



**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 дисциплине «Технология Машинного обучения»**

**Выполнил:
студент группы ИУ5-62
Миронов Святослав
подпись, дата**

2019 г.

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 своей репозитории на github.

Lab1_end

June 3, 2019

```
In [1]: !pip install seaborn
```

DEPRECATION: Python 2.7 will reach the end of its life on January 1st, 2020. Please upgrade your Python as Python 2.7 is deprecated and will not be supported by pip. See <https://pip.pypa.io/en/latest/learning-more/about-deprecated-features/#python-deprecation-warning> for more details.

Requirement already satisfied: seaborn in /usr/local/lib/python2.7/dist-packages (0.9.0)
Requirement already satisfied: matplotlib>=1.4.3 in /usr/lib/python2.7/dist-packages (from seaborn==0.9.0)
Requirement already satisfied: numpy>=1.9.3 in /usr/local/lib/python2.7/dist-packages (from seaborn==0.9.0)
Requirement already satisfied: pandas>=0.15.2 in /usr/local/lib/python2.7/dist-packages (from seaborn==0.9.0)
Requirement already satisfied: scipy>=0.14.0 in /usr/local/lib/python2.7/dist-packages (from seaborn==0.9.0)
Requirement already satisfied: pytz>=2011k in /usr/lib/python2.7/dist-packages (from pandas>=0.15.2->seaborn==0.9.0)
Requirement already satisfied: python-dateutil>=2.5.0 in /usr/lib/python2.7/dist-packages (from pandas>=0.15.2->seaborn==0.9.0)

```
In [2]: import numpy as np
import pandas as pd
from sklearn import datasets
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline
```

```
In [3]: df = pd.read_csv('heart.csv', sep=",")
```

1 Data Set Information:

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.

1.1 Attribute Information:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)

8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

In [4]: df.shape

Out[4]: (303, 14)

In [5]: df.head()

```
Out[5]:
```

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

	ca	thal	target
0	0	1	1
1	0	2	1
2	0	2	1
3	0	2	1
4	0	2	1

In [6]: df.dtypes

```
Out[6]:
```

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 [7]: df.isnull().sum()

```
Out[7]: 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 [8]: df['target'].unique()
```

```
Out[8]: array([1, 0])
```

```
In [9]: df.describe()
```

```
Out[9]:
```

	age	sex	cp	trestbps	chol	fbs	\
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	

	restecg	thalach	exang	oldpeak	slope	ca	\
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	
mean	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	
std	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	
min	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	
50%	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	
75%	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	
max	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	

	thal	target
count	303.000000	303.000000

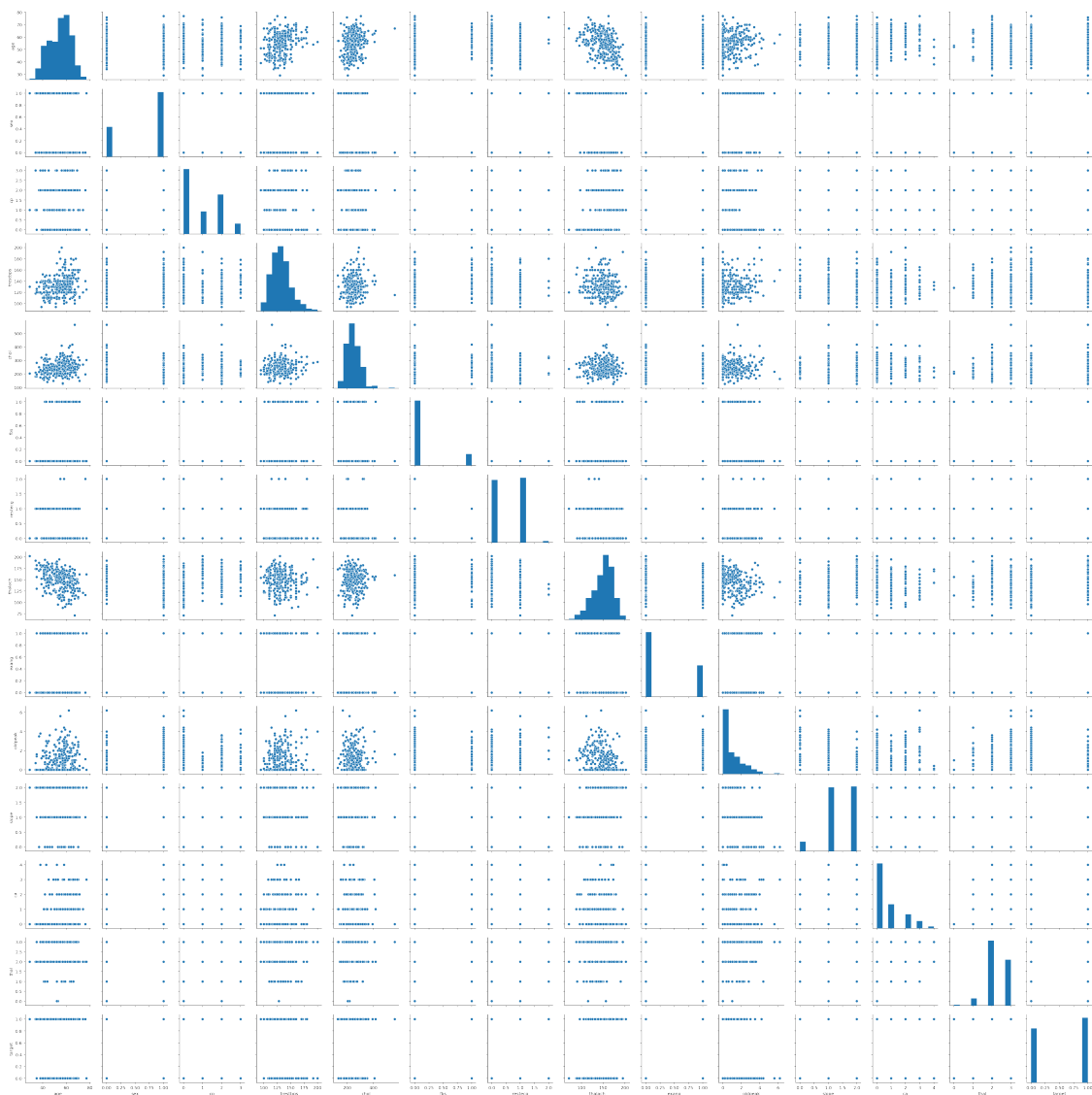
mean	2.313531	0.544554
std	0.612277	0.498835
min	0.000000	0.000000
25%	2.000000	0.000000
50%	2.000000	1.000000
75%	3.000000	1.000000
max	3.000000	1.000000

~,

In [10]: sns.pairplot(data= df)

Out[10]: <seaborn.axisgrid.PairGrid at 0x7f2287a89940>

Out[10]:



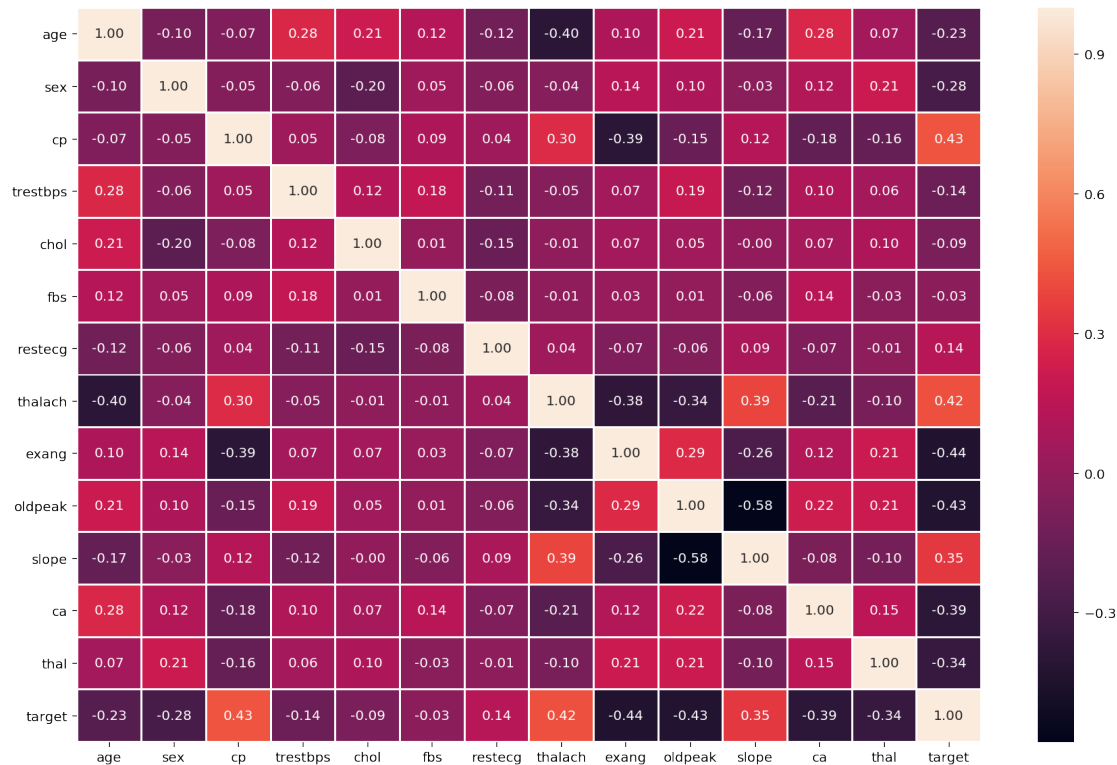
```
In [11]: #sns.pairplot(df, hue= "target")
```

```
:
```

```
In [12]: plt.figure(figsize=(15, 10))
sns.heatmap(df.corr(),annot=True, fmt='.2f', linewidths=1)
```

```
Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x7f228082e0f0>
```

```
Out[12]:
```



```
In [13]: plt.figure(figsize=(15, 10))
sns.heatmap(df.corr(method='spearman'),annot=True, fmt='.3f')
```

```
Out[13]: <matplotlib.axes._subplots.AxesSubplot at 0x7f227ed99390>
```

```
Out[13]:
```

