

МГТУ имени Баумана  
Факультет «Информатика и Системы управления»  
Кафедра «Системы обработки информации и управления»  
Дисциплина «Теория машинного обучения»

Отчет по рубежному контролю №1

Выполнил:  
Студент группы ИУ5-61Б  
Гапчук Л.Д.

Преподаватель:  
Гапанюк Ю.Е.

Москва, 2020г.

Цель лабораторной работы: проверка изученных тем.

Задание:

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.

```
In [1]: import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from pandas.plotting import scatter_matrix
import warnings
warnings.filterwarnings('ignore')
sns.set(style="ticks")
%matplotlib inline
```

```
In [2]: data = pd.read_csv('heart.csv')
```

```
In [3]: data.head()
```

```
Out[3]:
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

```
In [4]: data.dtypes
```

```
Out[4]: age          int64
sex            int64
cp             int64
trestbps       int64
chol           int64
fbs            int64
restecg        int64
thalach        int64
exang          int64
oldpeak       float64
slope          int64
ca             int64
thal           int64
target         int64
dtype: object
```

```
In [5]: data.isnull().sum()
# проверим есть ли пропущенные значения
```

```
Out[5]: age          0
sex            0
cp             0
trestbps       0
chol           0
fbs            0
restecg        0
thalach        0
exang          0
oldpeak       0
slope          0
ca             0
thal           0
target         0
dtype: int64
```

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

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
#   Column      Non-Null Count  Dtype
---  ---
0    age         303 non-null    int64
1    sex         303 non-null    int64
2    cp          303 non-null    int64
3    trestbps    303 non-null    int64
4    chol        303 non-null    int64
5    fbs         303 non-null    int64
6    restecg     303 non-null    int64
7    thalach     303 non-null    int64
8    exang       303 non-null    int64
9    oldpeak     303 non-null    float64
10   slope       303 non-null    int64
11   ca          303 non-null    int64
12   thal        303 non-null    int64
13   target      303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

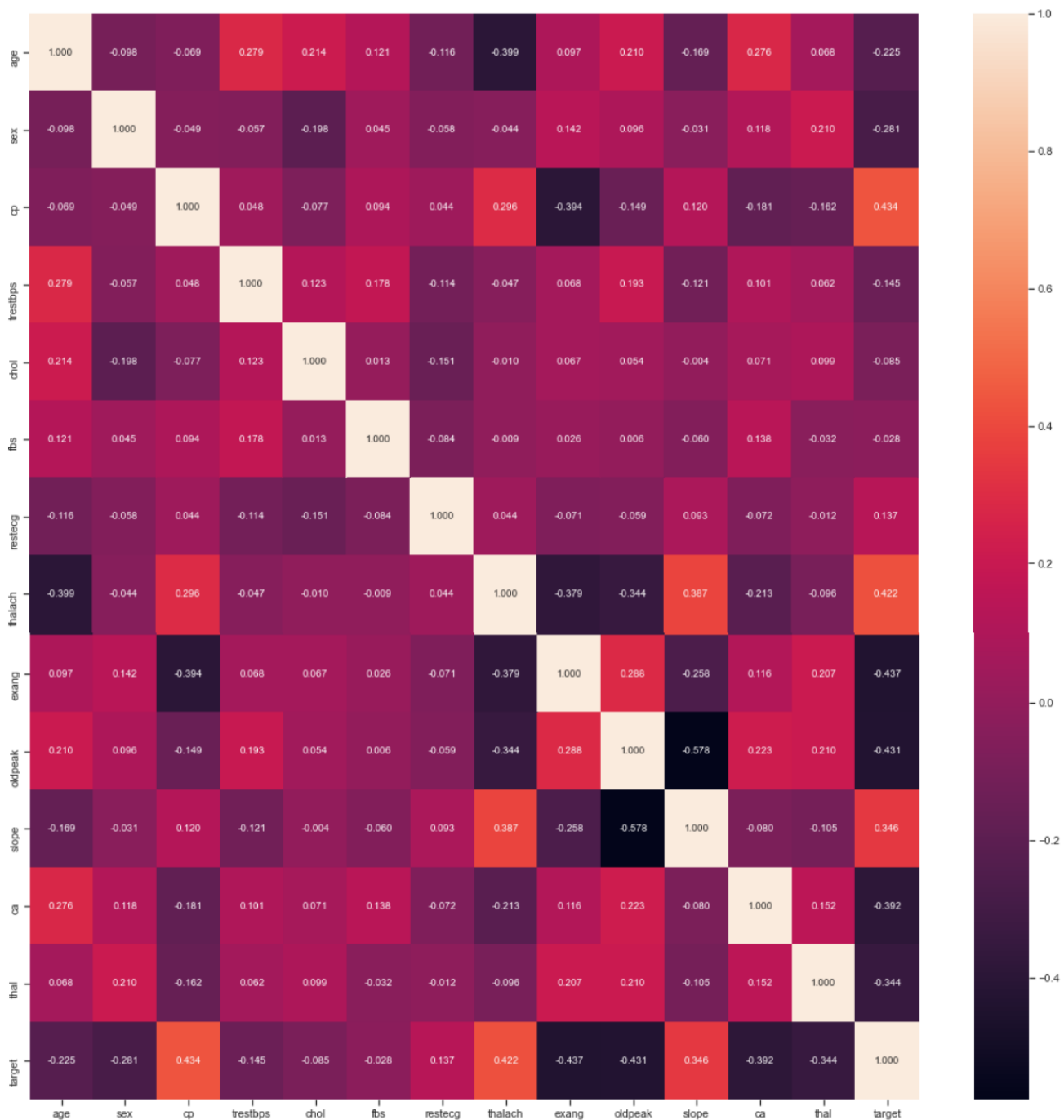
```
In [7]: ## Корр. анализ
corr_matrix = data.corr()
```

```
In [8]: corr_matrix['age']
```

```
Out[8]: age         1.000000
sex        -0.098447
cp         -0.068653
trestbps   0.279351
chol       0.213678
fbs        0.121308
restecg    -0.116211
thalach    -0.398522
exang      0.096801
oldpeak    0.210013
slope     -0.168814
ca         0.276326
thal       0.068001
target    -0.225439
Name: age, dtype: float64
```

```
In [9]: plt.figure(figsize=(20,20))
sns.heatmap(corr_matrix, annot=True, fmt='.3f')
```

```
Out[9]: <AxesSubplot:>
```



```
In [10]: # Увеличенные диаграммы рассеяния
sns.jointplot(x = "age", y = "trestbps", kind="scatter", data = data)
```

```
Out[10]: <seaborn.axisgrid.JointGrid at 0x2ba87b07910>
```

