Рубежный контроль №1

Бессонова Ксения ИУ5-61Б

Задание.

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. Для пары произвольных колонок данных построить график "Диаграмма рассеяния"

In [1]:
```
#Импорт библиотек
import pandas as pd
from sklearn.preprocessing import LabelEncoder
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

В качестве датасета возьмём данные о болезнях сердца

In [2]:
```
#Загрузка датасета
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	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

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

Out[4]:
```
(1025, 14)
```

Целевым признаком будет наличие сердечных заболеваний у пациента (поле "target")

Проверка типов данных и наличие пропусков

In [5]:
```
#Проверка типов
data.dtypes
```

Out[5]:
```
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
```

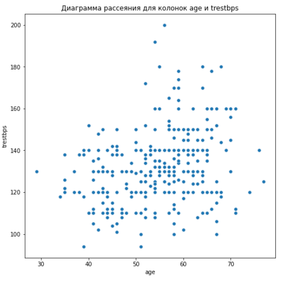
In [6]:
```
data.isnull().sum()
```

Out[6]:
```
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
```

Диаграмма рассеяния

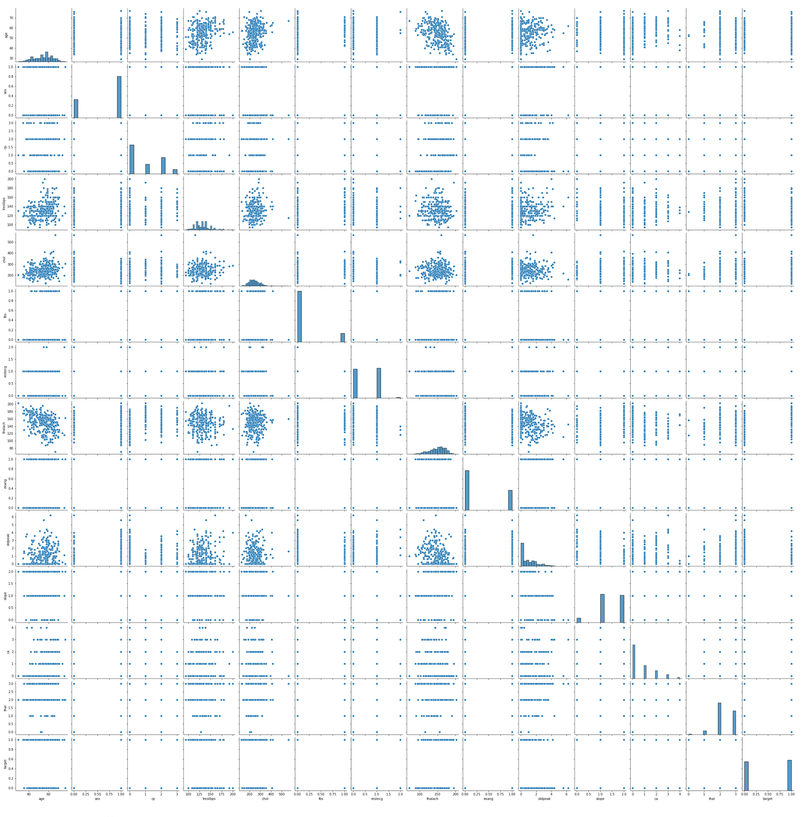
In [7]:
```
fig, ax = plt.subplots(figsize=(8,8))
ax.set_title("Диаграмма рассеяния для колонок age и trestbps")
sns.scatterplot(ax=ax, x='age', y='trestbps', data=data)
```

Out[7]:
```
<AxesSubplot:title={'center':'Диаграмма рассеяния для колонок age и trestbps'}, xlabel='age', ylabel='trestbps'>
```



In [8]:
```
sns.pairplot(data)
```

Out[8]:
```
<seaborn.axisgrid.PairGrid at 0x2bfabb00e50>
```

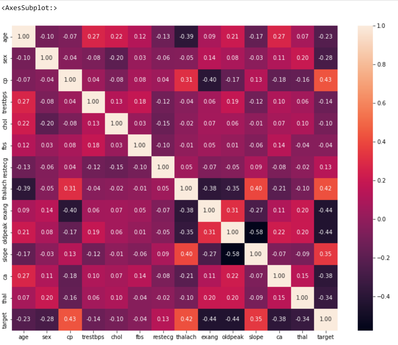


Корреляционный анализ

Создадим корреляционную матрицу используя коэффициент Пирсона

In [9]:
```
fig, ax = plt.subplots(1,1, figsize=(13,10))
sns.heatmap(data.corr("pearson"), annot=True, fmt=".2f")
```

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



На основе корреляционной матрицы можно сделать следующие выводы.

Все признаки слабо коррелируют с целевым признаком target. Наиболее сильно коррелируют с целевым признаком поля cp, thalach и slope.

Следовательно для обучения модели лучше выбрать поля cp, thalach и slope.

In [ ]:
```
```