Московский Государственный технический университет
имени Н. Э. Баумана

Лабораторная работа №1 по курсу:
«Технология машинного обучения»

Работу выполнил студент группы ИУ5-63
Федорова Антонина_____

Работу проверил:
Гапанюк Ю.Е._____

Москва
2019

Задание:

- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов 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.

Текст программы с примерами выполнения программы:

В качестве набора данных я выбрала датасет, демонстрирующий качество вина по содержанию в нем разных веществ (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009/version/2).
Эта задача будет полезна в различных ресторанах, алкогольных магазинах, а также на различных винодельнях.
Датасет состоит пока что из одного файла, но для обучения модели я разобью его на 3 части.
В моем датасете хранятся следующие данные:
1)fixed acidit - фиксированная кислотность
2)volatile acidity - летучая кислотность
3)citric acid - лимонная кислота
4)residual sugar - остаточный сахар
5)chlorides - хлориды
6)free sulfur dioxide - свободный диоксид серы
7)total sulfur dioxide - общий диоксид серы
8)density - плотность
9)pH - водородный показатель
10)sulphates - сульфаты
11)alcohol - алкоголь
12)quality - качество - целевой признак - оценка вина от 0 до 10 баллов
Более подробное описание этого датасета по ссылке.

```
#Импорт библиотек
import numpy as np
import pandas as pd
import seaborn as sns
from google.colab import drive
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
#Монтирую гугл диск, чтобы взять оттуда датасет
drive.mount("/content/gdrive", force_remount=True)
#Загружаю данные с гугл диска
data = pd.read_csv('/content/gdrive/My Drive/winequality-red.csv', sep=«,")
data.head()
```

data.shape

```
(1599, 12)
```

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

```
Всего строк: 1599
fixed acidity - 0
volatile acidity - 0
citric acid - 0
residual sugar - 0
chlorides - 0
free sulfur dioxide - 0
total sulfur dioxide - 0
density - 0
pH - 0
sulphates - 0
alcohol - 0
quality - 0
```

#Различные метрики по моим данным

```
data.describe()
```

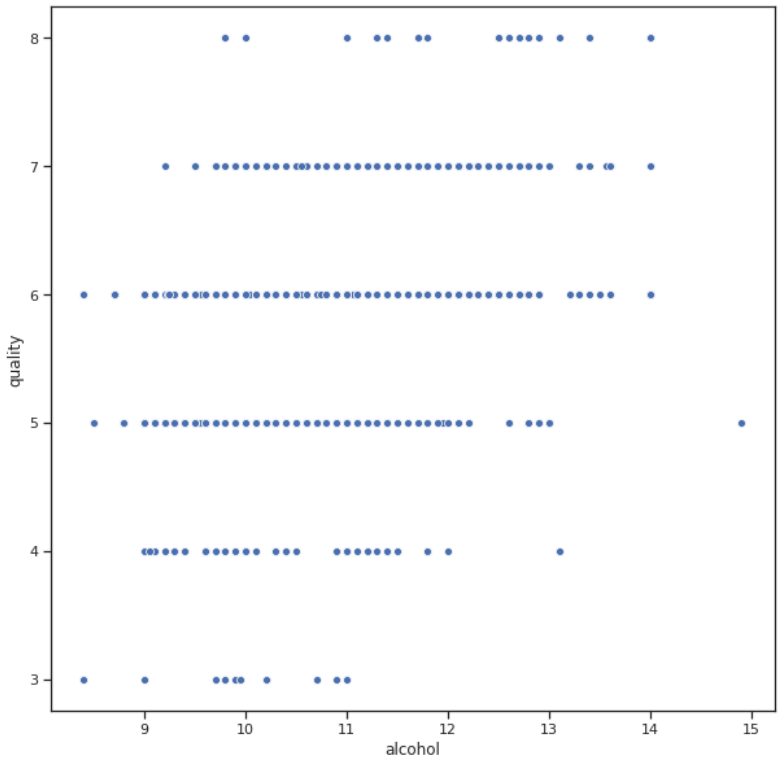
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467
std	1.741096	0.179060	0.194801	1.409928	0.047065
min	4.600000	0.120000	0.000000	0.900000	0.012000
25%	7.100000	0.390000	0.090000	1.900000	0.070000
50%	7.900000	0.520000	0.260000	2.200000	0.079000
75%	9.200000	0.640000	0.420000	2.600000	0.090000
max	15.900000	1.580000	1.000000	15.500000	0.611000

#Типы данных значений датасета
data.dtypes

```
fixed acidity           float64
volatile acidity        float64
citric acid             float64
residual sugar          float64
chlorides               float64
free sulfur dioxide     float64
total sulfur dioxide    float64
density                 float64
pH                      float64
sulphates               float64
alcohol                 float64
quality                   int64
dtype: object
```

#График рассеивания качества от алкоголя

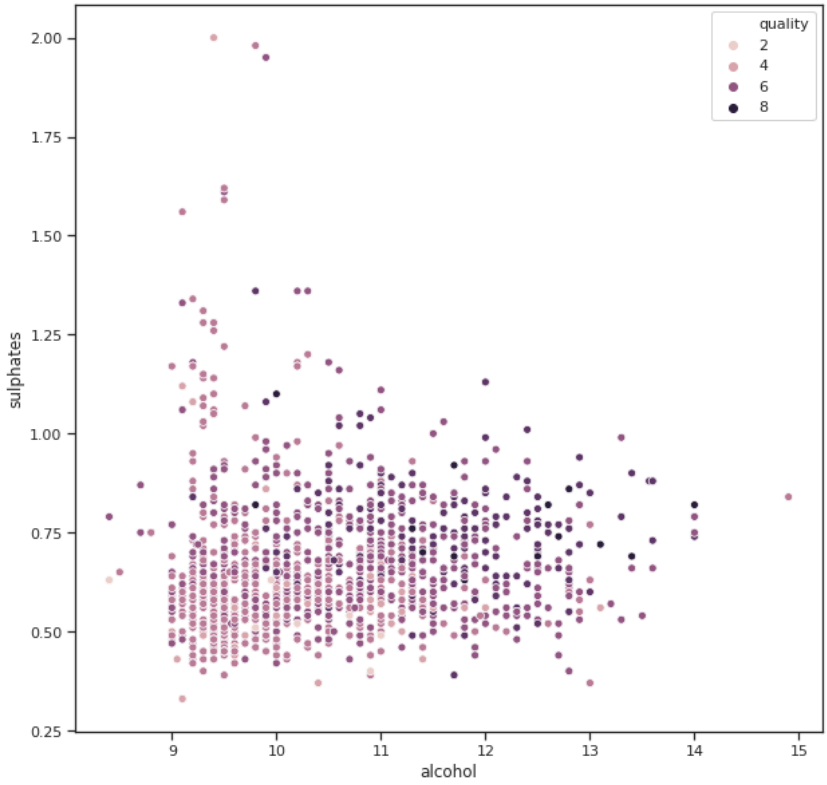
```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='alcohol', y='quality',
data=data)
```



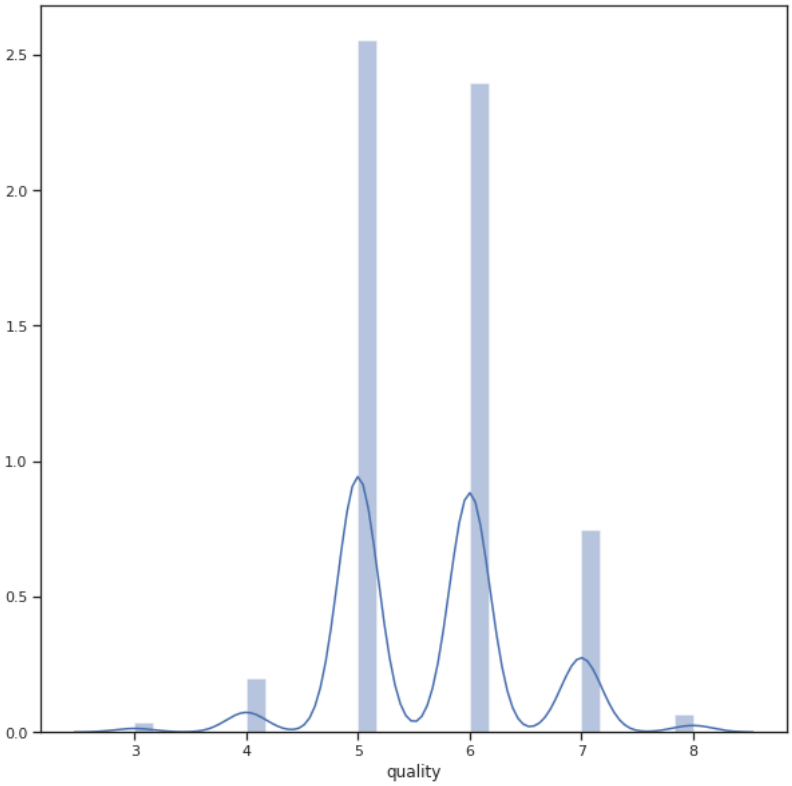
#Сульфаты и алкоголь+ окраска по качеству

```
fig, ax = plt.subplots(figsize=(10,10))
```

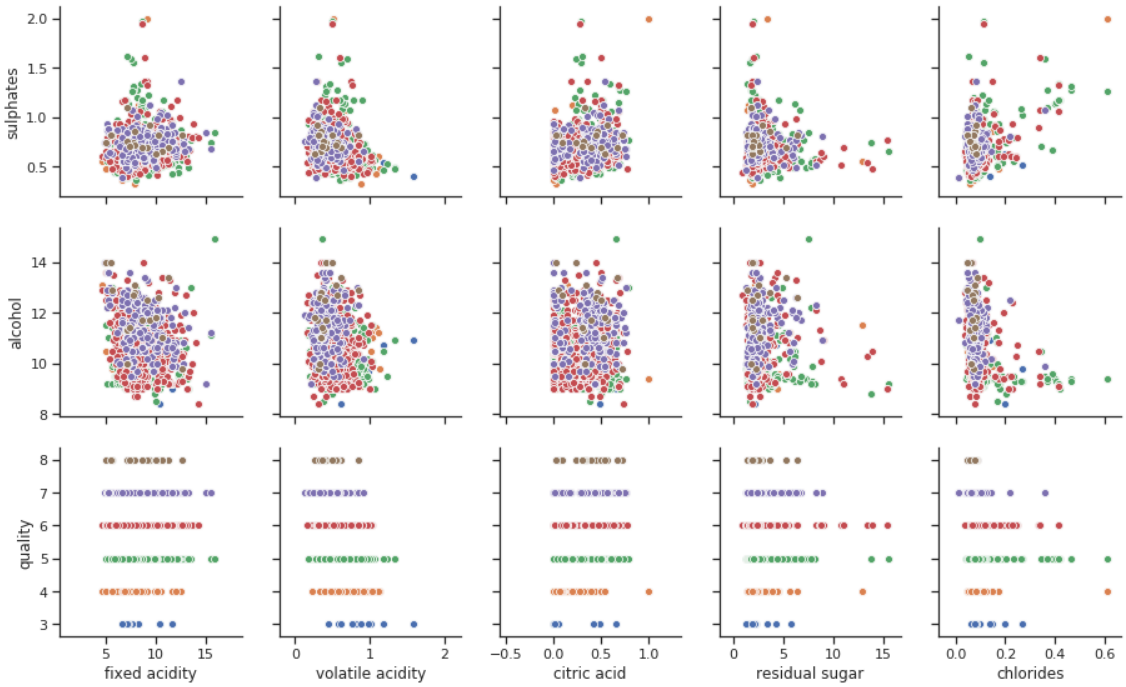
```
sns.scatterplot(ax=ax, x='alcohol', y='sulphates',
data=data, hue='quality')
```




```
#Гистограмма распределения целевого признака
fig, ax = plt.subplots(figsize=(10,10))
sns.distplot(data['quality'])
```



```
sns.pairplot(data, hue=«quality»)
```




```
fig, ax = plt.subplots(1, 3, sharex='col', sharey='row', figsize=(30,10))
sns.heatmap(data.corr(method='pearson'), ax=ax[0], annot=True, fmt='.2f')
sns.heatmap(data.corr(method='kendall'), ax=ax[1], annot=True, fmt='.2f')
sns.heatmap(data.corr(method='spearman'), ax=ax[2], annot=True, fmt='.
2f')
fig.suptitle('Корреляционные матрицы, построенные различными методами')
ax[0].title.set_text('Pearson')
ax[1].title.set_text('Kendall')
ax[2].title.set_text(‘Spearman')
```