Московский государственный технический университет им. Н. Э. Баумана

Курс «Технологии машинного обучения»

Отчёт по лабораторной работе №1

Выполнил:
Сергеев М. А.
группа ИУ5-64Б

Проверил:
Гапанюк Ю.Е.

Дата: 14.05.25

Дата:

Подпись:

Подпись:

Москва, 2025 г.

Цель лабораторной работы: изучение различных методов визуализация данных.

Краткое описание. Построение основных графиков, входящих в этап разведочного анализа данных.

Рекомендуемые инструментальные средства можно посмотреть здесь.

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.
4. Информация о корреляции признаков.

- Сформировать отчет и разместить его в своем репозитории на github.

Ход выполнения:

Каждый образец содержит следующие колонки:

1. **Alcohol** - содержание алкоголя.
2. **Malic acid** - содержание яблочной кислоты.
3. **Ash** - содержание золы.
4. **Alcalinity of ash** - щелочность золы.
5. **Magnesium** - содержание магния.
6. **Total phenols** - общее содержание фенолов.
7. **Flavanoids** - содержание флавоноидов.
8. **Nonflavanoid phenols** - содержание нефлавоноидных фенолов.
9. **Proanthocyanins** - содержание проантоцианидинов.
10. **Color intensity** - интенсивность цвета.
11. **Hue** - оттенок.
12. **OD280/OD315 of diluted wines** - оптическая плотность (отношение OD280/OD315).
13. **Proline** - содержание пролина.

Целевой признак:

- **Class** - класс вина (0, 1 или 2), соответствующий одному из трех культиваров.

1. Импорт необходимых библиотек

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

[1] ✓ 8.8s Python

Будем анализировать данные только на обучающей выборке

```
def make_dataframe(ds_function):
    ds = ds_function()
    df = pd.DataFrame(data= np.c_[ds['data'], ds['target']],
                      columns= list(ds['feature_names']) + ['target'])
    return df
data = make_dataframe(load_wine)
```

[27] ✓ 0.0s Python

2. Основные характеристики датасета

```
data.head()
```

✓ 0.0s Python

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od315_of_diluted_wines	proline	target
0	14.23	1.71	2.43	15.6	127.0	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065.0	0.0
1	13.20	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050.0	0.0
2	13.16	2.36	2.67	18.6	101.0	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185.0	0.0
3	14.37	1.95	2.50	16.8	113.0	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480.0	0.0
4	13.24	2.59	2.87	21.0	118.0	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735.0	0.0

```
data.shape
```

✓ 0.0s Python

```
(178, 14)
```

```
total_count = data.shape[0]
print('Всего строк: {}'.format(total_count))
```

✓ 0.0s Python

```
Всего строк: 178
```

```
# Список колонок
data.columns
```

[6] ✓ 0.0s Python

```
Index(['alcohol', 'malic_acid', 'ash', 'alcalinity_of_ash', 'magnesium',
       'total_phenols', 'flavanoids', 'nonflavanoid_phenols',
       'proanthocyanins', 'color_intensity', 'hue',
       'od280/od315_of_diluted_wines', 'proline', 'target'],
      dtype='object')
```

```
# Список колонок с типами данных
data.dtypes
```

[7] ✓ 0.0s Python

```
alcohol                         float64
malic_acid                      float64
ash                             float64
alcalinity_of_ash               float64
magnesium                       float64
total_phenols                   float64
flavanoids                      float64
nonflavanoid_phenols            float64
proanthocyanins                 float64
color_intensity                 float64
hue                             float64
od280/od315_of_diluted_wines    float64
proline                         float64
target                          float64
dtype: object
```

```
# Проверим наличие пустых значений
# Цикл по колонкам датасета
for col in data.columns:
    # Количество пустых значений - все значения заполнены
    temp_null_count = data[data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

8] ✓ 0.0s Python

```
alcohol - 0
malic_acid - 0
ash - 0
alcalinity_of_ash - 0
magnesium - 0
total_phenols - 0
flavanoids - 0
nonflavanoid_phenols - 0
proanthocyanins - 0
color_intensity - 0
hue - 0
od280/od315_of_diluted_wines - 0
proline - 0
target - 0
```

```
# Основные статистические характеристки набора данных
data.describe()
```

9] ✓ 0.0s Python

c_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od315_of_diluted_wines	proline	target
00000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000
36348	2.366517	19.494944	99.741573	2.295112	2.029270	0.361854	1.590899	5.058090	0.957449	2.611685	746.893258	0.938202
17146	0.274344	3.339564	14.282484	0.625851	0.998859	0.124453	0.572359	2.318286	0.228572	0.709990	314.907474	0.775035
40000	1.360000	10.600000	70.000000	0.980000	0.340000	0.130000	0.410000	1.280000	0.480000	1.270000	278.000000	0.000000
02500	2.210000	17.200000	88.000000	1.742500	1.205000	0.270000	1.250000	3.220000	0.782500	1.937500	500.500000	0.000000
65000	2.360000	19.500000	98.000000	2.355000	2.135000	0.340000	1.555000	4.690000	0.965000	2.780000	673.500000	1.000000
82500	2.557500	21.500000	107.000000	2.800000	2.875000	0.437500	1.950000	6.200000	1.120000	3.170000	985.000000	2.000000
00000	3.230000	30.000000	162.000000	3.880000	5.080000	0.660000	3.580000	13.000000	1.710000	4.000000	1680.000000	2.000000

```
# Определим уникальные значения для целевого признака
data['target'].unique()
```

10] ✓ 0.0s Python

```
array([0., 1., 2.])
```

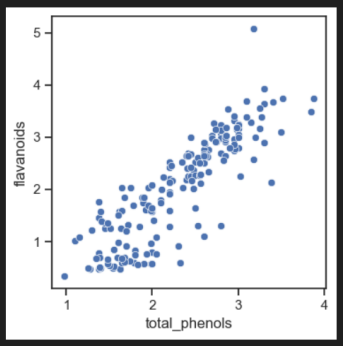
3. Визуальное исследование датасета

Диаграмма рассеяния

```
fig, ax = plt.subplots(figsize=(4,4))
sns.scatterplot(ax=ax, x='total_phenols', y='flavanoids', data=data)
```

[54] ✓ 0.0s Python

```
<Axes: xlabel='total_phenols', ylabel='flavanoids'>
```



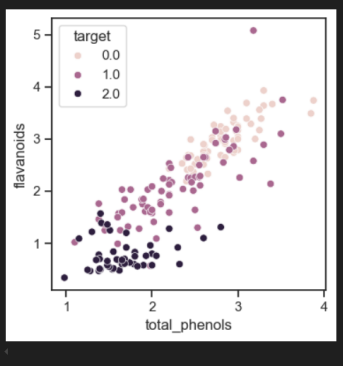
Можно заметить что **между содержанием фенола и флаваноидов сильная положительная корелляция**

Посмотрим на сколько эти величины влияют на целевой признак:

```
fig, ax = plt.subplots(figsize=(4,4))
sns.scatterplot(ax=ax, x='total_phenols', y='flavanoids', data=data, hue='target')
```

[5] ✓ 0.1s Python

```
<Axes: xlabel='total_phenols', ylabel='flavanoids'>
```



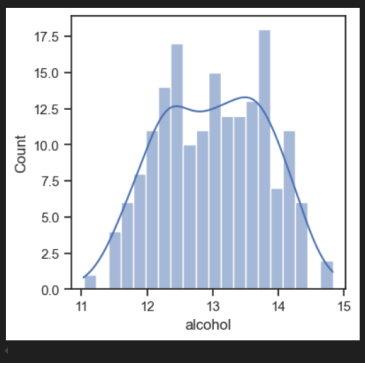
Видно что **чем выше класс вина, тем меньше в нем флаваноидов и фенола**

Гистограмма

```
fig, ax = plt.subplots(figsize=(4,4))
sns.histplot(data['alcohol'], kde=True, bins=20)
```

[56] ✓ 0.1s Python

```
<Axes: xlabel='alcohol', ylabel='Count'>
```

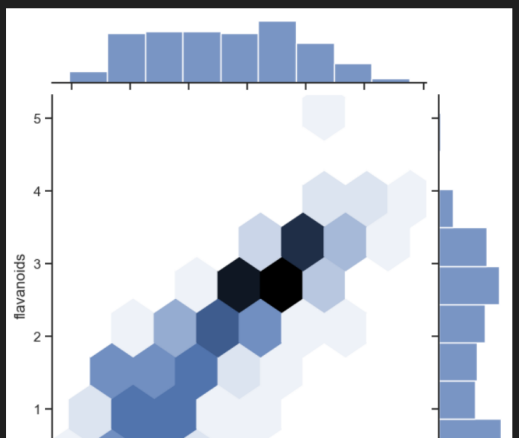


Jointplot

```
sns.jointplot(x='total_phenols', y='flavanoids', data=data, kind = 'hex')
```

✓ 0.2s Python

```
<seaborn.axisgrid.JointGrid at 0x2160f808500>
```

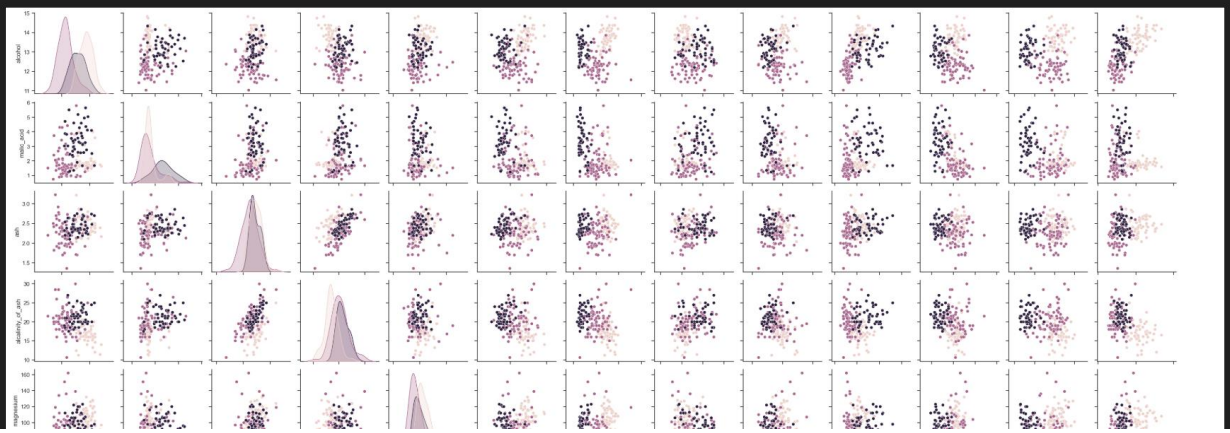


Парные диаграммы

```
sns.pairplot(data, hue="target")
```

[40] ✓ 29.8s Python

```
<seaborn.axisgrid.PairGrid at 0x2160d8338f0>
```

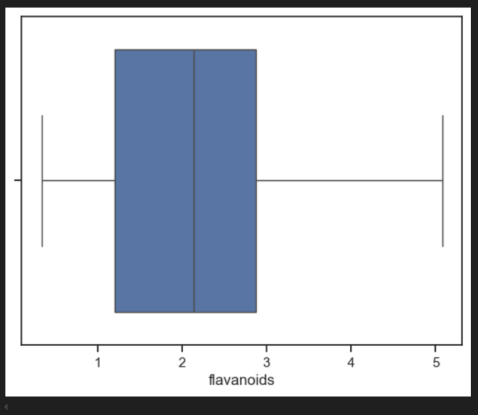


Ящик с усами

```
sns.boxplot(x=data['flavanoids'])
```

[46] ✓ 0.0s Python

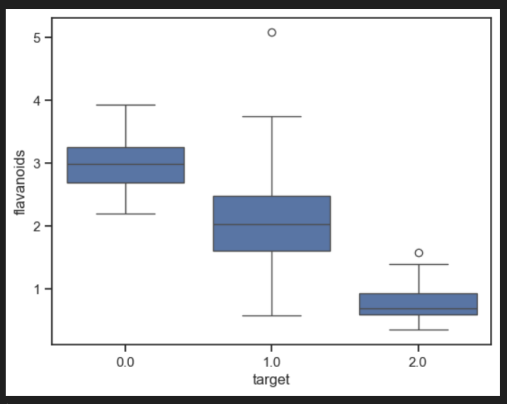
```
<Axes: xlabel='flavanoids'>
```



```
# Распределение параметра flavanoids сгруппированные по target.
sns.boxplot(x='target', y='flavanoids', data=data)
```

[47] ✓ 0.0s Python

```
<Axes: xlabel='target', ylabel='flavanoids'>
```



4. Информация о корелляции признаков

```
data.corr()
```

[49] ✓ 0.0s Python

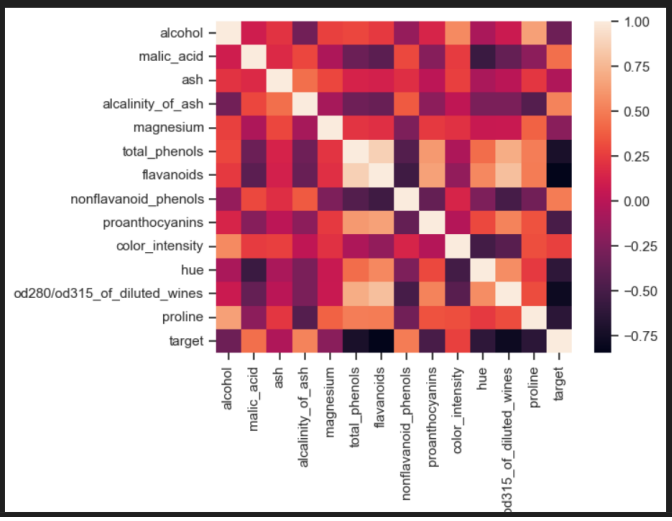
	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od280/od315_of_diluted_wines	proline	target
)	0.094397	0.211545	-0.310235	0.270798	0.289101	0.236815	-0.155929	0.136698	0.546364	-0.071747	0.072343	0.643720	-0.328222
7	1.000000	0.164045	0.288500	-0.054575	-0.335167	-0.411007	0.292977	-0.220746	0.248985	-0.561296	-0.368710	-0.192011	0.437776
5	0.164045	1.000000	0.443367	0.286587	0.128980	0.115077	0.186230	0.009652	0.258887	-0.074667	0.003911	0.223626	-0.049643
5	0.288500	0.443367	1.000000	-0.083333	-0.321113	-0.351370	0.361922	-0.197327	0.018732	-0.273955	-0.276769	-0.440597	0.517859
3	-0.054575	0.286587	-0.083333	1.000000	0.214401	0.195784	-0.256294	0.236441	0.199950	0.055398	0.066004	0.393351	-0.209179
l	-0.335167	0.128980	-0.321113	0.214401	1.000000	0.864564	-0.449935	0.612413	-0.055136	0.433681	0.699949	0.498115	-0.719163
5	-0.411007	0.115077	-0.351370	0.195784	0.864564	1.000000	-0.537900	0.652692	-0.172379	0.543479	0.787194	0.494193	-0.847498
)	0.292977	0.186230	0.361922	-0.256294	-0.449935	-0.537900	1.000000	-0.365845	0.139057	-0.262640	-0.503270	-0.311385	0.489109
3	-0.220746	0.009652	-0.197327	0.236441	0.612413	0.652692	-0.365845	1.000000	-0.025250	0.295544	0.519067	0.330417	-0.499130
4	0.248985	0.258887	0.018732	0.199950	-0.055136	-0.172379	0.139057	-0.025250	1.000000	-0.521813	-0.428815	0.316100	0.265668
7	-0.561296	-0.074667	-0.273955	0.055398	0.433681	0.543479	-0.262640	0.295544	-0.521813	1.000000	0.565468	0.236183	-0.617369
3	-0.368710	0.003911	-0.276769	0.066004	0.699949	0.787194	-0.503270	0.519067	-0.428815	0.565468	1.000000	0.312761	-0.788230
)	-0.192011	0.223626	-0.440597	0.393351	0.498115	0.494193	-0.311385	0.330417	0.316100	0.236183	0.312761	1.000000	-0.633717
2	0.437776	-0.049643	0.517859	-0.209179	-0.719163	-0.847498	0.489109	-0.499130	0.265668	-0.617369	-0.788230	-0.633717	1.000000

В случае большого количества признаков анализ числовой корреляционной матрицы становится неудобен.

```
sns.heatmap(data.corr())
```

0.2s

```
<Axes: >
```



Из корреляционной матрицы можно сделать следующие выводы:

1. между flavanoids и total_phenols сильная положительная корреляция, поэтому будем использовать только один
2. целевой признак слабо коррелирует с color_intensity, nonflavanoid_phenols, alcohol, malic_acid, ash, magnesium и hue, поэтому эти признаки стоит исключить из модели
3. из flavanoids и total_phenols сильнее коррелирует с ключевым признаком flavanoids, поэтому лучше оставить его

Таким образом в модели останутся признаки **alcalinity_of_ash, flavanoids, od280/od315_of_diluted_wines и proline**