МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. Н.Э. Баумана

Факультет «Информатика и системы управления»
Кафедра «Систем обработки информации и управления»

ОТЧЕТ

Рубежный контроль №__1__
по дисциплине «Методики машинного обучения»

Вариант 5

ИСПОЛНИТЕЛЬ: Егоров С.А.
ФИО
группа ИУ5-22М
_____
подпись
"__"_____2020 г.

ПРЕПОДАВАТЕЛЬ: Гапанюк Ю.Е.
ФИО
_____
подпись
"__"_____2020 г.

Москва - 2020

Задание

1. Для заданного набора данных постройте основные графики, входящие в этап разведочного анализа данных с использованием библиотек Matplotlib и Seaborn.
2. В случае наличия пропусков в данных удалите строки или колонки, содержащие пропуски.
3. Какие графики Вы построили и почему?
4. Какие выводы о наборе данных Вы можете сделать на основании построенных графиков?
5. Проведите корреляционный анализ.
6. Сделайте выводы о возможности построения моделей машинного обучения и о возможном вкладе признаков в модель.

Реализация задания

Подключив набор данных, проверим его на наличие пустых значений:

```
# Проверим наличие пустых значений
for col in data.columns:
    temp_null_count = data[data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
```

```
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
```

Посмотрим на типы данных данного набора данных:

```
# Список колонок с типами данных
data.dtypes
```

```
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
```

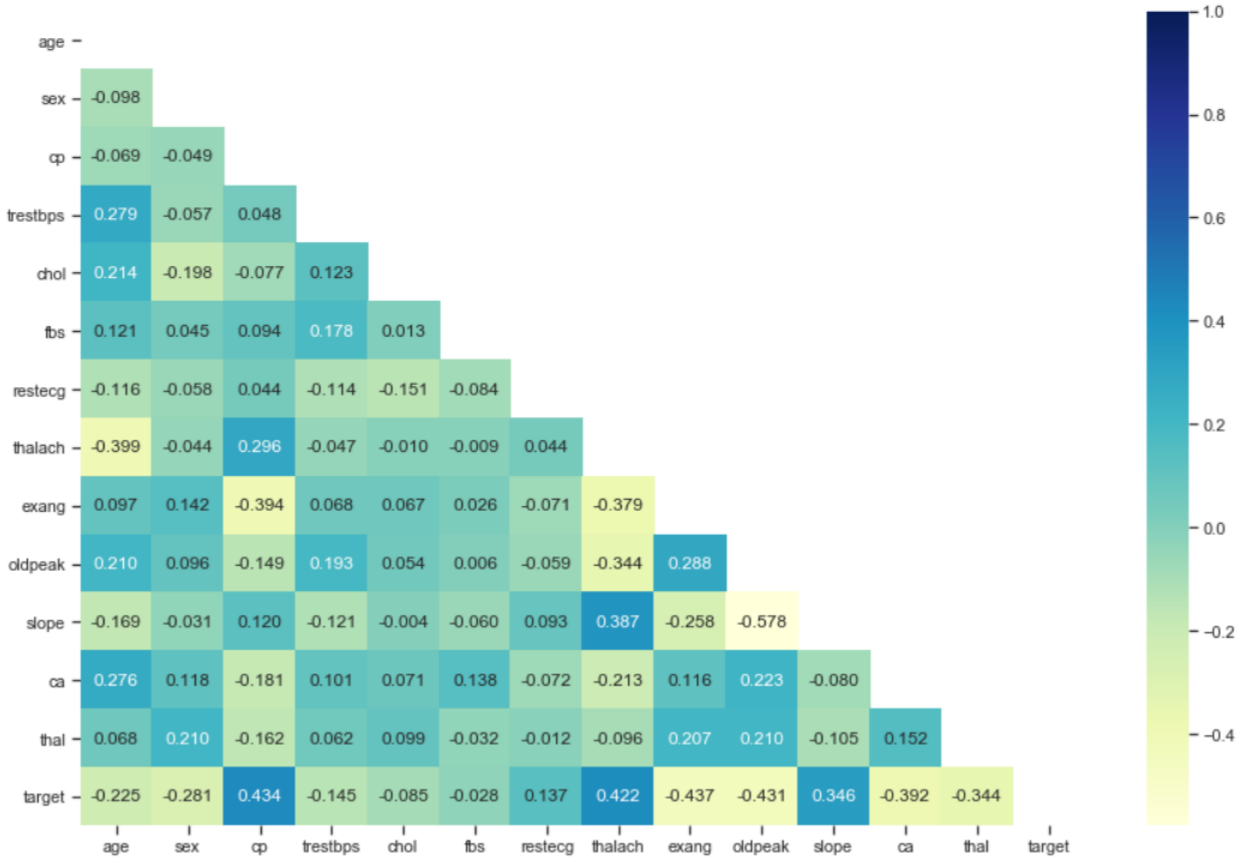
Так как пустых значений не обнаружено и все данных имеют числовой формат, то можно приступать к разведочному анализу.

Выведет основные статистические характеристик данных:

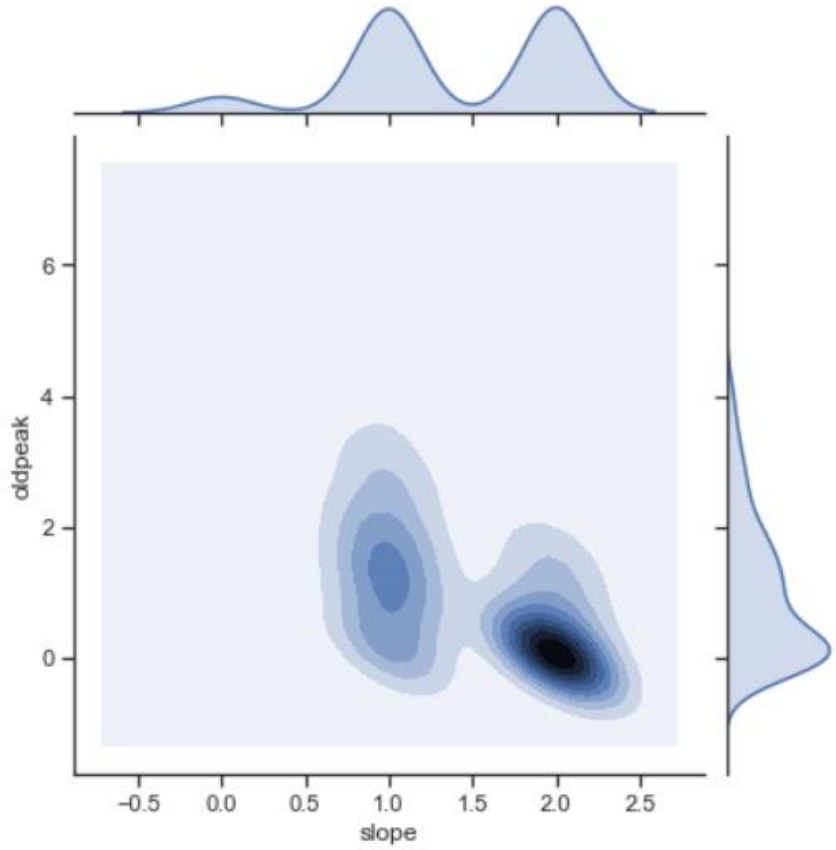
	age	sex	cp	trestbps	chol	fbs	restecg
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000

	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Чтобы попарно сравнивать показатели проведем корреляционный анализ и построим треугольную матрицу корреляции для набора данных:

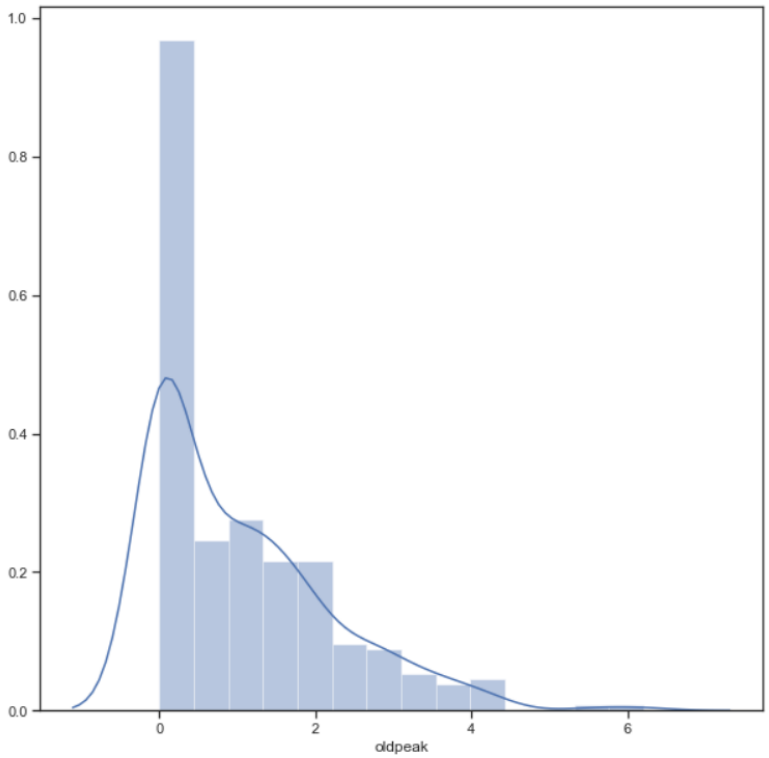


Как видно лучшую статистическую взаимосвязь имеют параметры “slope” и “oldpeak” (0,578) для них и построим диаграмму рассеивания с гистограммой:



Рассматривать другие графики не имеет большого смысла, так как коэффициент корреляции у них меньше 0,5, то есть данные имеет слабую взаимосвязь.

Построим гистограмму для параметра “oldpeak”.



Выводы

Исходя из корреляционно анализа сложно выделить какой-то целевой признак, так как взаимосвязи между признаками малы, а если связь имеет от 0,4 до 0,6 значение, то такие связи присутствуют у разных признаков и выделить какой-то признак затруднительно.