



Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования «Московский государственный
технический университет имени Н.Э. Баумана»
(МГТУ им. Н.Э. Баумана)

Рубежный контроль №1
по курсу «Методы машинного обучения»

Выполнил
студент группы ИУ5-25М
Козинов Олег

Москва, 2023

ИУ5-25М Вариант 7

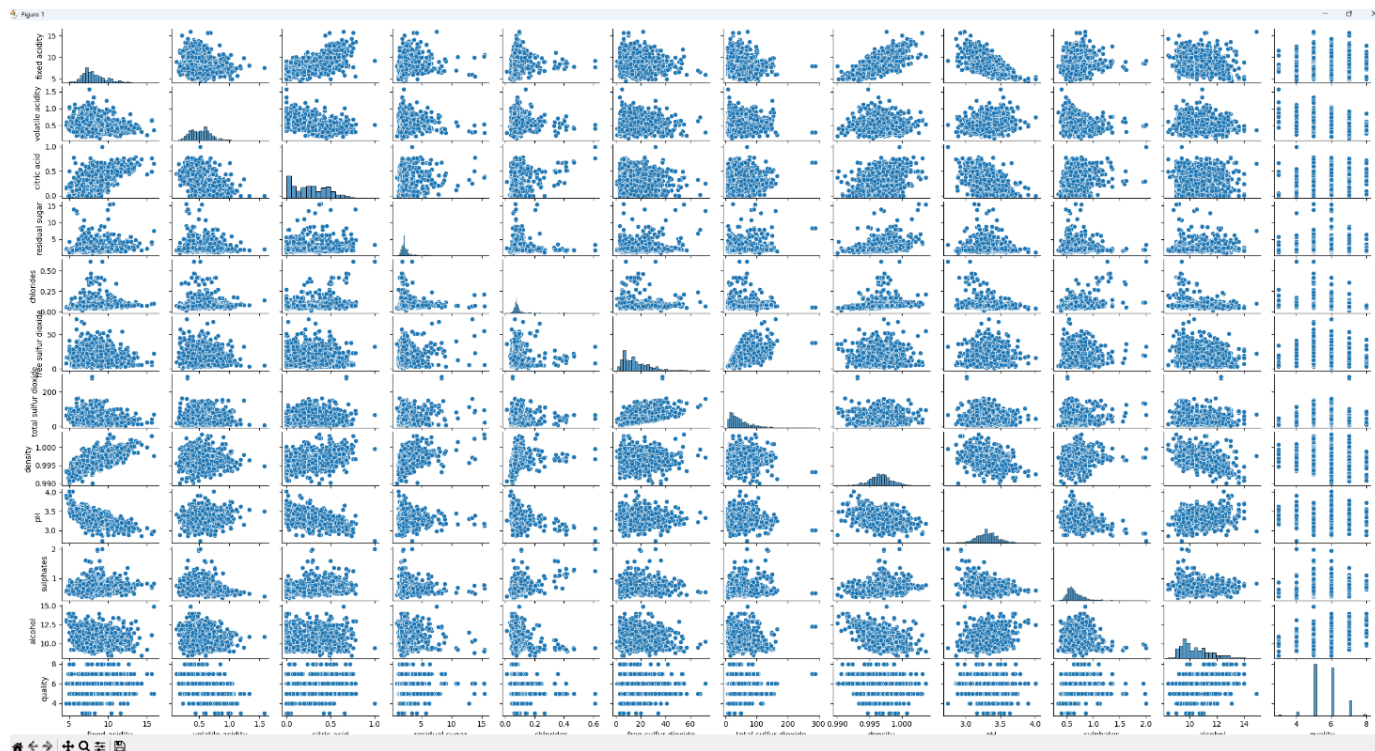
Доп. задание: для студентов групп ИУ5-25М- для произвольной колонки данных построить гистограмму.

Задача №7

Для набора данных проведите устранение пропусков для одного (произвольного) числового признака с использованием метода заполнения медианой.

```
main.py  rk1.py ×
1 import pandas as pd
2 import seaborn as sns
3 import matplotlib.pyplot as plt
4
5 # загрузить набор данных в объект pandas DataFrame
6 data = pd.read_csv('C:/Users/Олег/PycharmProjects/RPSgame/winequality-red.csv')
7
8 # до заполнения
9 print(data.isnull().sum())
10
11 # определить, какой столбец содержит числовой признак, для которого требуется заполнить пропуски
12 numeric_feature = 'fixed acidity'
13
14 # рассчитать медиану числового признака
15 median = data[numeric_feature].median()
16
17 # заменить все пропущенные значения в столбце на вычисленную медиану
18 data[numeric_feature].fillna(median, inplace=True)
19
20 # проверить, что все пропуски были заполнены
21 print(data.isnull().sum())
22
23 # сохранить изменения в исходном наборе данных
24 data.to_csv('C:/Users/Олег/PycharmProjects/RPSgame/winequality-red.csv', index=False)
25
26 # загрузить данные
27 df = pd.read_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv', delimiter=';')
28
29 # построить парные диаграммы
30 sns.pairplot(df, vars=['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides',
31                       'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol', 'quality'])
32
33 # показать диаграммы
34 plt.show()
```

Гистограмма:



Задача №27

```
main.py  rk1.py ×
1 import pandas as pd
2 # import seaborn as sns
3 # import matplotlib.pyplot as plt
4
5 # загрузить данные
6 df = pd.read_csv('C:/Users/Олег/PycharmProjects/RPSgame/winequality-red.csv')
7
8 # выбрать колонку для обработки
9 col = 'fixed acidity'
10
11 # определить верхнюю и нижнюю границы выбросов
12 q1 = df[col].quantile(0.05)
13 q2 = df[col].quantile(0.95)
14
15 # заменить выбросы на верхнюю и нижнюю границы
16 df.loc[df[col] < q1, col] = q1
17 df.loc[df[col] > q2, col] = q2
18
19 print(df[col].describe())
```

```
count    1599.000000
mean       8.294309
std        1.563319
min        6.100000
25%        7.100000
50%        7.900000
75%        9.200000
max       11.800000
Name: fixed acidity, dtype: float64
```