МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МСК

ОТЧЕТ

по лабораторной работе №1

по дисциплине «Машинное обучение»

Тема: Предобработка данных

Студенка гр. 8303 _____ Самойлова А.С.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2021

Цель работы

Ознакомиться с методами предобработки данных из библиотеки *Scikit Learn*.

Ход выполнения работы

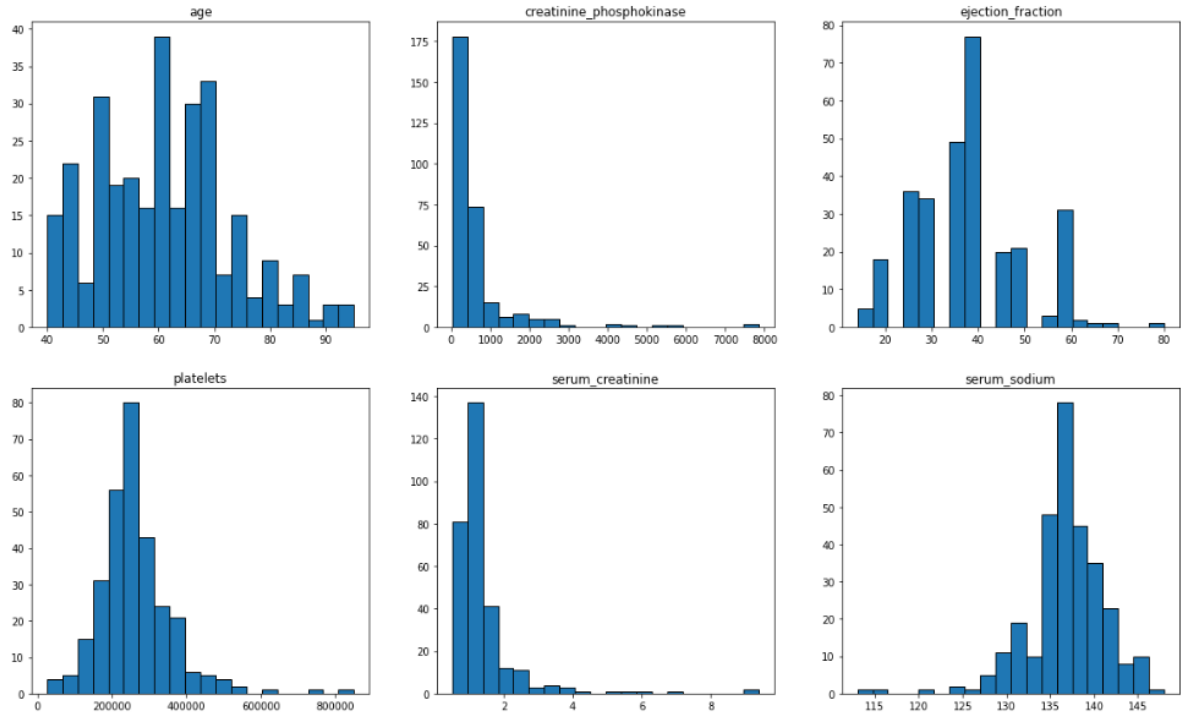
Загрузка данных

1. Загрузить датасет (https://www.kaggle.com/andrewmvd/heart-failure-clinical-data).
2. Создать Python скрипт. Загружзить датасет в датафрейм, и исключить бинарные признаки и признак времени:

	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium
0	40.0	23	14	25100.0	0.5	113
1	40.0	30	15	47000.0	0.6	116
2	40.0	47	15	51000.0	0.6	121
3	40.0	47	17	62000.0	0.6	124
4	40.0	47	17	70000.0	0.6	125
...	...	...	...	...	...	...
294	90.0	4540	62	533000.0	5.8	145
295	90.0	5209	62	543000.0	6.1	145
296	94.0	5882	65	621000.0	6.8	145
297	95.0	7702	70	742000.0	9.0	146
298	95.0	7861	80	850000.0	9.4	148

299 rows × 6 columns

3. Построить гистограммы признаков:



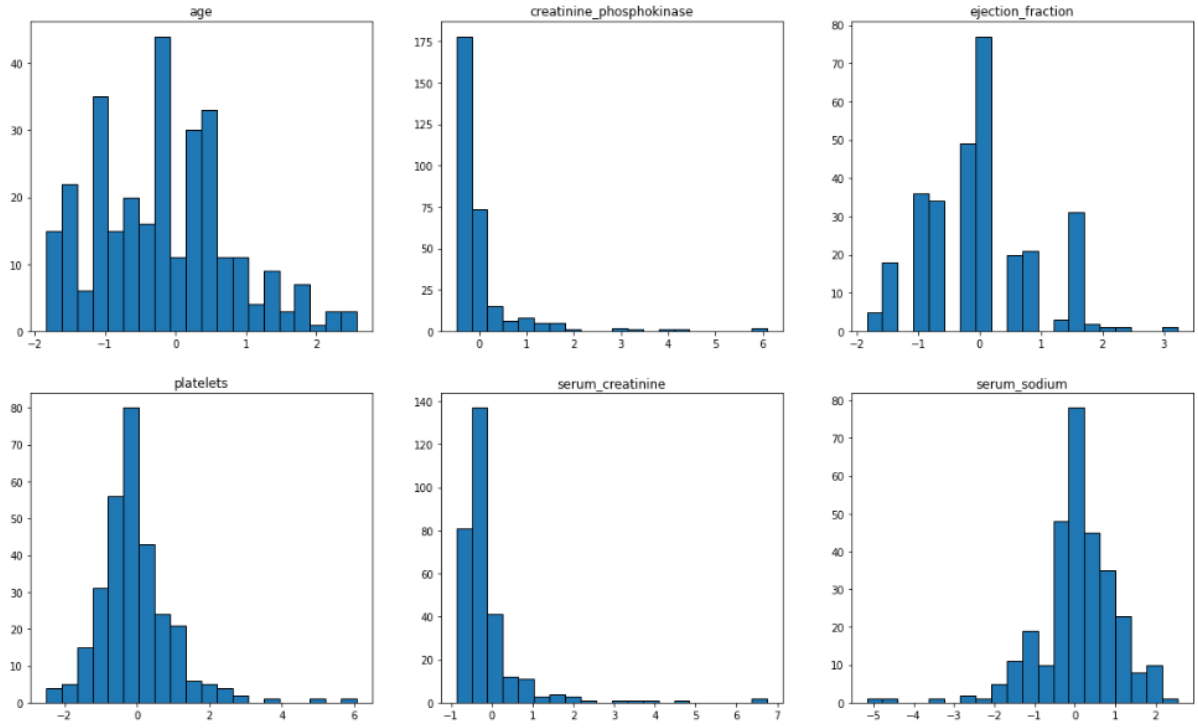
4. На основании гистограмм определить диапазоны значений для каждого из признаков, а так же возле какого значения лежит наибольшее количество наблюдений:

	Диапазон	Наибольшее количество наблюдений
age	(40, 95)	60
creatinine_phosphokinase	(0, 7900)	0
ejection_fraction	(14, 80)	40
platelets	(25000, 850000)	225000
serum_creatine	(0.5, 9.2)	1
serum_sodium	(113, 148)	136

5. Так как библиотека *Sklearn* работает с *NumPy* массива, преобразовать датафрейм к двумерному массиву *NumPy*, где строка соответствует наблюдению, а столбец признаку.

Стандартизация данных

1. Подключить модуль *Sklearn*. Настроить стандартизацию на основе первых 150 наблюдений используя *StandardScaler*.
2. Стандартизировать данные.
3. Построить гистограммы признаков:



4. Сравнить данные до и после:

	Диапазон	Наибольшее количество наблюдений
age	(-1.75, 2.6)	-0.2
creatinine_phosphokinase	(-0.5, 6.1)	-0.4
ejection_fraction	(-2.2, 3.2)	0.1
platelets	(-2.5, 6)	-0.2
serum_creatine	(-0.9, 6.6)	0.1
serum_sodium	(-5.1, 2.5)	0

Из таблицы видно, что наибольшее количество наблюдений находится возле 0, а сам разброс значений находится в пределах одного порядка.

5. Рассчитать мат. ожидание и СКО до и после стандартизации. На основе этих значений вывести формулы по которым они стандартизировались.

	μ до	σ до	μ после	σ после
age	60.83	11.89	-0.17	0.96
creatinine_phosphokinase	581.84	970.29	-0.02	0.82
ejection_fraction	38.08	11.83	0.01	0.91
platelets	263358	97804.24	-0.04	1.02
serum_creatine	1.39	1.03	-0.11	0.89
serum_sodium	136.63	4.41	0.04	0.97

Исходя из таблицы можно сделать вывод, что стандартизация данных приводит к тому, что математическое ожидание становится равным 0, а среднеквадратичное отклонение равным 1. Следовательно формула преобразования данных имеет вид:

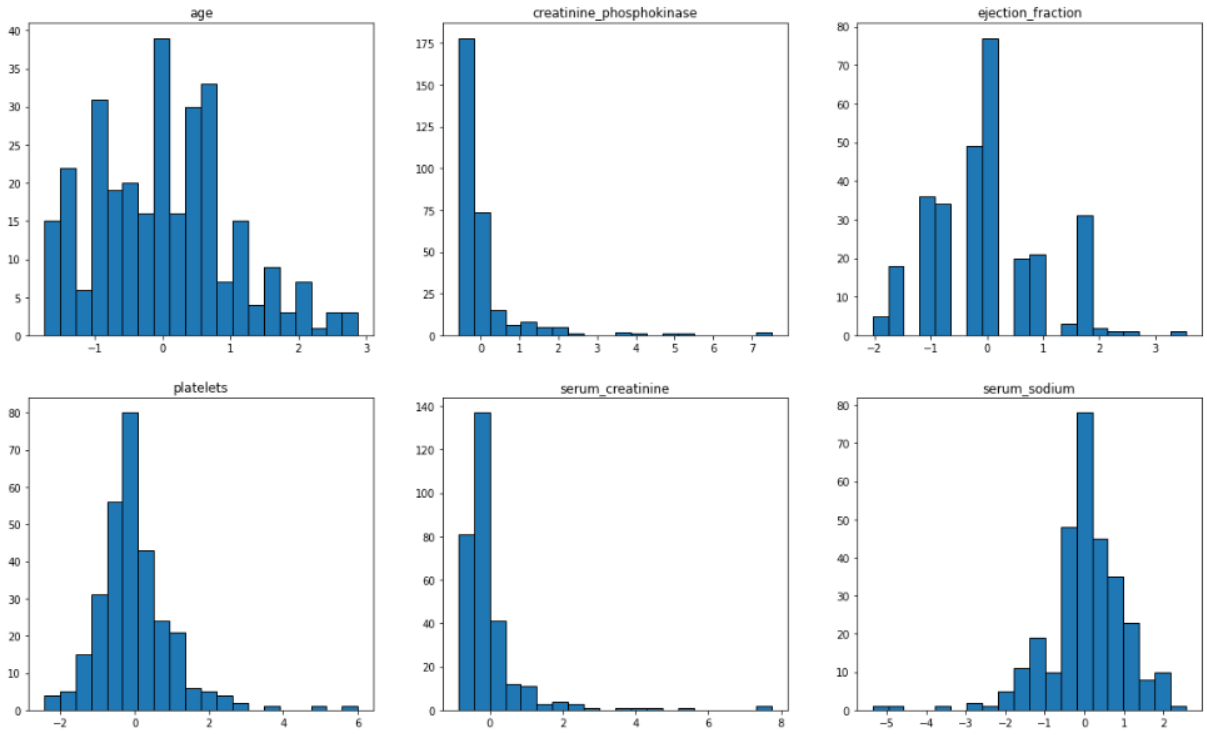
$$y = \frac{x - \mu(X)}{\sigma(X)}$$

где X — исходный набор данных.

6. Сравнить значения формул с полями *mean_* и *var_* объекта *scaler*:

	mean_	var_	σ
age	62.95	155.00	12.45
creatinine_phosphokinase	607.15	1415489	1189.74
ejection_fraction	37.95	170.02	13.04
platelets	266746.75	9.252860e+9	96191.79
serum_creatine	1.52	1.36	1.17
serum_sodium	136.45	20.61	4.54

7. Провести настройку стандартизации на всех данных и сравнить с результатаами настройки на основании 150 наблюдений:

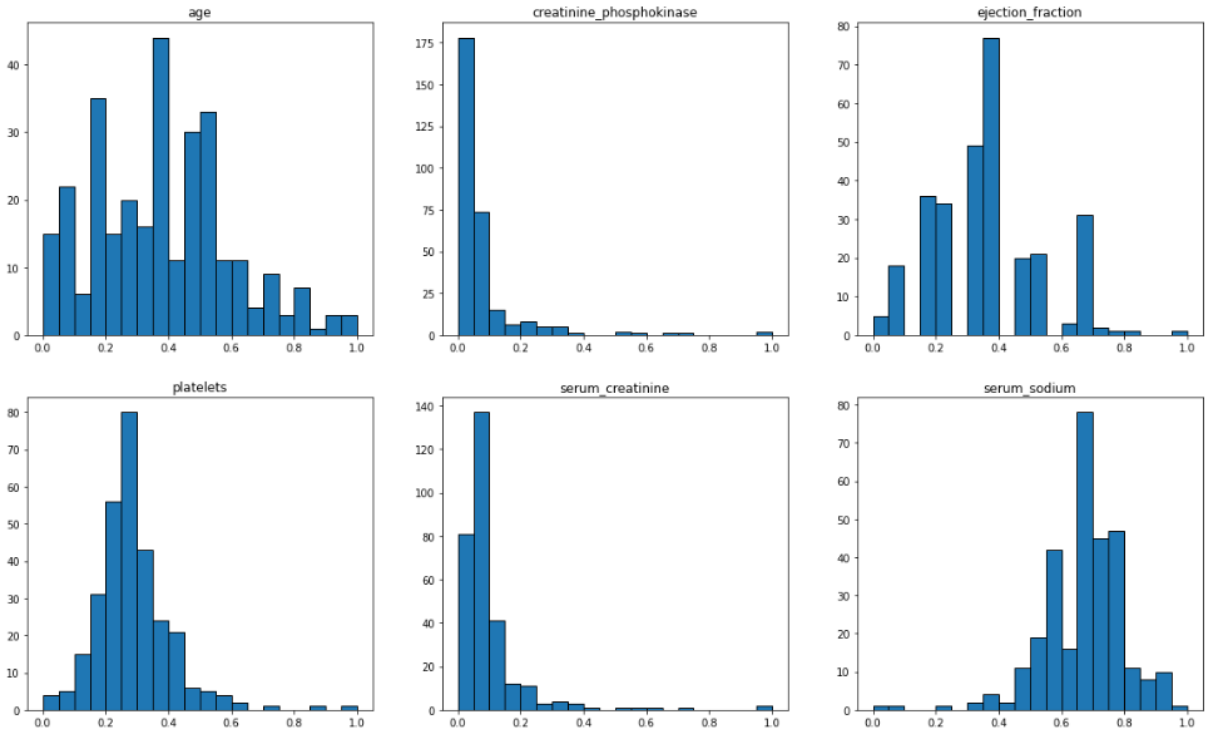


	μ до	σ до	μ после	σ после
age	60.83	11.89	0.00	1.00
creatinine_phosphokinase	581.84	970.29	0.00	1.00
ejection_fraction	38.08	11.83	0.00	1.00
platelets	263358	97804.24	0.00	1.00
serum_creatine	1.39	1.03	0.00	1.00
serum_sodium	136.63	4.41	0.00	1.00

	mean_	var_	σ
age	60.83	141.01	11.87
creatinine_phosphokinase	581.84	938309.9	968.66
ejection_fraction	38.08	139.60	11.82
platelets	263358	9.533677e+9	97640.55
serum_creatine	1.39	1.07	1.03
serum_sodium	136.63	1.94	4.41

Приведение к диапазону

1. Привести данные к диапазону используя *MinMaxScaler*.
2. Построить гистограммы для признаков:

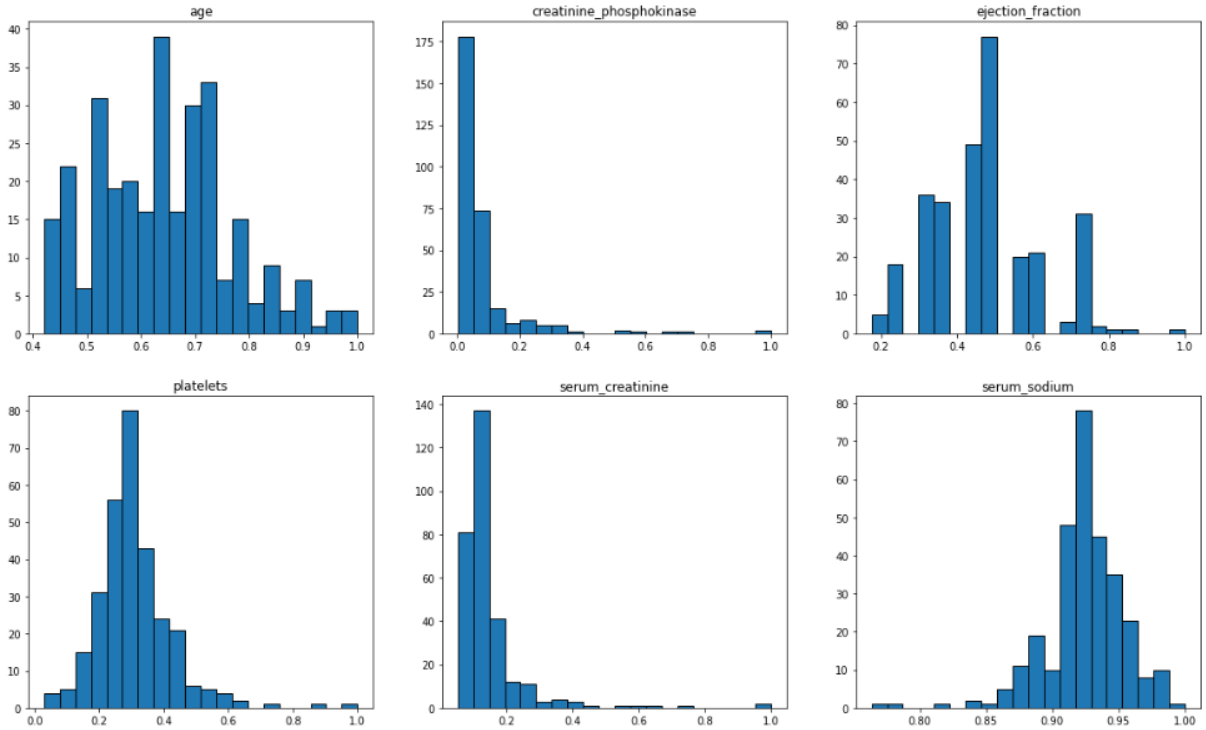


3. Чераз параметры *MinMaxScaler* определить инимальное и максимальное значение в данных для каждого признака.

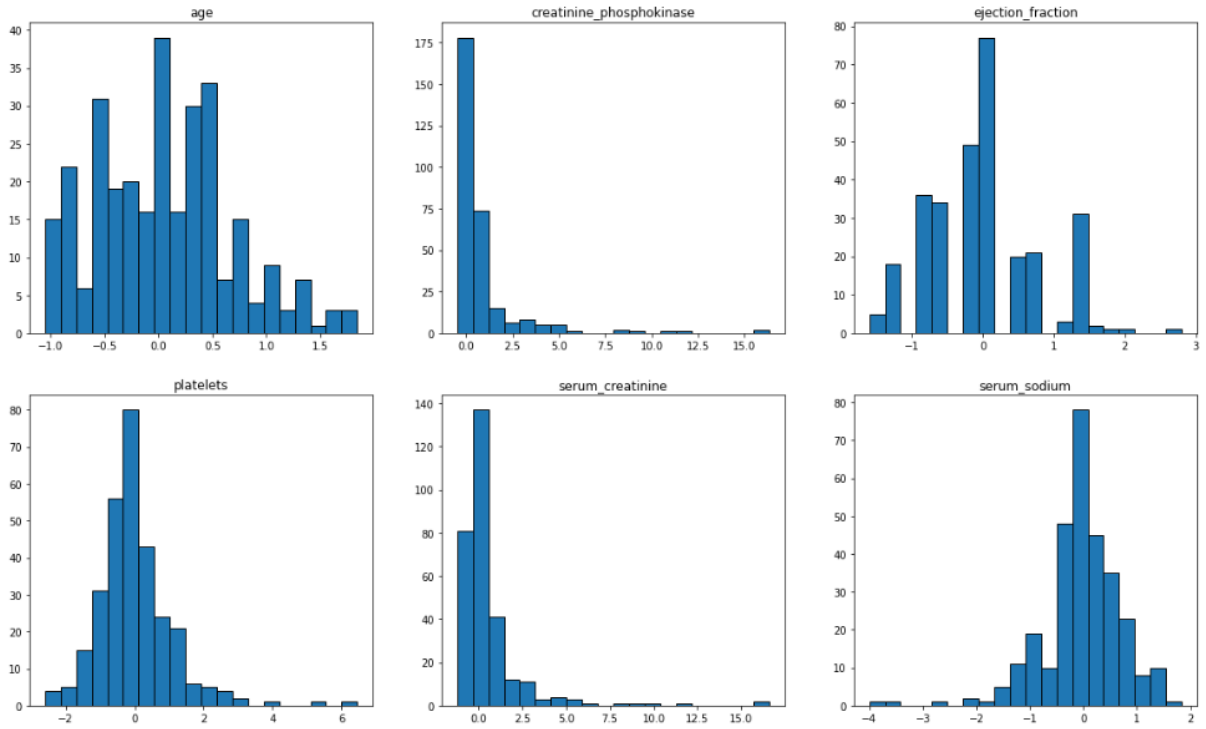
	min до	max до	min после	max после
age	40.0	95.0	0.0	1.0
creatinine_phosphokinase	23.0	7861.0	0.0	1.0
ejection_fraction	14.0	80.0	0.0	1.0
platelets	25100.0	850000.0	0.0	1.0
serum_creatine	0.5	9.4	0.0	1.0
serum_sodium	113.0	148.0	0.0	1.0

4. Аналогично трансформировать данные используя *MaxAbsScaler* и *RobustScaler*. Построить гистограммы. Определить к какому диапазону приводяттся данные.

MaxAbsScaler



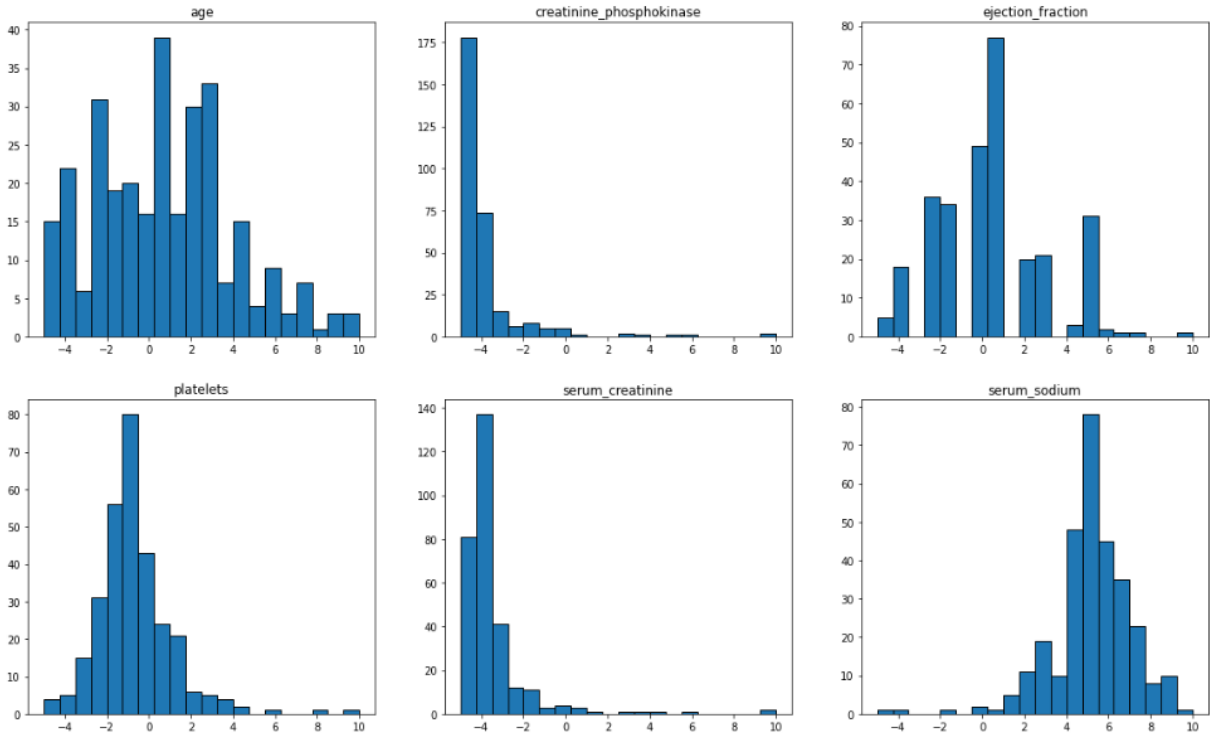
RobustScaler



5. Написать функцию, которая приводит данные к диапазону [-5, 10].

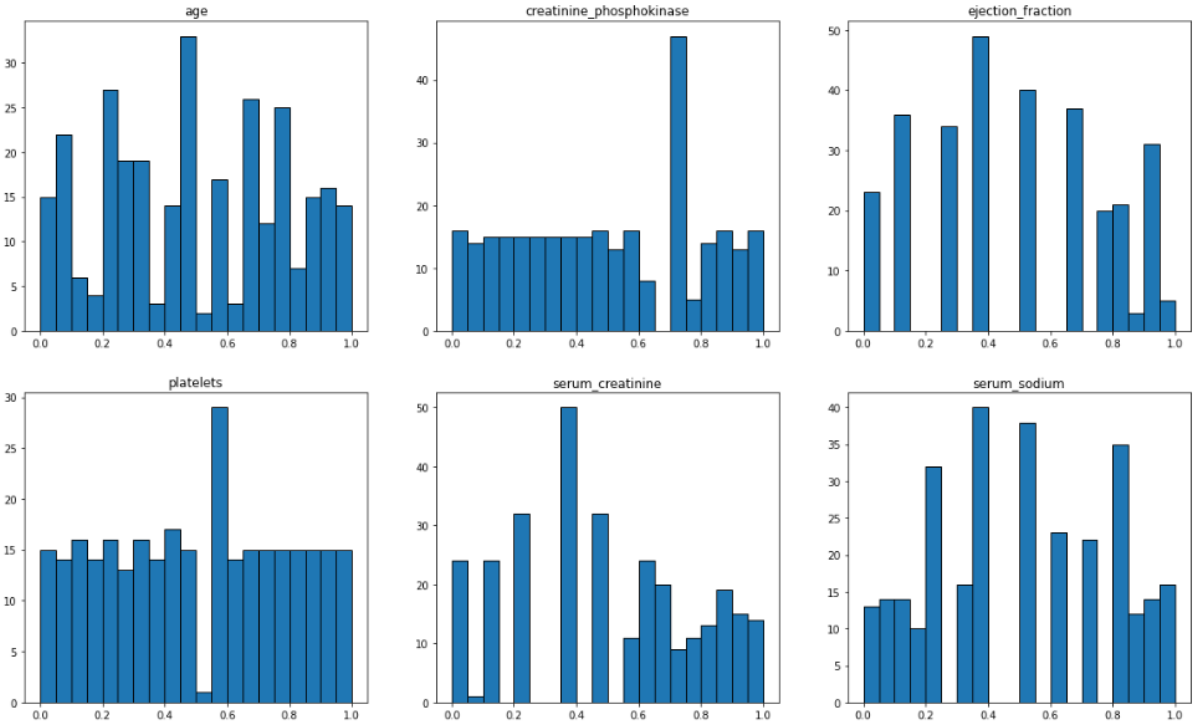
   Для приведения к диапазону [-5, 10] будет использована формула:

$$y = \frac{15 \cdot (x - min(X))}{max(X) - min(X)} - 5$$



Нелинейные преобразования

1. Привести данные к равномерному распределению используя *QuantileTransformer*.
2. Построить гистограммы и сравнить с исходными данными.

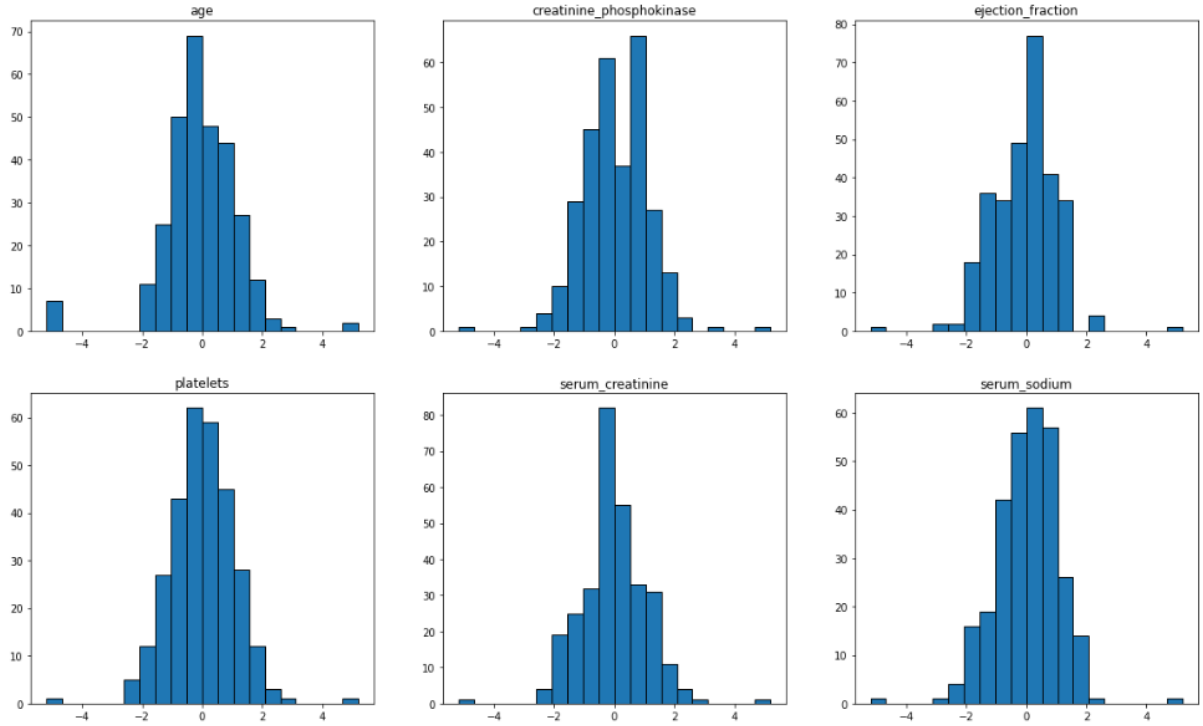


3. Определить как и на что влияет значение параметра *n_quantiles*.

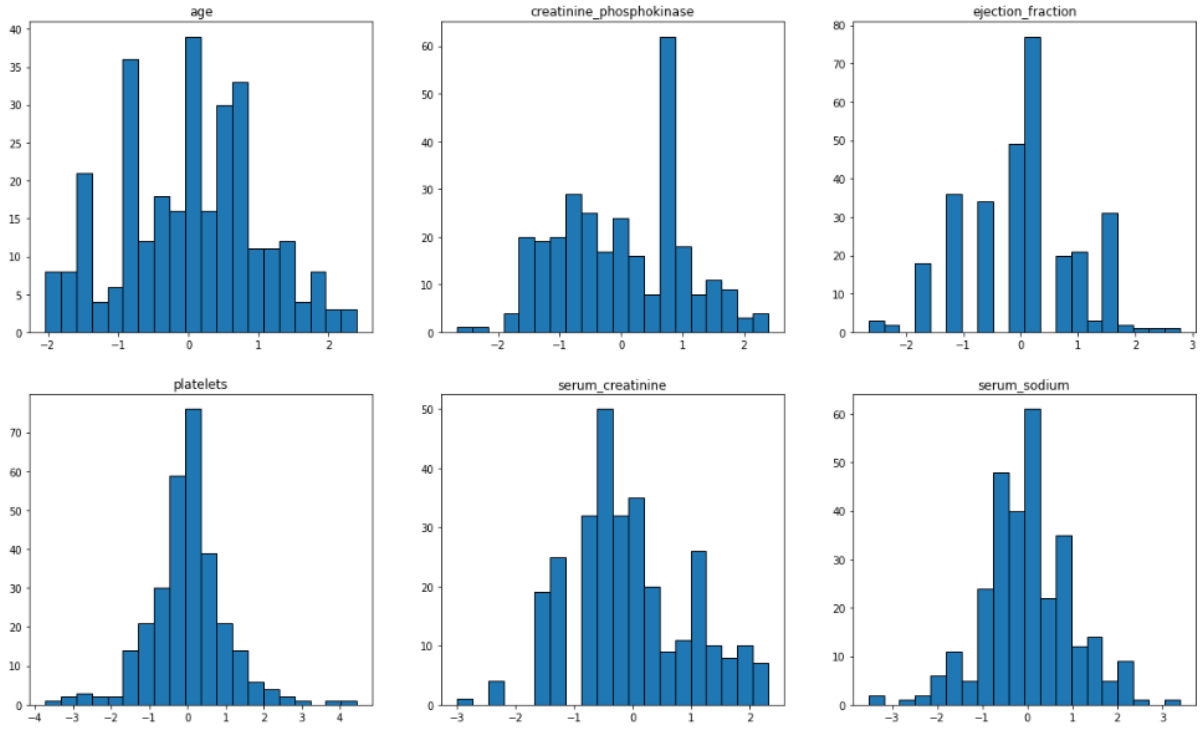
   Параметр определяет количество квантилей, которое будет использовано для дискретизации функции распределения.

4. Привести данные к нормальному распределению передав в *QuantileTransformer* параметр *output_distribution=“normal“*.

5. Построить гистограммы:

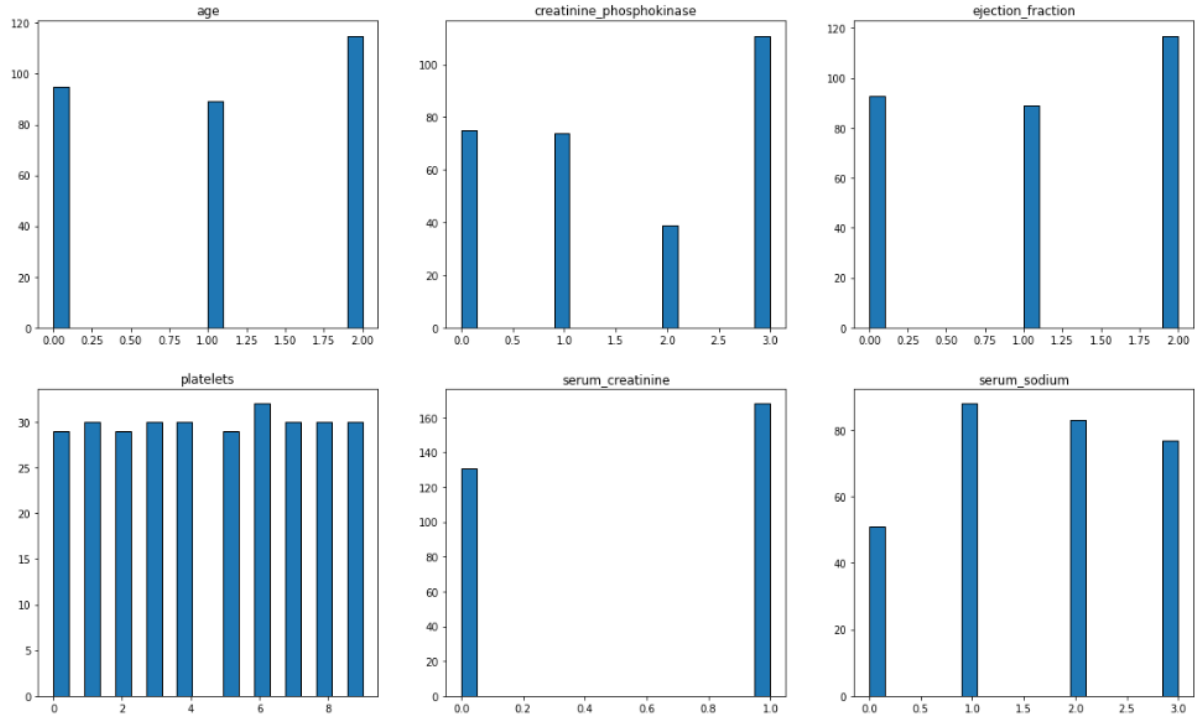


6. Привести данные к нормальному распределению используя *PowerTransformer*.



Дискретизация признаков

1. Провести дискретизацию признаков, используя *KbinsDiscretizer* на следующее колиество диапазонов:
   - age — 3
   - creatinine_phosphokinase — 4
   - ejection_fraction — 3
   - platelets — 10
   - serum_creatine — 2
   - serum_sodium — 4
2. Построить гистограммы:



3. Через параметр *bin_edges_* вывести диапазоны каждого интервала для каждого признака.

age	[40, 55, 65, 95]
creatinine_phosphokinase	[23, 116.5, 250, 582, 7861]
ejection_fraction	[14, 35, 40, 80]
platelets	[25100, 153000, 196000, 221000, 237000, 262000, 265000, 285200, 319800, 374600, 850000]
serum_creatine	[0.5, 1.1, 9.4]
serum_sodium	[113, 134, 137, 140, 148]

Вывод

В ходе выполнения лабораторной работы были изучены методы предобработки данных, на практике изучен метод Scikit Learn. Исходя из полученных результатов сделаны выводы:

- стандартизация по неполным данным даёт неточный результат;
- при линейных преобразованиях распределение не меняется, меняется только диапазон значений;
- нелинейные преобразования позволяют привести данные к любой другой форме.