Московский авиационный институт (национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №0 по курсу «Искусственный интеллект»
Тема: Анализ и подготовка данных

Студент: А. В. Тимофеев
Преподаватель: Самир Ахмед
Группа: М8О-407Б-19

Дата:
Оценка:
Подпись:

Москва, 2022

Задача

Задача: В данной лабораторной работе, вы выступаете в роли предприимчивого начинающего стартапера в области машинного обучения. Вы заинтересовались этим направлением и хотите предложить миру что-то новое и при этом неплохо заработать. От вас требуется определить задачу которую вы хотите решить и найти под нее соответствующие данные. Так как вы не очень богаты, вам предстоит руками проанализировать данные, визуализировать зависимости, построить новые признаки и сказать хватит ли вам этих данных, и если не хватит найти еще. Вы готовитесь представить отчет ваши партнерам и спонсорам, от которых зависит дальнейшая ваша судьба. Поэтому тщательно работайте:) И главное, день промедления и вас опередит ваш конкурент, да и сплагиаченная работа отразится на репутации По сути в данной лабораторной работе вы выполняете часть работы BI системы. Если вы заинтересовались этим направлением, то можно будет в дальнейшем что-то придумать)

1 Описание

В качестве датасета я выбрал «Heart Attack Analysis Prediction Dataset» с сайта kaggle.
Он находится по ссылке https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset?select=heart.csv.
В данном датасете собраны некоторые признаки, влияющие на возникновение сердечного приступа.

В этом наборе данных приведены признаки:

1. Age : Возраст пациента
2. Sex: Пол пациента
3. exang: стенокардия, вызванная физической нагрузкой (1 = да; 0 = нет)
4. cp: тип боли в груди
    4.1 Value 1: типичная стенокардия
    4.2 Value 2: атипичная стенокардия
    4.3 Value 3: неангинальная боль
    4.4 Value 4: бессимптомное течение
5. trtbps: артериальное давление в состоянии покоя (в мм рт. ст.)
6. chol: холесторал в мг / дл, определяемый с помощью датчика ИМТ
7. fbs: (уровень сахара в крови натощак> 120 мг / дл) (1 = истина; 0 = ложь)
8. rest_ecg : результаты электрокардиографии в состоянии покоя
    8.1 Value 0: нормальное
    8.2 Value 1: аномалия зубца ST-T (инверсия зубца T и / или повышение или понижение ST> 0,05 мВ)
    8.3 Value 2: отображение вероятной или определенной гипертрофии левого желудочка по критериям Эстеса
9. thalach: достигнутая максимальная частота сердечных сокращений
10. target : 0 = меньше шансов сердечного приступа 1 = больше шансов сердечного приступа

2 Ход работы

Сначала проверим датасет на наличие в нем пустых ячеек, с помощью info(). Таковых там не оказалось.

```
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   trtbps     303 non-null     int64
4   chol       303 non-null     int64
5   fbs        303 non-null     int64
6   restecg    303 non-null     int64
7   thalachh   303 non-null     int64
8   exng       303 non-null     int64
9   oldpeak    303 non-null     float64
10  slp        303 non-null     int64
11  caa        303 non-null     int64
12  thall      303 non-null     int64
13  output     303 non-null     int64
dtypes: float64(1),int64(13)
memory usage: 33.3 KB
```

Так как все данные записаны в численном виде, мне не пришлось придумывать, как символьным данным сопоставить числа.
Дальше я проверил существуют ли повторяющиеся строки и если существуют, то дублирующаяся строка удаляется.

```
1 print(data.shape)
```

```
(303,14)
```

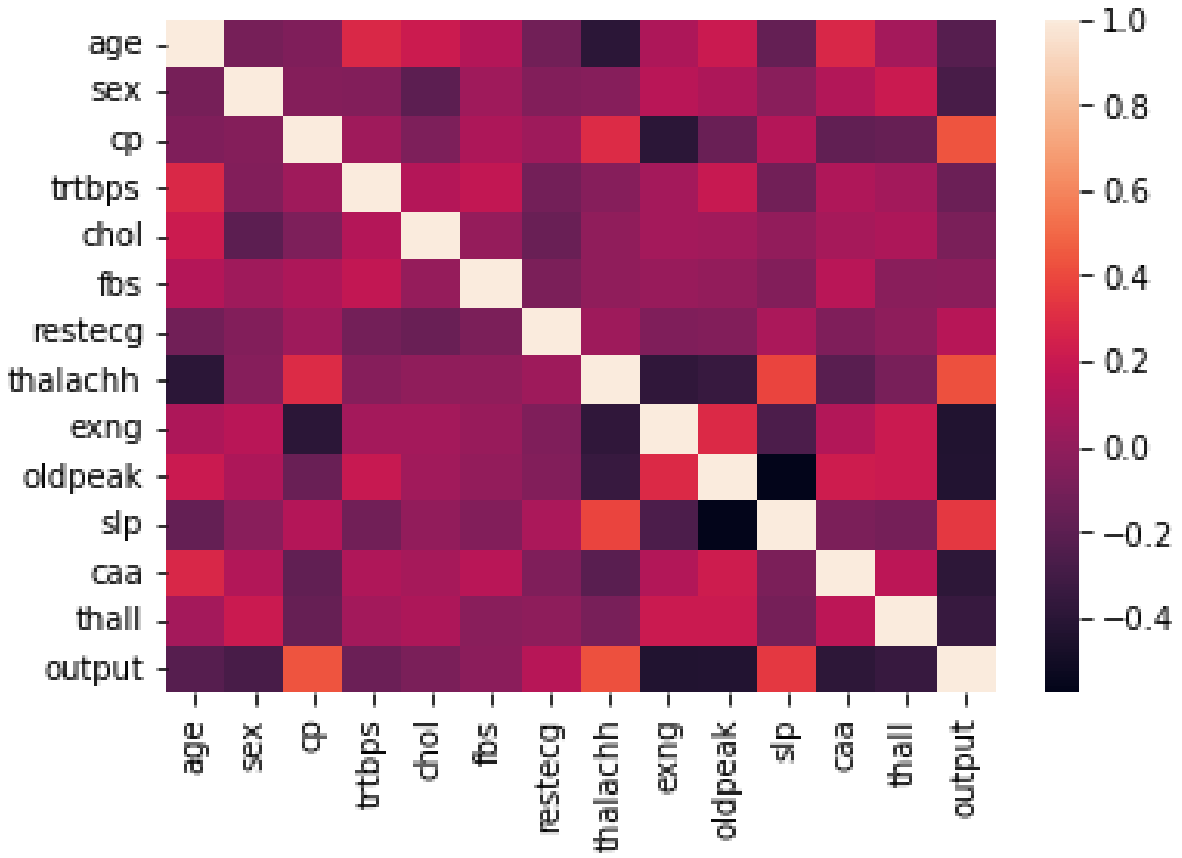
	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

302 rows × 14 columns

Далее я перешел непосредственно к анализу зависимостей в данных, построил корреляционную матрицу.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
age	1.000000	-0.098447	-0.068653	0.279351	0.213678	0.121308	-0.116211	-0.398522	0.096801	0.210013	-0.168814	0.276326	0.068001	-0.225439
sex	-0.098447	1.000000	-0.049353	-0.056769	-0.197912	0.045032	-0.058196	-0.044020	0.141664	0.096093	-0.030711	0.118261	0.210041	-0.280937
cp	-0.068653	-0.049353	1.000000	0.047608	-0.076904	0.094444	0.044421	0.295762	-0.394280	-0.149230	0.119717	-0.181053	-0.161736	0.433798
trtbps	0.279351	-0.056769	0.047608	1.000000	0.123174	0.177531	-0.114103	-0.046698	0.067616	0.193216	-0.121475	0.101389	0.062210	-0.144931
chol	0.213678	-0.197912	-0.076904	0.123174	1.000000	0.013294	-0.151040	-0.009940	0.067023	0.053952	-0.004038	0.070511	0.098803	-0.085239
fbs	0.121308	0.045032	0.094444	0.177531	0.013294	1.000000	-0.084189	-0.008567	0.025665	0.005747	-0.059894	0.137979	-0.032019	-0.028046
restecg	-0.116211	-0.058196	0.044421	-0.114103	-0.151040	-0.084189	1.000000	0.044123	-0.070733	-0.058770	0.093045	-0.072042	-0.011981	0.137230
thalachh	-0.398522	-0.044020	0.295762	-0.046698	-0.009940	-0.008567	0.044123	1.000000	-0.378812	-0.344187	0.386784	-0.213177	-0.096439	0.421741
exng	0.096801	0.141664	-0.394280	0.067616	0.067023	0.025665	-0.070733	-0.378812	1.000000	0.288223	-0.257748	0.115739	0.206754	-0.436757
oldpeak	0.210013	0.096093	-0.149230	0.193216	0.053952	0.005747	-0.058770	-0.344187	0.288223	1.000000	-0.577537	0.222682	0.210244	-0.430696
slp	-0.168814	-0.030711	0.119717	-0.121475	-0.004038	-0.059894	0.093045	0.386784	-0.257748	-0.577537	1.000000	-0.080155	-0.104764	0.345877
caa	0.276326	0.118261	-0.181053	0.101389	0.070511	0.137979	-0.072042	-0.213177	0.115739	0.222682	-0.080155	1.000000	0.151832	-0.391724
thall	0.068001	0.210041	-0.161736	0.062210	0.098803	-0.032019	-0.011981	-0.096439	0.206754	0.210244	-0.104764	0.151832	1.000000	-0.344029
output	-0.225439	-0.280937	0.433798	-0.144931	-0.085239	-0.028046	0.137230	0.421741	-0.436757	-0.430696	0.345877	-0.391724	-0.344029	1.000000

Далее я построил тепловую карту корреляционной матрицы.

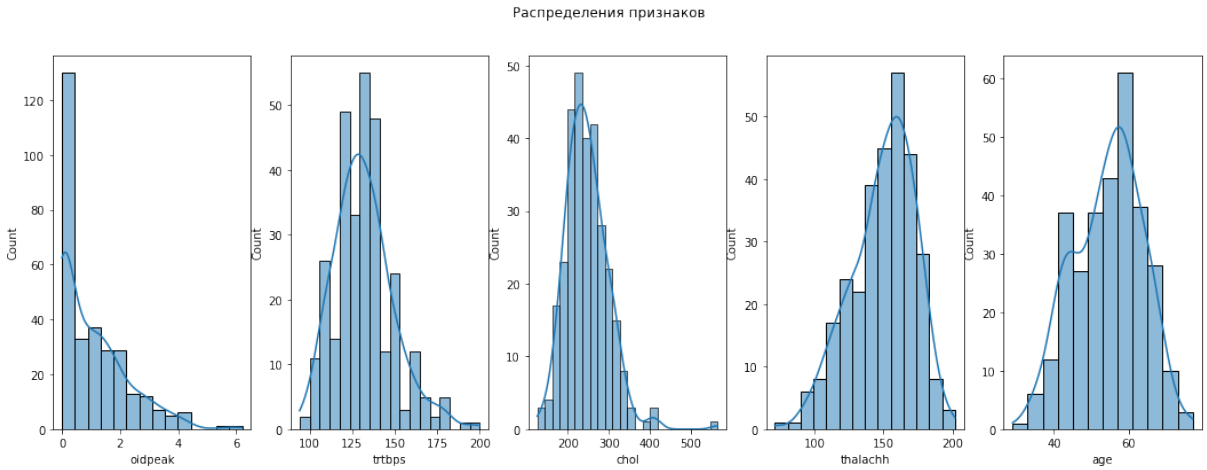


Так как целевым параметром является параметр output (0 = меньше вероятность сердечного приступа 1 = больше вероятность сердечного приступа) проанализируем последнюю колонку или строку. Из матрицы видно, что больше всего на этот параметр влияют: тип боли в груди (cp), достигнутая максимальная частота сердечных сокращений (thalachh), стенокардия, вызванная физической нагрузкой (exng), подавление сегмента ST, вызванное упражнением относительно отдыха. (oldpeak). А меньше всего - холесторал в мг / дл (chol) и уровень сахара в крови натощак (fbs). Также из матрицы видно, что остальные параметры сильно между собой некоррелируют, следовательно, можно не удалять/объединять их. Дальше я рассмотрел параметры моих признаков в датасет

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	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	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Отсюда видно, что данные ненормированы. Придется это иправить по мере реализации модели. Также можно подтвердить, что данные находятся в указанных диапазонах, как и говорится в описании к датасету

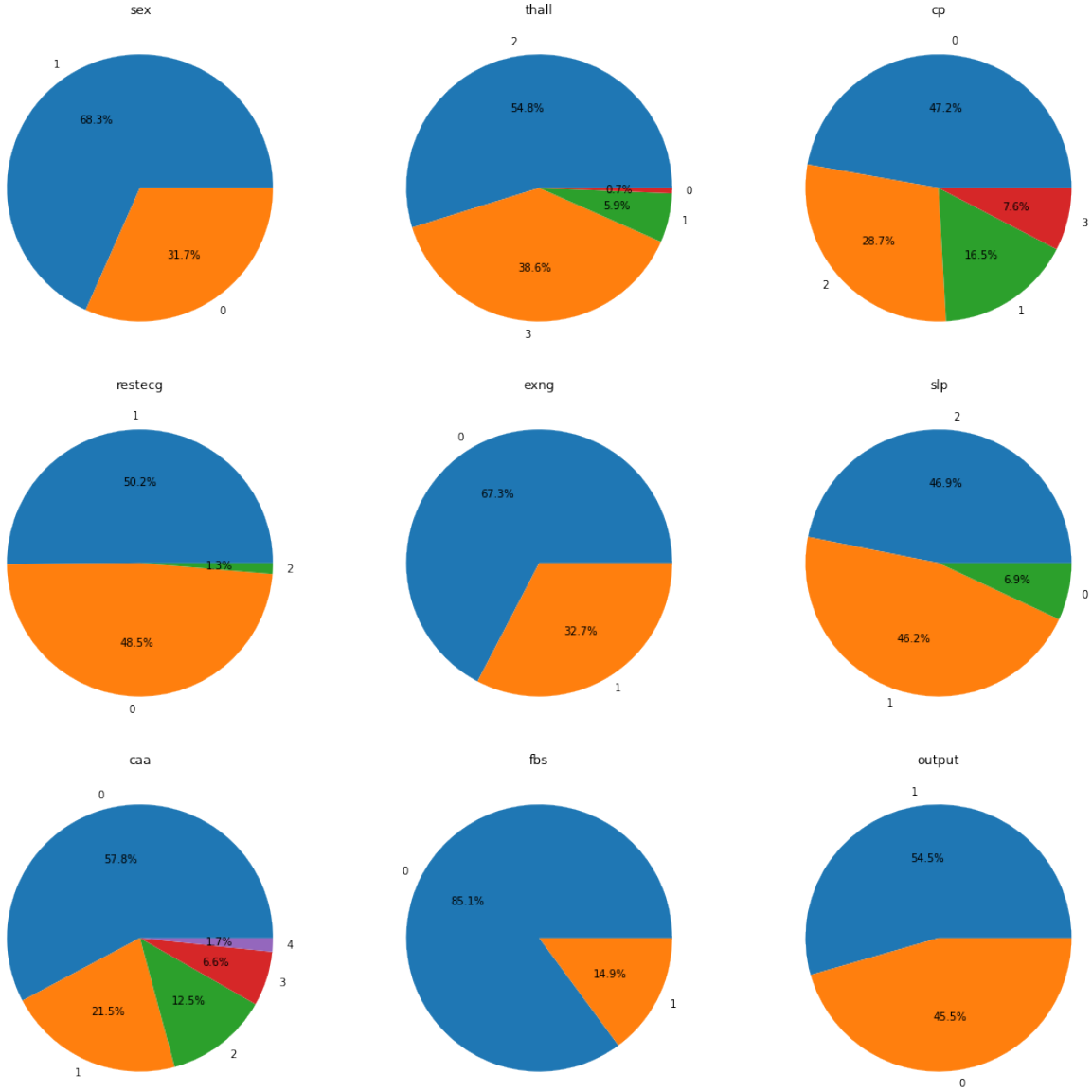
Далее построим гистаграммы распределения признаков. Для колличественных построим графики, где по оси X указывается значение параметра, а по Y - его количество



Из графиков видно, что значения распределены неравномерно, но эти распределения похожи на нормальные. Также на данных гистаграммах мной не были замечены выбросы, следовательно никаких изменений в датасет вносить не надо.

Также на графике oldpeak видно, что много людей с ST = 0, что является патологией. Скорее всего, датасет был основан на людях с больным сердцем, это косвено подтверждает возраст указанный в выборке. Все пациенты из датасета в примерно попадают в диапазон от 40 до 60 лет

Далее построим диаграммы для категориальных признаков.



Из данных диаграмм видно, что данные распределены неравномерно. Из диаграммы outout видно, что больных и здоровых людей примерно поровну, так что оверсемплинг делать не требуется.

Построим парные графики, при этом выделим цветом данные параметра HeartDisease.

Данные распределены достаточно хаотично. Добавлять новые признаки не требуется. Данные готовы к обучению.

3 Выводы

Данная лабораторная работа помогла мне лучше понять способы анализа и обработки данных из датасетов. После нейросетей непривычно, что надо преобразовывать дадасет, делать различные графики и рисунки.
Пожалуй главной проблемой было выбрать удачный датасет, но у меня получилось и в нем даже почти не пришлось ничего менять и добавлять. Также проблемы возникали с литературой, потому что она в основе своей написана на английском.