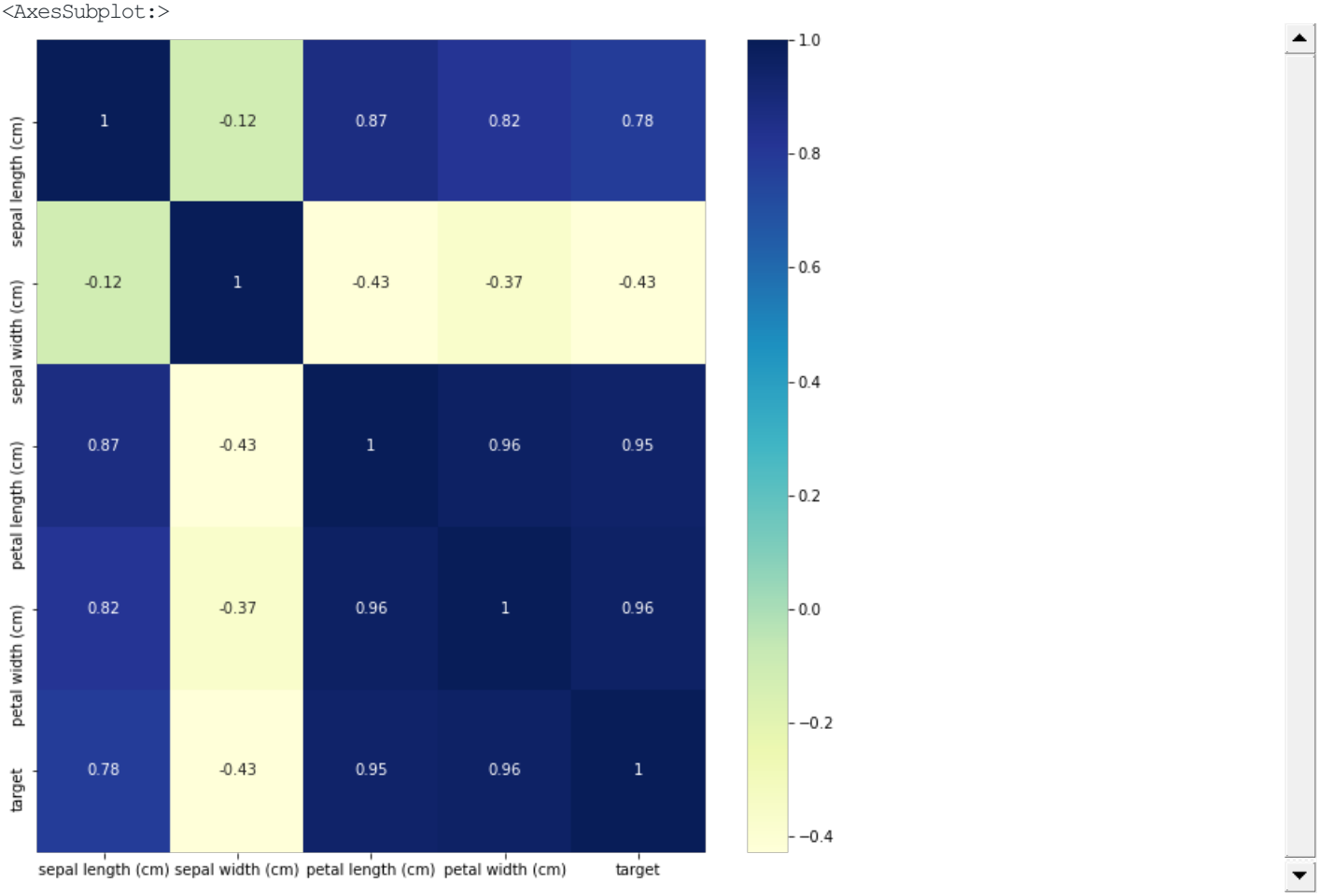
```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [7]:

```
fig, ax = plt.subplots(figsize=(10,10))
sns.heatmap(iris.corr(), ax=ax, annot=True, cmap="YlGnBu")
```

Out[7]:

```
<AxesSubplot:>
```



- Очень сильно (положительно) коррелируют признаки petal width и petal length (0.96) и целевой признак с petal width и petal length (0.96 и 0.95 соответственно).
- Достаточно сильно (положительно) коррелируют признаки sepal length и petal length (0.87), sepal length и petal width (0.82), целевая фича и sepal length (0.78).
- Довольно слабо (отрицательно) коррелируют petal length и sepal width (-0.43), petal width и sepal width (-0.37), целевой признак и sepal width (-0.43).
- Слабая отрицательная корреляция наблюдается между признаками sepal width и sepal length (-0.12).

Пропуски в данных

Поскольку в задании используется игрушечный датасет, никаких пропусков не должно быть.

In [8]:

```
iris.isnull().any()
```

Out[8]:

```
sepal length (cm)    False
sepal width (cm)     False
petal length (cm)    False
petal width (cm)     False
target               False
dtype: bool
```

Наше предположение подтвердилось, удалять нечего.

Возможный вклад признаков в модель и возможность построения модели

Очевидно, поскольку датасет игрушечный, никаких серьёзных проблем с ним нет и быть не может.

Для построения наиболее эффективно работающей модели необходимо убедиться, что никакая пара признаков в наборе данных не коррелирует слишком сильно. Таким образом, можно было бы убрать из модели колонку petal width или petal length, поскольку они очень сильно коррелируют как между собой, так и с другими колонками.

Jointplot

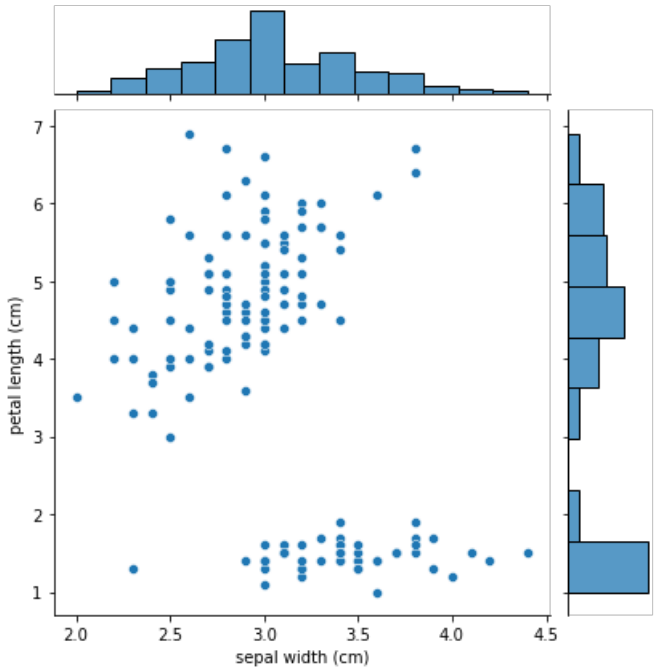
Построим график jointplot для признаков sepal width и petal length.

In [9]:

```
sns.jointplot(data=iris,x="sepal width (cm)",y="petal length (cm)")
```

Out[9]:

```
<seaborn.axisgrid.JointGrid at 0x2393c6a0d00>
```



In [ ]: