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»

Отчет по РК № 1

Технологии машинного обучения

Студент:

группы ИУ5-64Б

Ведьгун Е.А.

2021 г.

Москва

Задача №1

Для заданного набора данных проведите корреляционный анализ. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски. Сделайте выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

Набор данных:

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html#sklearn.datasets.load_iris

Дополнительные требования по группам:

Для студентов группы ИУ5-64Б - для произвольной колонки данных построить график "Скрипичная диаграмма (violin plot)".

1) Основные характеристики датасета

```
# импорт библиотек
from sklearn.datasets import load_iris
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

```
# преобразование и загрузка данных
iris = load_iris()
X = iris.data
target = iris.target
names = iris.target_names
df = pd.DataFrame(X, columns=iris.feature_names)
df['species'] = iris.target
df['species'] = df['species'].replace(to_replace= [0, 1, 2], value = ['setosa', 'versicolor', 'virginica'])
```

```
# первые 5 столбцов таблицы
df.head()
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	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

```
# проверим есть ли пропущенные значения
df.isnull().sum()
```

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

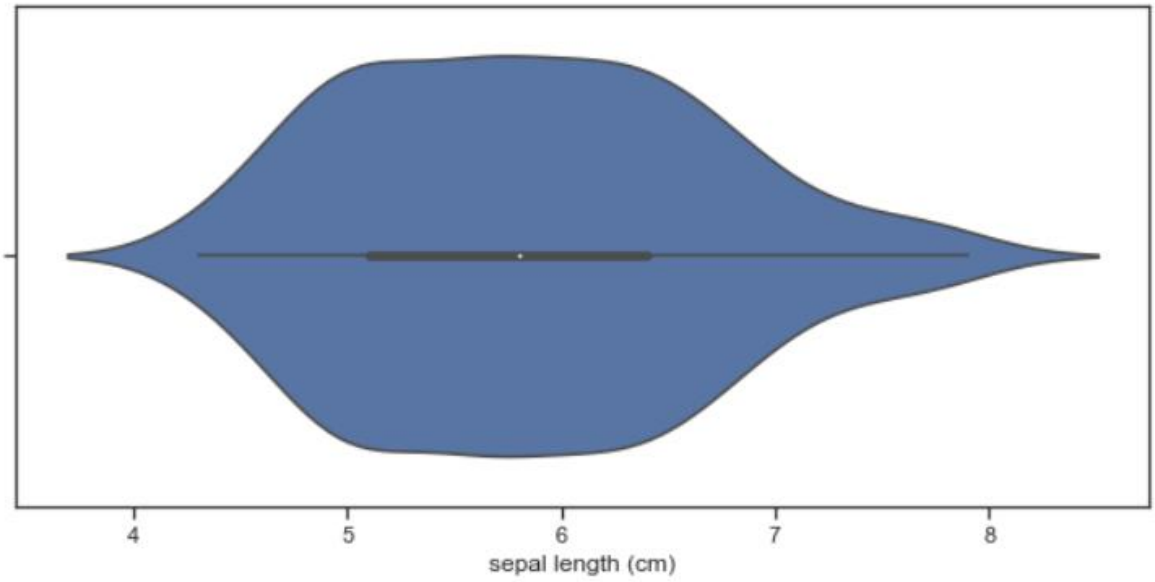
```
# размер датасета
df.shape
```

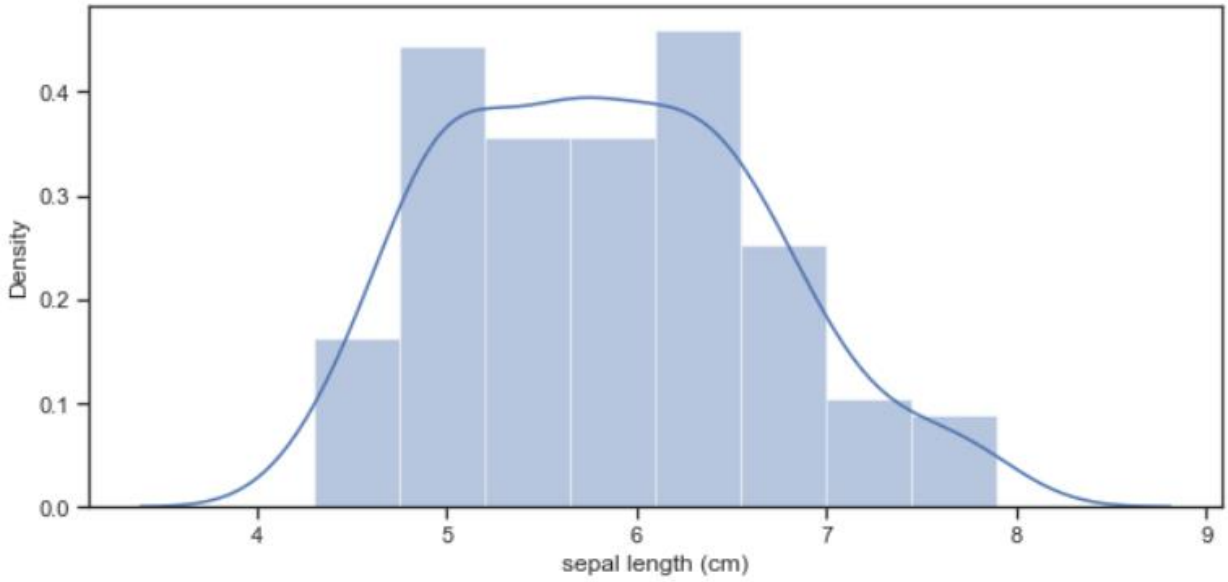
```
(150, 5)
```

2) Визуальный анализ данных

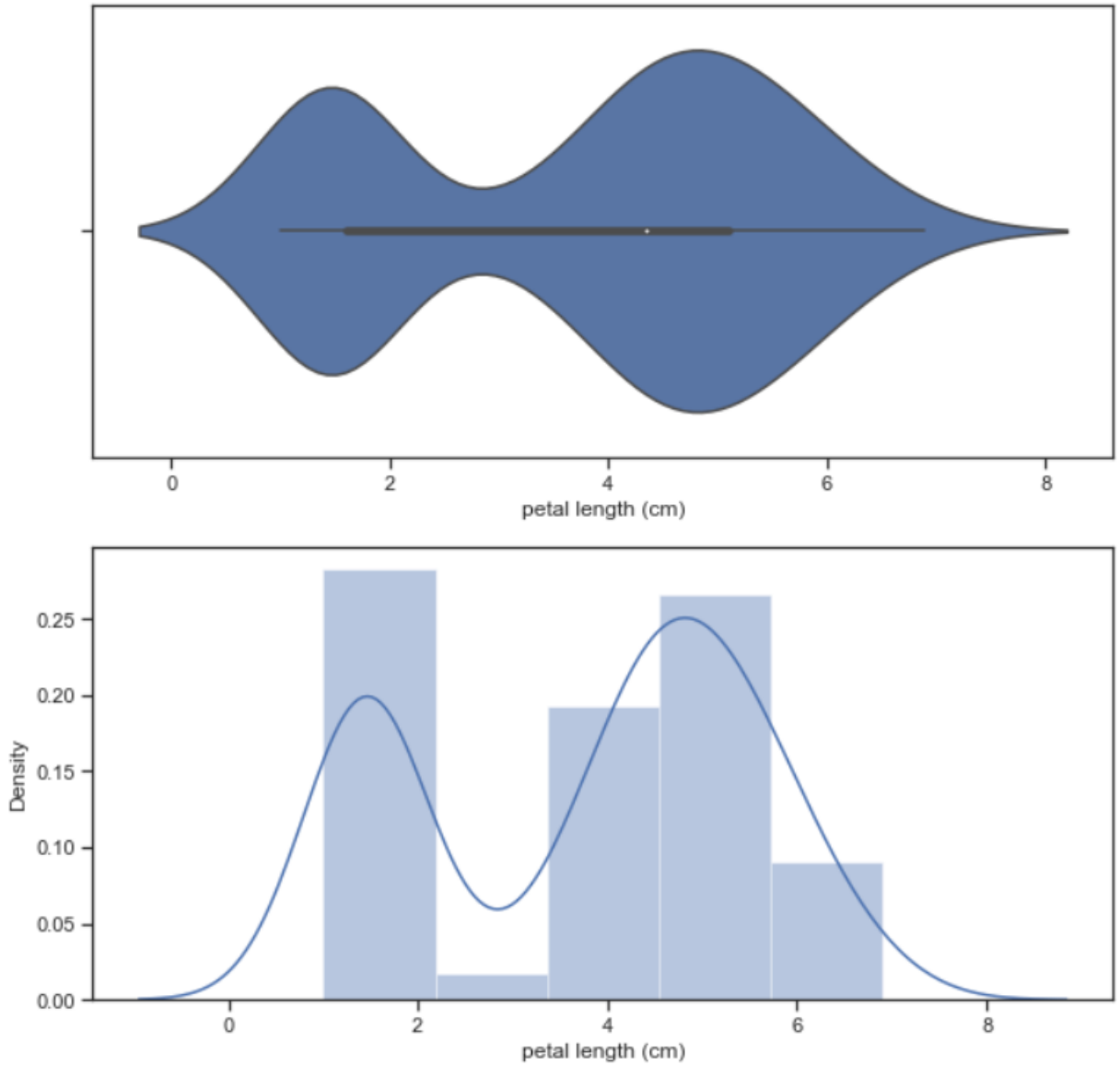
Скрипичная диаграмма

```
fig, ax = plt.subplots(2, 1, figsize=(10,10))
sns.violinplot(ax=ax[0], x=df['sepal length (cm)'])
sns.distplot(df['sepal length (cm)'], ax=ax[1])
```


4
5
6
7
8
sepal length (cm)


Density
0.0
0.1
0.2
0.3
0.4
4
5
6
7
8
9
sepal length (cm)

```
fig, ax = plt.subplots(2, 1, figsize=(10,10))
sns.violinplot(ax=ax[0], x=df['petal length (cm)'])
sns.distplot(df['petal length (cm)'], ax=ax[1])
```



3) Информация о корреляции признаков

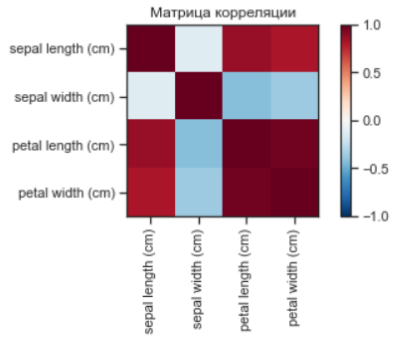
```
df.corr()
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
sepal length (cm)	1.000000	-0.117570	0.871754	0.817941
sepal width (cm)	-0.117570	1.000000	-0.428440	-0.366126
petal length (cm)	0.871754	-0.428440	1.000000	0.962865
petal width (cm)	0.817941	-0.366126	0.962865	1.000000

```
feature_names = iris["feature_names"]
ax = plt.axes()
im = ax.imshow(np.corrcoef(X.T), cmap="RdBu_r", vmin=-1, vmax=1)

ax.set_xticks([0, 1, 2, 3])
ax.set_xticklabels(list(feature_names), rotation=90)
ax.set_yticks([0, 1, 2, 3])
ax.set_yticklabels(list(feature_names))

plt.colorbar(im).ax.set_ylabel("", rotation=0)
ax.set_title("Матрица корреляции")
plt.tight_layout()
```



Возьмем в качестве целевого признака 'sepal length (cm)', тогда на основе корреляционной матрицы можно сделать следующие выводы:

- Целевой признак наиболее сильно коррелирует с petal length (0.87) и petal width (0.82). Эти признаки обязательно следует оставить в модели.
- Целевой признак слабо коррелирует с sepal width. Скорее всего этот признак стоит исключить из модели, он только ухудшит ее качество.