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: Лошманов Ю.А.
Преподаватель: Самир Ахмед
Группа: М8О-306Б-19
Дата:
Оценка:
Подпись:

Москва, 2022

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 Failure Prediction Dataset» (очень долго не мог найти подходящий, в начале находил только с 303 строками). Находится он здесь: `https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction`. Датасет направлен на выявление сердечной недостаточности по некоторым признакам, а именно:

1. Age: возраст пациента

2. Sex: пол пациента (M: Мужчина, F: Женщина)

3. ChestPainType: тип боли в груди

- TA: Типичная стенокардия

- ATA: Атипичная стенокардия

- NAP: Неангинальная боль

- ASY: Бессимптомная

4. RestingBP: кровяное давление в покое (мм рт. ст.)

5. Cholesterol: холестерин (мм/дл)

6. FastingBS: уровень сахара в крови натощак (1: если уровень сахара в крови натощак $>$ 120 мг/дл, 0: в противном случае)

7. RestingECG: результаты электрокардиограммы в покое

- Normal: Нормальный

- ST: Наличие аномальных зубцов S и T (отрицательный зубец T и/или отрицательное или положительное отклонение ST $>$ 0.05 мВ)

- LVH: наличие гипертрофии левого желудочка по критериям ЭхоКГ

8. MaxHR: достигнутая максимальная частота сердечных сокращений (числовое значение от 60 до 202)

9. ExerciseAngina: стенокардия, вызванная физической нагрузкой (Y: да, N: нет)

10. Oldpeak: числовое значение ST, измеренное в депрессии

11. ST_Slope: наклон сегмента ST при пиковой нагрузке

- Down: наклон вниз

- Flat: нет наклона (плоский)

- Up: наклон вверх

12. HeartDisease: Сердечная недостаточность: выходной класс (1: болезнь сердца, 0: норма)

2 Ход выполнения

Для начала я проверил датасет на присутствие в нем нулей (пустых ячеек), для этого вывел информацию о нем с помощью метода info().

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             918 non-null    int64
 1   Sex             918 non-null    object
 2   ChestPainType   918 non-null    object
 3   RestingBP       918 non-null    int64
 4   Cholesterol     918 non-null    int64
 5   FastingBS       918 non-null    int64
 6   RestingECG      918 non-null    object
 7   MaxHR           918 non-null    int64
 8   ExerciseAngina  918 non-null    object
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object
 11  HeartDisease    918 non-null    int64
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
None
```

Как видно, признаков с пустыми значениями не наблюдается. Если посмотреть на таблицу, можно заметить еще одну проблему - некоторые данные записаны в виде символов или слов:

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

Я исправил это следующим образом: каждому слову/символу для каждого столбца сопоставил число следующим образом:

- пол пациена (0 - мужчина, 1 - женщина)
- тип боли в груди (0 - TA, 1 - ATA, 2 - NAP, 3 - ASY)

- результаты ЭКГ в покое (0 - Normal, 1 - ST, 2 - LVH)
- стенокардия, вызванная физической нагрузкой (1 - да, 0 - нет)
- наклон сегмента ST при пиковой нагрузке (0 - Down, 1 - Flat, 2 - Up)

Получилось следующее:

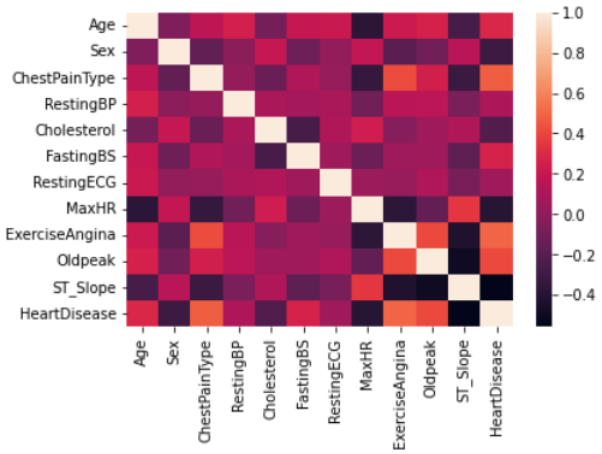
	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	0	1	140	289	0	0	172	0	0.0	2	0
1	49	1	2	160	180	0	0	156	0	1.0	1	1
2	37	0	1	130	283	0	1	98	0	0.0	2	0
3	48	1	3	138	214	0	0	108	1	1.5	1	1
4	54	0	2	150	195	0	0	122	0	0.0	2	0

Потом я попробовал удалить повторяющиеся строки, но таких не оказалось.

Далее перешел непосредственно к анализу данных. Построил корреляционную матрицу.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
Age	1.000000	-0.055750	0.165896	0.254399	-0.095282	0.198039	0.213152	-0.382045	0.215793	0.258612	-0.268264	0.282039
Sex	-0.055750	1.000000	-0.168254	-0.005133	0.200092	-0.120076	0.018343	0.189186	-0.190664	-0.105734	0.150693	-0.305445
ChestPainType	0.165896	-0.168254	1.000000	0.022168	-0.136139	0.116703	0.031383	-0.343654	0.416625	0.245027	-0.317480	0.471354
RestingBP	0.254399	-0.005133	0.022168	1.000000	0.100893	0.070193	0.097661	-0.112135	0.155101	0.164803	-0.075162	0.107589
Cholesterol	-0.095282	0.200092	-0.136139	0.100893	1.000000	-0.260974	0.112095	0.235792	-0.034166	0.050148	0.111471	-0.232741
FastingBS	0.198039	-0.120076	0.116703	0.070193	-0.260974	1.000000	0.050707	-0.131438	0.060451	0.052698	-0.175774	0.267291
RestingECG	0.213152	0.018343	0.031383	0.097661	0.112095	0.050707	1.000000	0.048552	0.036119	0.114428	-0.078807	0.061011
MaxHR	-0.382045	0.189186	-0.343654	-0.112135	0.235792	-0.131438	0.048552	1.000000	-0.370425	-0.160691	0.343419	-0.400421
ExerciseAngina	0.215793	-0.190664	0.416625	0.155101	-0.034166	0.060451	0.036119	-0.370425	1.000000	0.408752	-0.428706	0.494282
Oldpeak	0.258612	-0.105734	0.245027	0.164803	0.050148	0.052698	0.114428	-0.160691	0.408752	1.000000	-0.501921	0.403951
ST_Slope	-0.268264	0.150693	-0.317480	-0.075162	0.111471	-0.175774	-0.078807	0.343419	-0.428706	-0.501921	1.000000	-0.558771
HeartDisease	0.282039	-0.305445	0.471354	0.107589	-0.232741	0.267291	0.061011	-0.400421	0.494282	0.403951	-0.558771	1.000000

или тоже самое картинкой



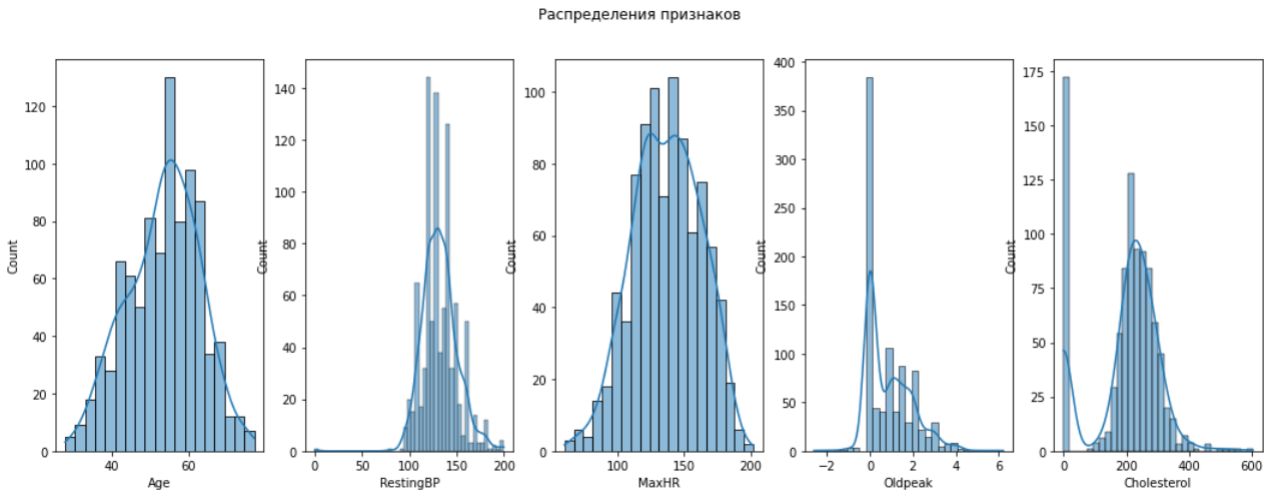
Так как меня больше всего интересует параметр HeartDisease, я начал анализировать его. Из матрицы видно, что больше всего на этот параметр влияют наклон сегмента ST при пиковой нагрузке, наличие стенокардии, а также тип боли в груди. А меньше всего - результаты электрокардиограммы в покое. Также из матрицы видно, что остальные параметры сильно между собой некоррелируют, следовательно, можно не удалять/объединять их.

Дальше я рассмотрел параметры моих признаков в датасете.

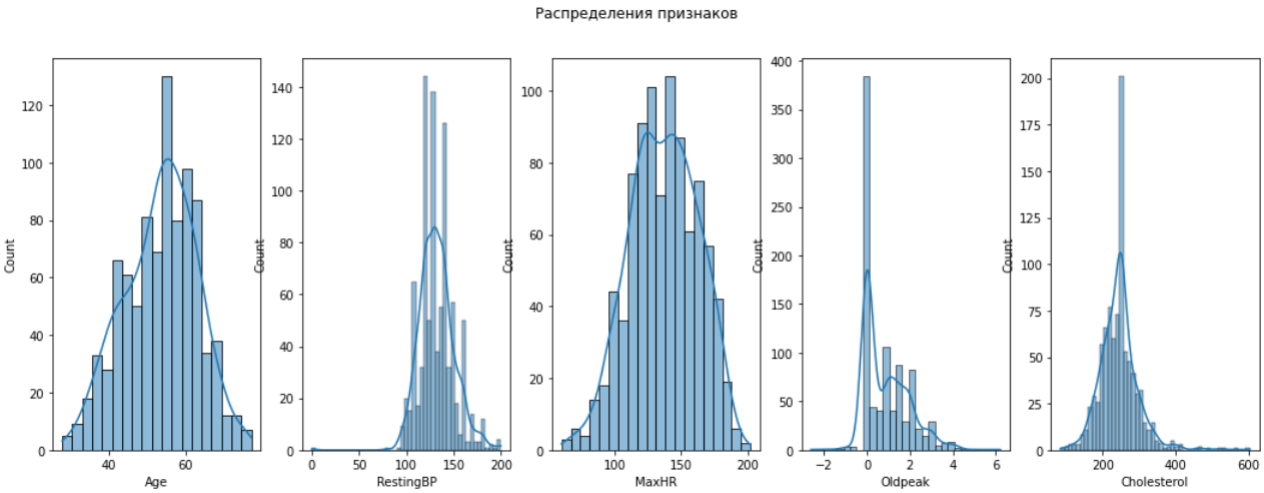
	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	0.210240	2.251634	132.396514	198.799564	0.233115	0.603486	136.809368	0.404139	0.887364	1.361656	0.553377
std	9.432617	0.407701	0.931031	18.514154	109.384145	0.423046	0.805968	25.460334	0.490992	1.066570	0.607056	0.497414
min	28.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	60.000000	0.000000	-2.600000	0.000000	0.000000
25%	47.000000	0.000000	2.000000	120.000000	173.250000	0.000000	0.000000	120.000000	0.000000	0.000000	1.000000	0.000000
50%	54.000000	0.000000	3.000000	130.000000	223.000000	0.000000	0.000000	138.000000	0.000000	0.600000	1.000000	1.000000
75%	60.000000	0.000000	3.000000	140.000000	267.000000	0.000000	1.000000	156.000000	1.000000	1.500000	2.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	603.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	1.000000

Отсюда видно, что данные ненормированы. Это придется сделать по мере реализации модели. Также можно подтвердить, что данные находятся в указанных диапазонах, как и говорится в описании к датасету (например, MaxHR от 60 до 202).

Потом я построил распределение признаков. Для колличественных я построил графики, где по оси x указывается значение параметра, а по y - его количество.

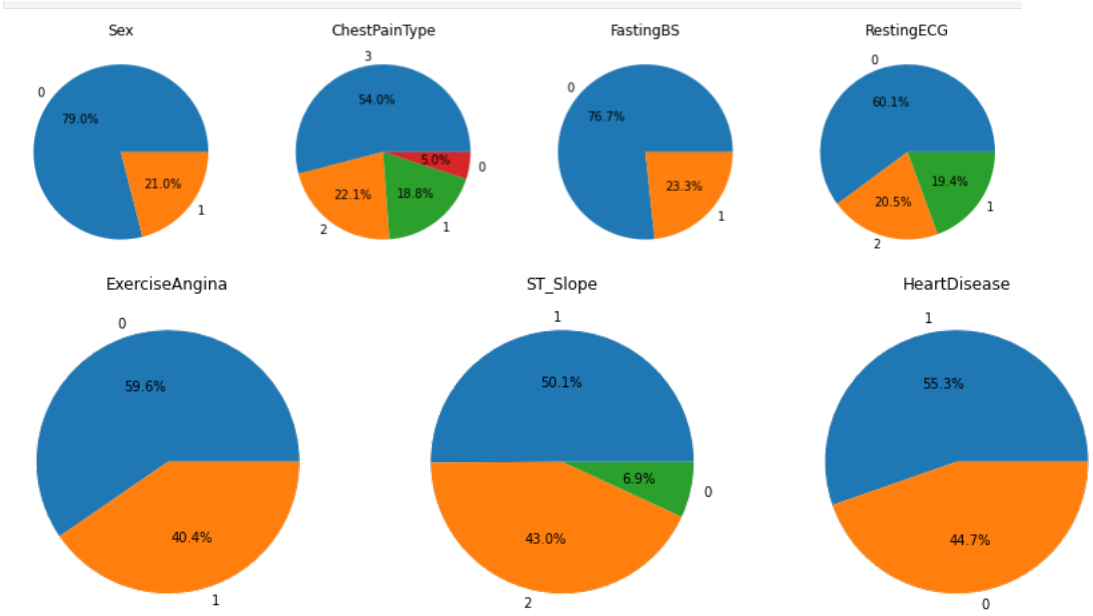


Из графиков видно, что значения распределены неравномерно. Их распределение отдалённо более похоже на нормальное .Однако на графике с холестерином прослеживается большое количество нулей, чего быть не может. Поэтому, так как данных у меня мало, я для каждого из исходов(здоровый или больной) посчитал средние значения холестерина, и заменил нули ими. Получилось следующее:



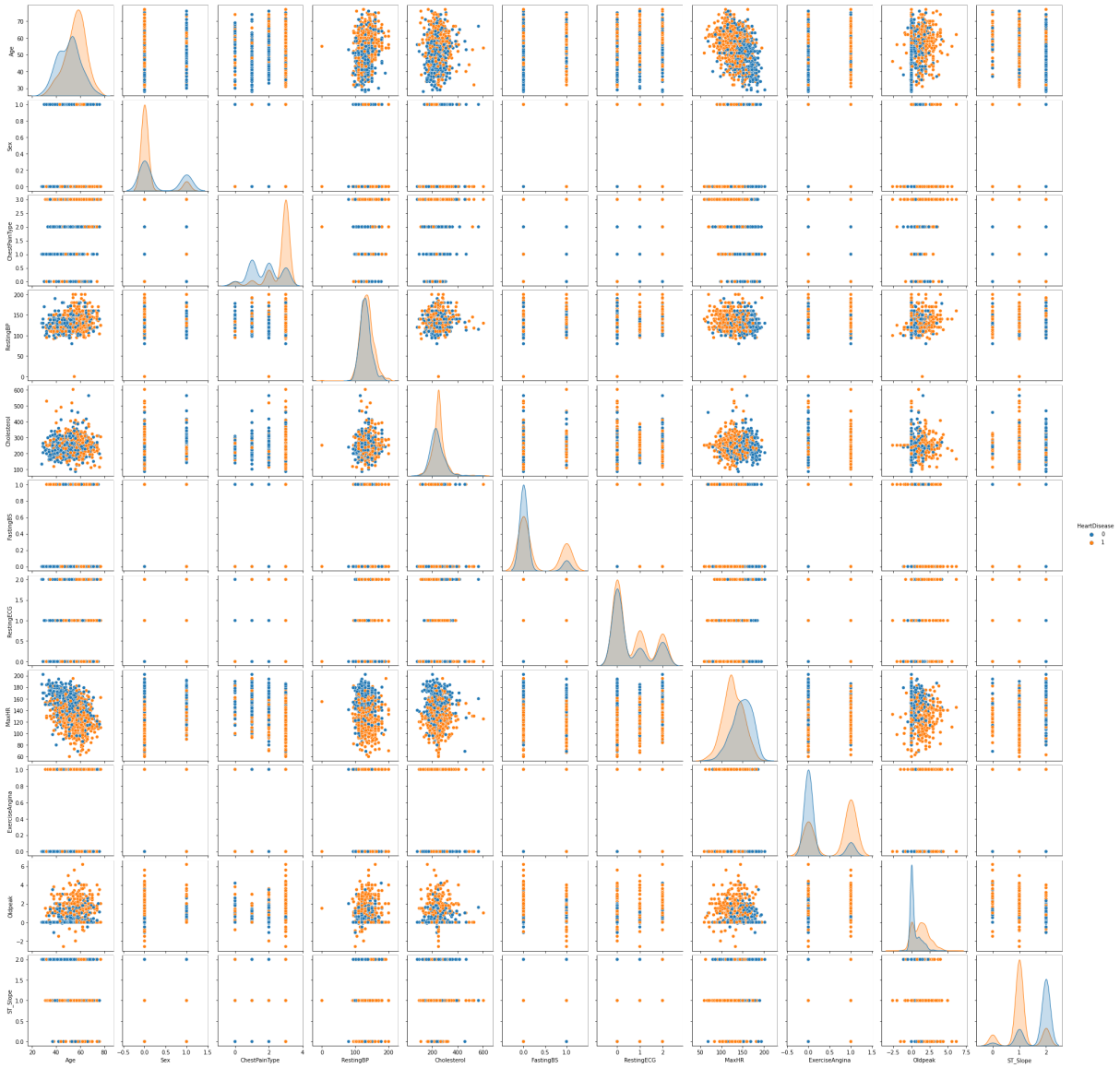
Данный график халестерина более соответствует действительности. Также на графике oldpeak видно, что много людей с ST = 0, что является патологией. Скорее всего, датасет был основан на людях с больным сердцем.

Далее я построил диаграммы для категориальных признаков.



Тут тоже многие данные распределены неравномерно. Я обратил внимание, что больных и здоровых людей примерно одинаково, так что оверсемплинг делать не требуется.

После этого я построил парные графики, при этом выделил цветом данные параметра HeartDisease.



Данные распределены достаточно хаотично, но где-то всё-таки можно как-то провести разделяющую прямую. Добавлять новые признаки не требуется. Данные готовы к обучению.

3 Выводы

Данная лабораторная работа дала мне интересный опыт в анализе данных. После нейросетей непревычно, что надо как-то преобразовывать дадасет, делать различные графики и рисунки. Однако, это позоляет глубже углубиться в процесс, лучше понять что происходит сейчас, и будет происходить при обучении модели. Тут я столкнулся с такими проблемами, как выбор качественного датасета, преобразования данных в нём, устранение выбросов путём усреднения других данных параметра. Во время работы уже приходили идеи по реализации непосредственно самой модели, например, использование кросс валидации. В целом, я улучшил навыки работы с данными.