Московский государственный технический университет имени Н. Э. Баумана

Кафедра «Системы обработки информации и управления»

Лабораторная работа №1

по курсу

«Методы машинного обучения»

на тему:

«Разведочный анализ данных. Исследование и визуализация данных»

Выполнил:

Студент ИУ5-24М

Черната Н. С.

Москва, 2020

Задание

- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.
- Для первой лабораторной работы рекомендуется использовать датасет без пропусков в данных, например из Scikit-learn.
- Пример преобразования датасетов Scikit-learn в Pandas Dataframe можно посмотреть здесь.

Для лабораторных работ не рекомендуется выбирать датасеты большого размера.

- Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета. Необходимо использовать не менее 2 различных библиотек и не менее 5 графиков.
4. Информация о корреляции признаков.

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

The famous Iris database, first used by Sir R.A. Fisher. The dataset is taken from Fisher's paper. Note that it's the same as in R, but not as in the UCI Machine Learning Repository, which has two wrong data points.

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

In [2]:
```
from sklearn.datasets import load_iris
```

In [3]:
```
raw_data = load_iris()
features = pd.DataFrame(data=raw_data['data'],columns=raw_data['feature_names'])
data = features
data['target']=raw_data['target']
data['class']=data['target'].map(lambda ind: raw_data['target_names'][ind])
data.head()
```

Out[3]:

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

In [4]:
```
data.describe()
```

Out[4]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333	1.000000
std	0.828066	0.435866	1.765298	0.762238	0.819232
min	4.300000	2.000000	1.000000	0.100000	0.000000
25%	5.100000	2.800000	1.600000	0.300000	0.000000
50%	5.800000	3.000000	4.350000	1.300000	1.000000
75%	6.400000	3.300000	5.100000	1.800000	2.000000
max	7.900000	4.400000	6.900000	2.500000	2.000000

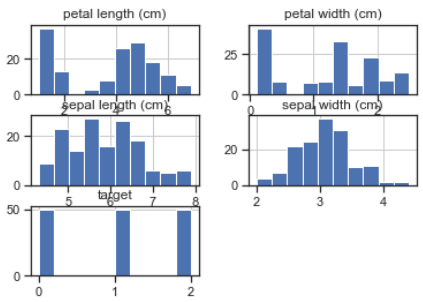
In [5]:
```
data.shape
```

Out[5]:
```
(150, 6)
```

Датасет включает в себя 5 атрибутов: Sepal length (cm) Sepal width (cm) Petal length (cm) Petal width (cm) Target

In [6]:
```
data.hist()
```

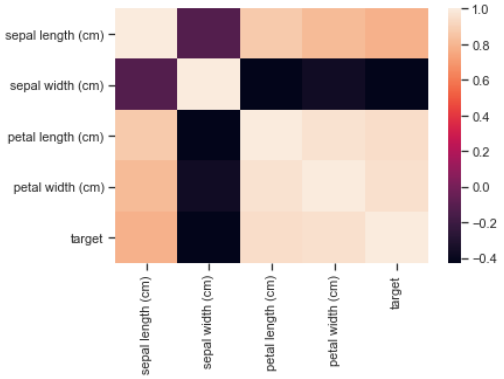
Out[6]:
```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x00000237606DD648>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000023762784408>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x00000237627BE288>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x00000237637C2FC8>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x00000237637FBF88>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x0000023763834E08>]],
      dtype=object)
```



In [7]:
```
corr = data.corr()
```

In [8]:
```
sns.heatmap(corr,
        xticklabels=corr.columns,
        yticklabels=corr.columns)
```

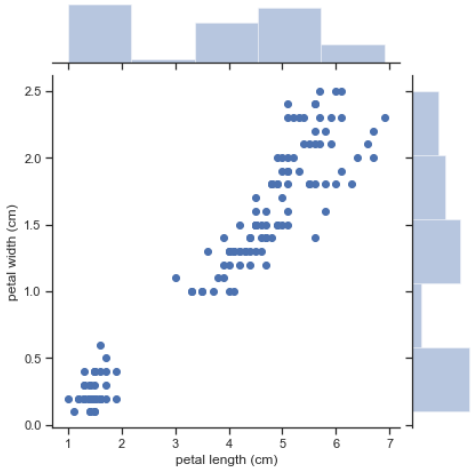
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x237639b8808>



Наибольшая корреляция наблюдается между парами параметров petal width (cm) - petal length (cm) и petal width (cm) - target

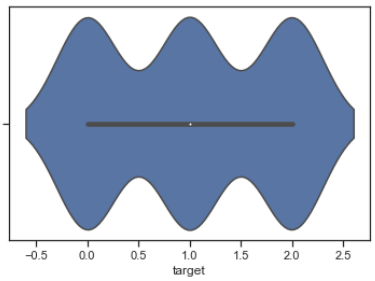
In [18]:
```
sns.jointplot(x='petal length (cm)', y='petal width (cm)', data=data)
```

Out[18]: <seaborn.axisgrid.JointGrid at 0x23764324e48>



In [19]:
```
sns.violinplot(x=data['target'])
```

Out[19]: <matplotlib.axes._subplots.AxesSubplot at 0x23763f1ff88>



In [21]:
```
for i in data.target.unique():
    sns.distplot(data['sepal length (cm)'][data.target==i],
                 kde=1,label='{}'.format(i))

plt.legend()
```

Out[21]: <matplotlib.legend.Legend at 0x23763b46dc8>