МИНОБРНАУКИ РОССИИ

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

ОТЧЁТ
по лабораторной работе №1
по дисциплине «Машинное обучение»

Тема: «Предобработка данных»

Студент гр. 6307 Гарифуллин В.Ф.

Преподаватель Жангиров Т.Р.

Санкт-Петербург
2020

Цель работы

Ознакомиться с методами предобработки данных из библиотеки Scikit Learn

Ход работы

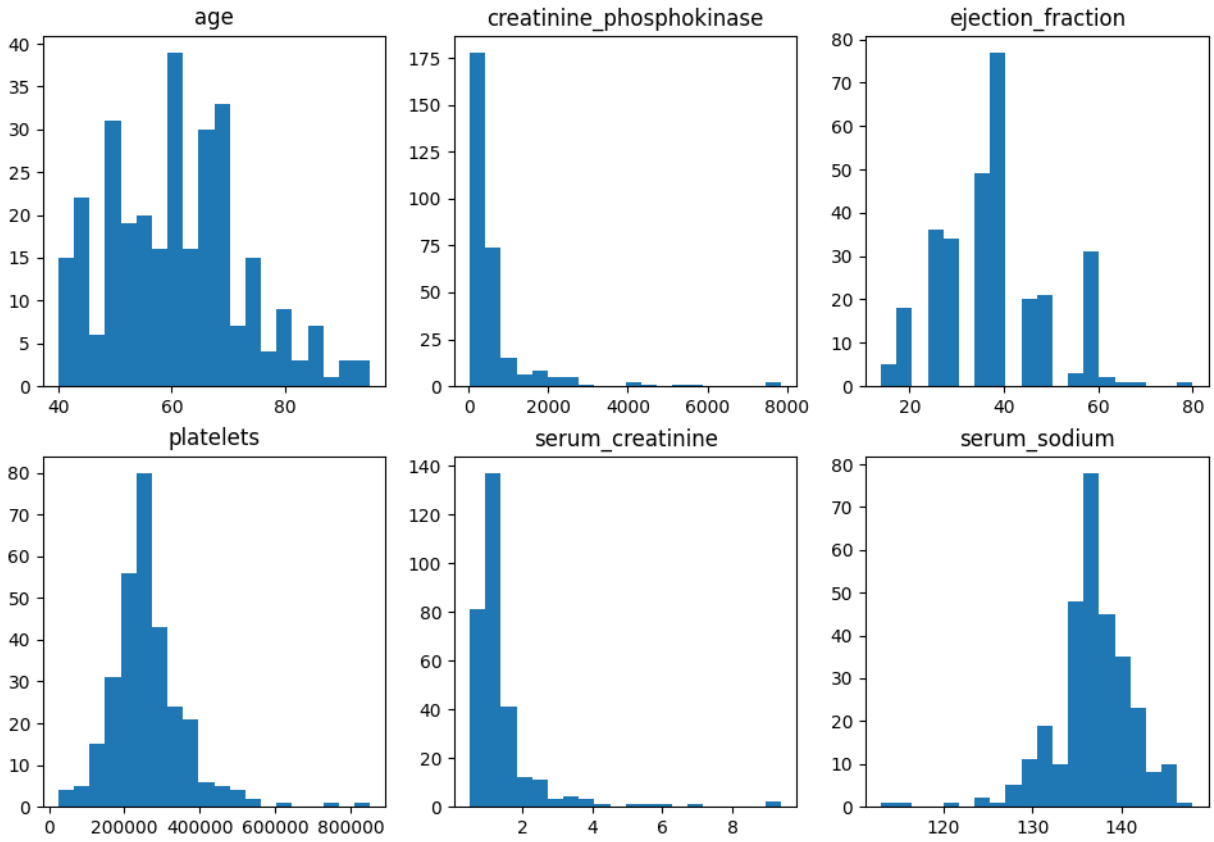
Загрузка данных

1. Вывод датафрейма с данными для лаб. работы. Должно быть 299 наблюдений и 6 признаков

	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium
0	75.0	582	20	265000.00	1.9	130
1	55.0	7861	38	263358.03	1.1	136
2	65.0	146	20	162000.00	1.3	129
3	50.0	111	20	210000.00	1.9	137
4	65.0	160	20	327000.00	2.7	116
..	...	...	...	...	...	...
294	62.0	61	38	155000.00	1.1	143
295	55.0	1820	38	270000.00	1.2	139
296	45.0	2060	60	742000.00	0.8	138
297	45.0	2413	38	140000.00	1.4	140
298	50.0	196	45	395000.00	1.6	136

[299 rows x 6 columns]

2. Гистограммы признаков

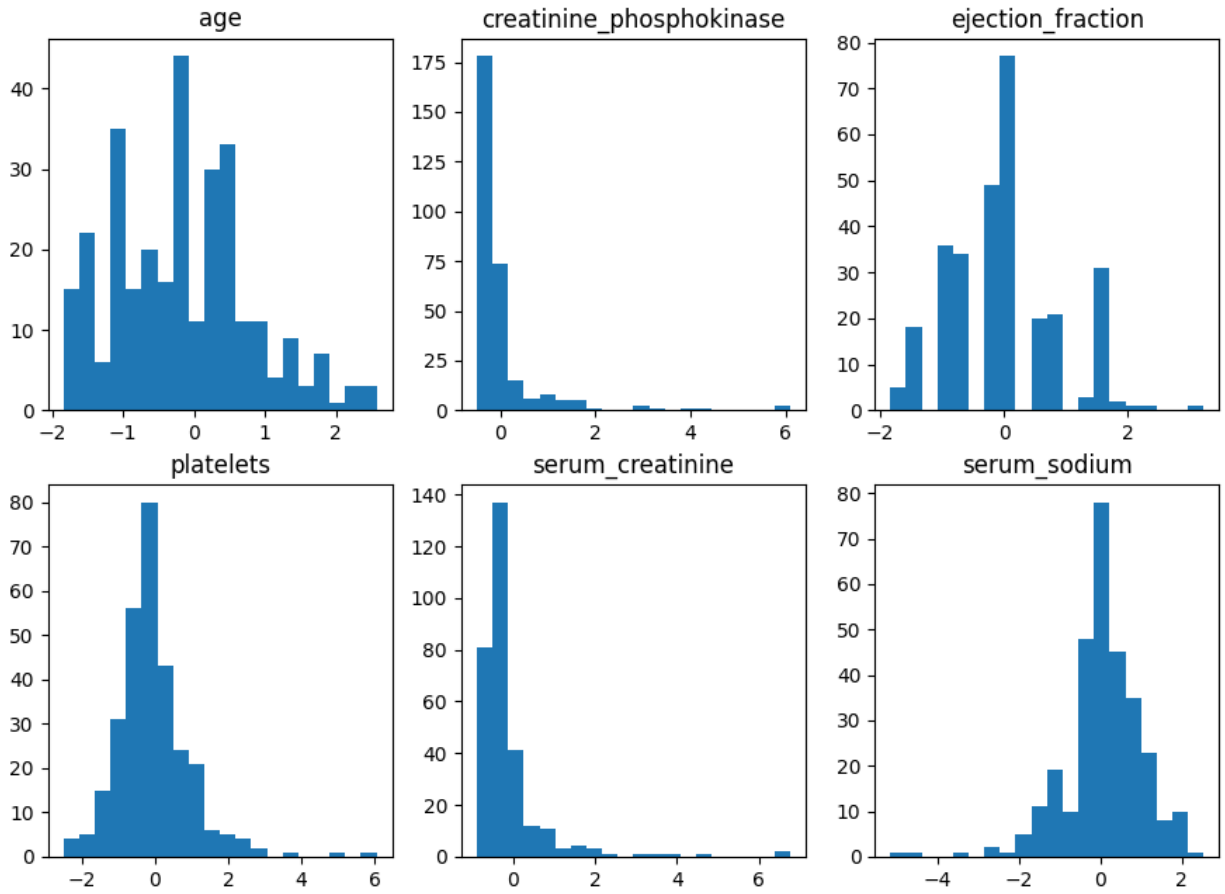


3. На основании гистограмм определите диапазоны значений для каждого из признаков, а также возле какого значения лежит наибольшее количество наблюдений.

age: 40 – 95, 60
creatinine_phosphokinase: 23 – 7850, 200
ejection_fraction: 14 – 80, 39
platelets: 25000 – 850000, 250000
serum_crearinine: 0.5 – 9.4, 1.2
serum_sodium: 113 – 148, 136

Стандартизация данных

1. Постройте гистограммы стандартизированных данных



2. Сравните данные до и после стандартизации. Опишите, что изменилось и почему.

Теперь на оси X значение, рядом с которым лежит наибольшее количество наблюдений соответствует нулю.

3. Рассчитайте мат. ожидание и СКО до и после стандартизации. На основании этих значений выведите для каждого признака формулы по которым они стандартизировались.

Признак	МО до	МО после	СКО до	СКО после
age	60	-0.169	11.87	0.95
creatinine_phosphokinase	581	-0.0212	968.66	0.81
ejection_fraction	38	0.0105	11.81	0.9
platelets	263358	-0.035	97640	1.01
serum_crearinine	1.39	-0.108	1.03	0.88
serum_sodium	136	0.0379	4.4	0.97

Значение = (исходное_значение - МО) / СКО

4. Сравните значений из формул с полями mean_ и var_ объекта scaler

Признак	mean_	var_
age	62.9	155
creatinine_phosphokinase	607	1415489
ejection_fraction	37.9	170
platelets	266746	9252860500
serum_crearinine	1.52	1.36
serum_sodium	136	21

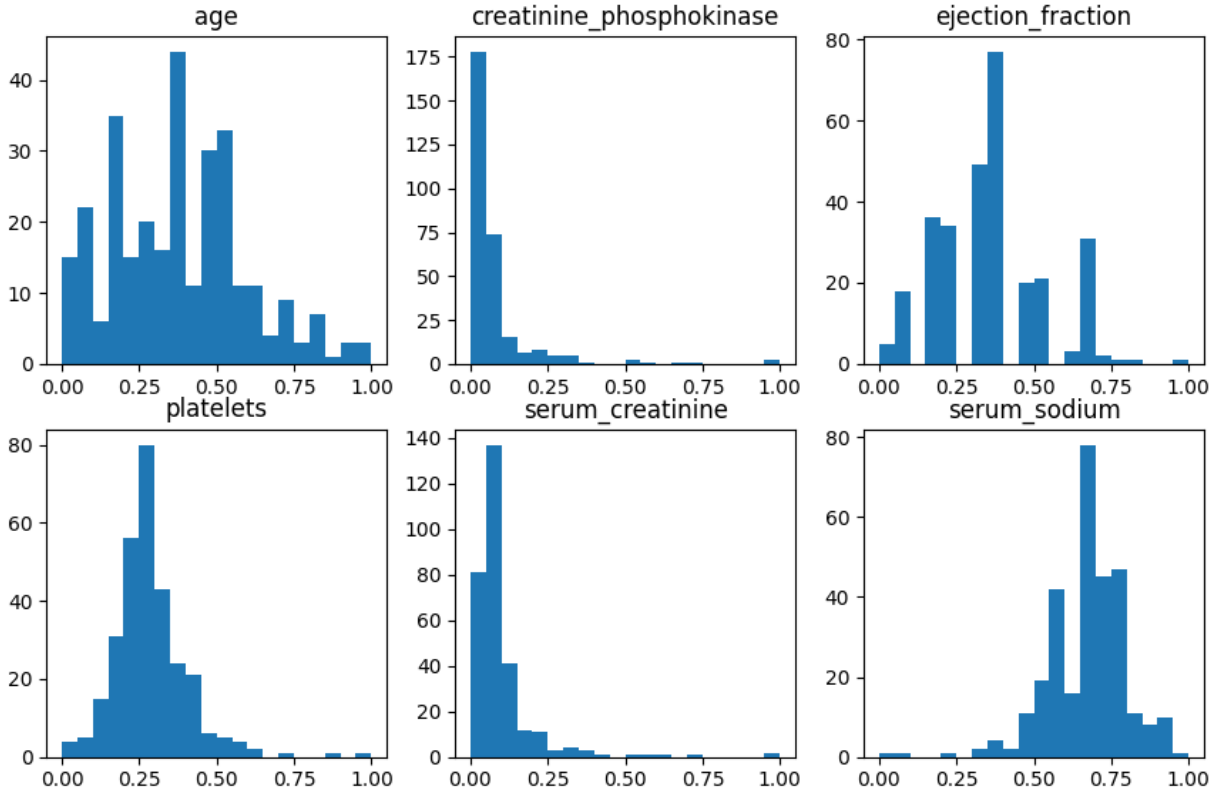
5. Проведите настройку стандартизации на всех данных и сравните с результатами настройки на основании 150 наблюдений

Признак	МО все	МО 150	СКО все	СКО 150
age	5.70335306e-16	-0.169	1	0.95
creatinine_phosphokinase	0	-0.0212	1	0.81
ejection_fraction	-3.26754603e-17	0.0105	1	0.9
platelets	7.72329061e-17	-0.035	1	1.01
serum_crearinine	1.42583827e-16	-0.108	1	0.88
serum_sodium	-8.67384945e-16	0.0379	1	0.97

Стандартизация на основе всех данных даёт более точный результат.

Приведение к диапазону

1. Постройте гистограммы для признаков и сравните с исходными данными

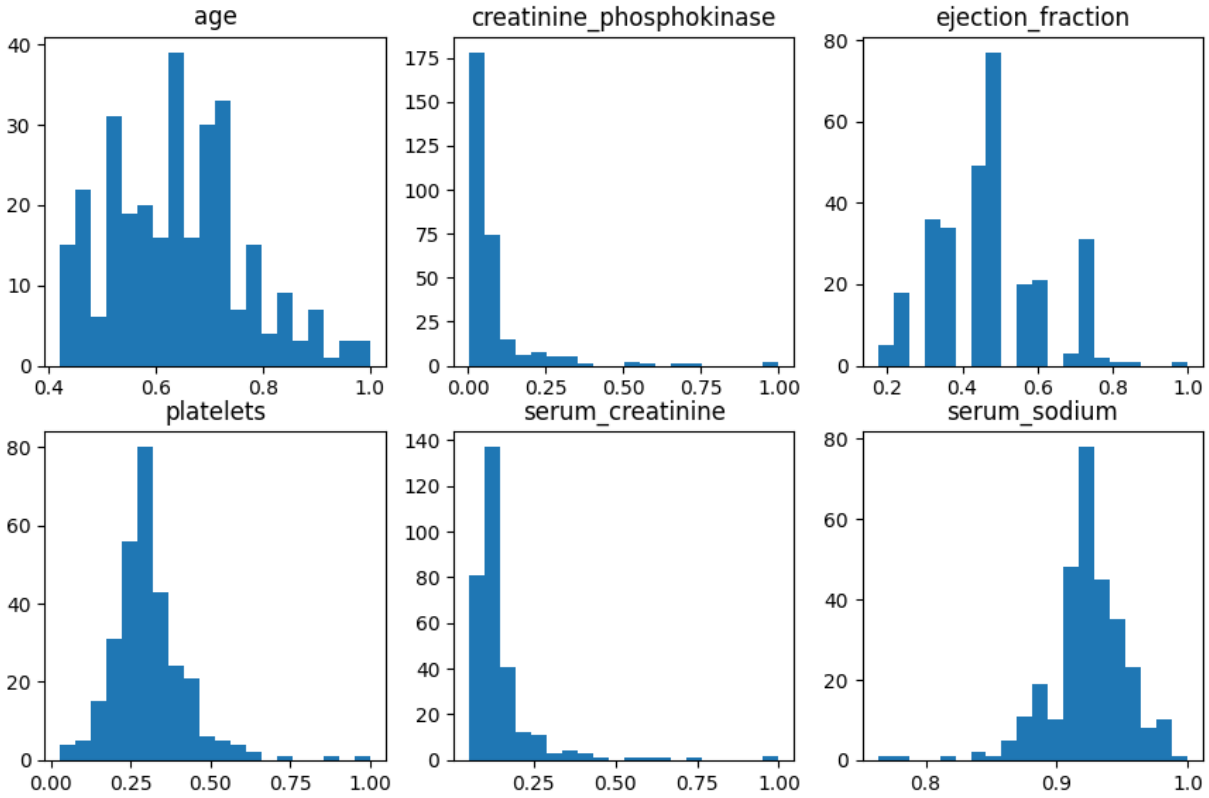


Теперь данные распределены на промежутке от 0 до 1.

2. Через параметры MinMaxScaler определите минимальное и максимальное значение в данных для каждого признака.

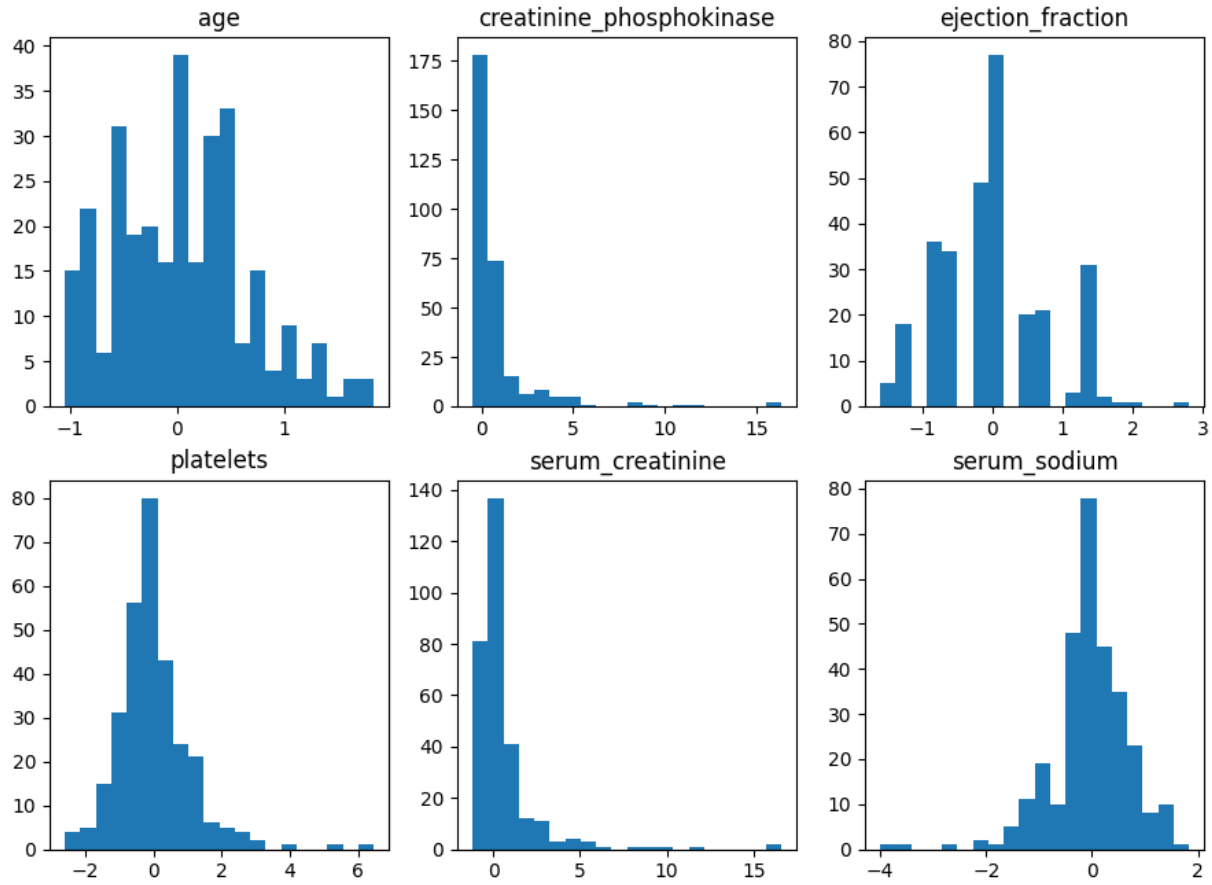
Признак	min	max
age	4.00e+01	9.500e+01
creatinine_phosphokinase	2.30e+01	7.861e+03
ejection_fraction	1.40e+01	8.000e+01
platelets	2.51e+04	8.500e+05
serum_crearinine	5.00e-01	9.400e+00
serum_sodium	1.13e+02	1.480e+02

3. MaxAbsScaler



Масштабирует данные так, что максимум будет единицей.

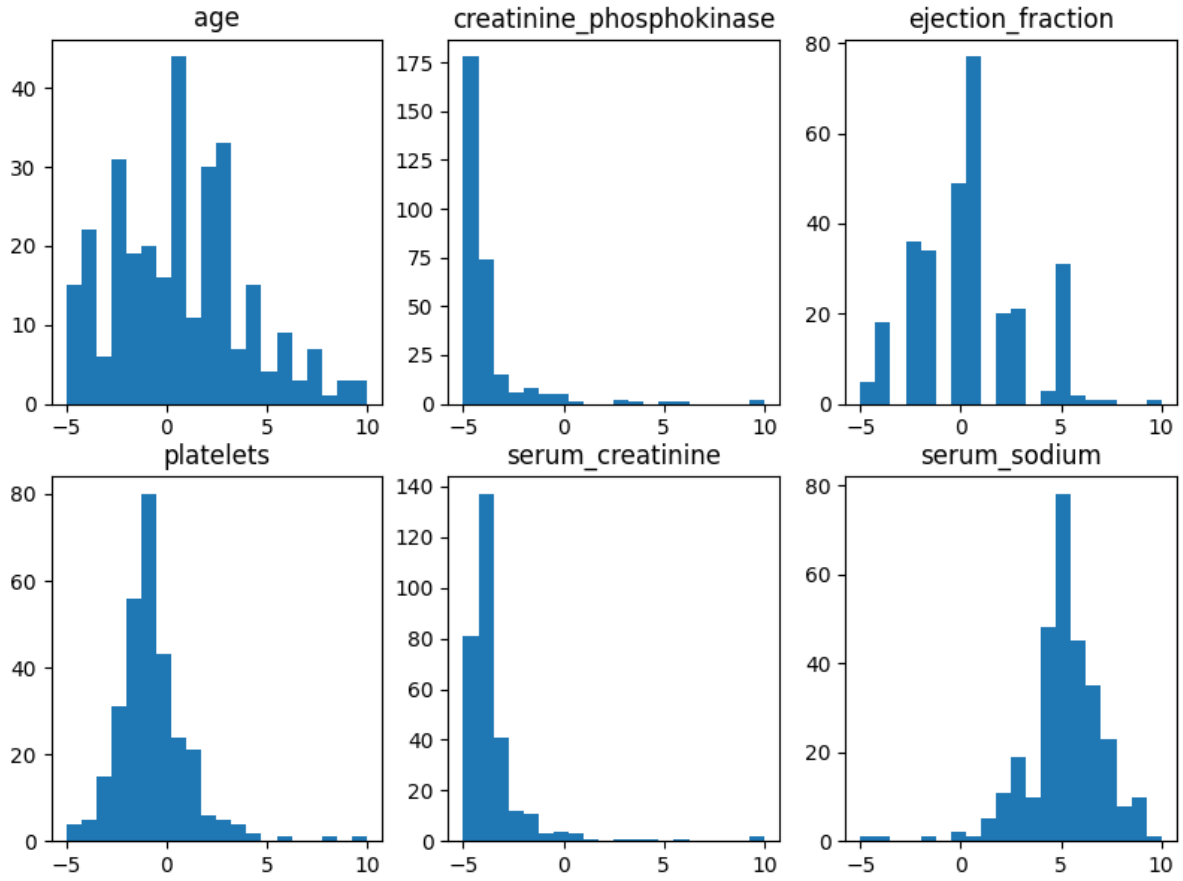
4. RobustScaler



Рассчитывается по формуле:
Значение = (Значение – Медианное значение) / (75-й перцентиль – 25-й)

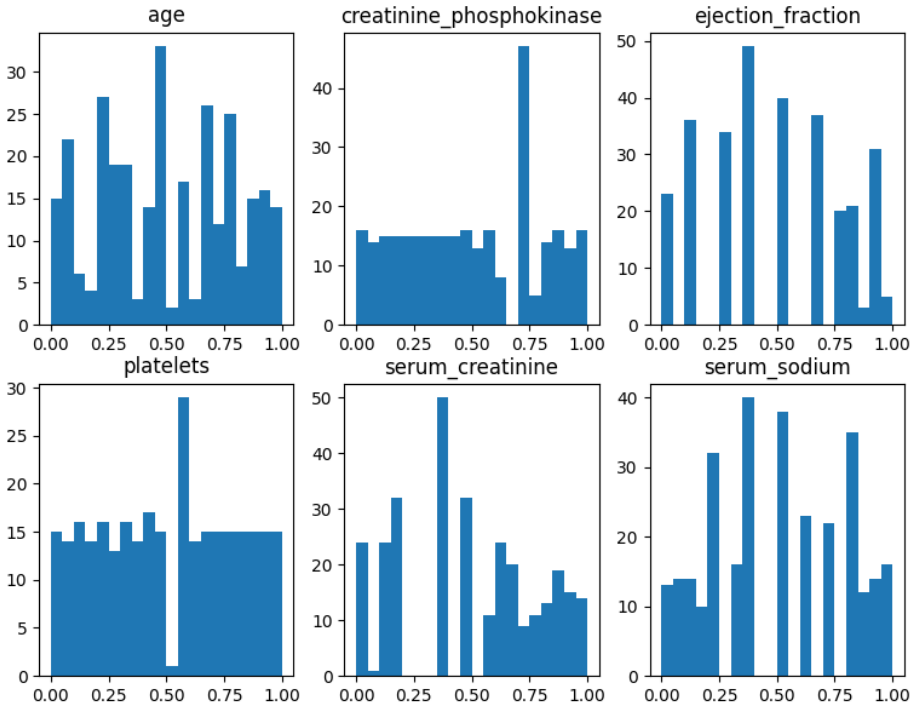
5. Напишите функцию, которая приводит все данные к диапазону [-5 10]

```
def scalefromMinus5to10(data):
  scaledData = preprocessing.MinMaxScaler(feature_range=(-5, 10)).fit_transform(data)
  return scaledData
```

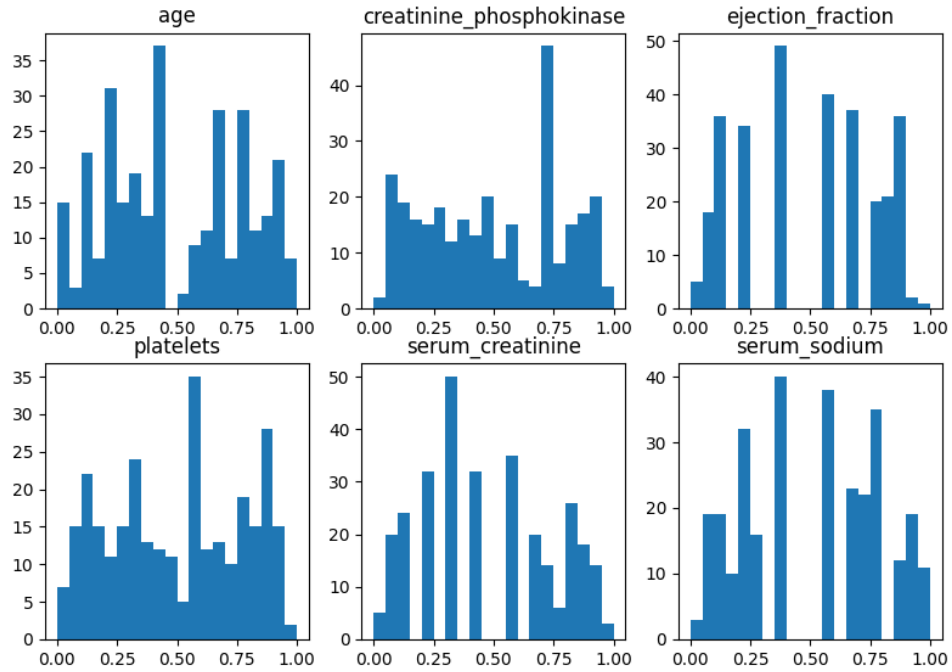


Нелинейные преобразования

1. Приведите данные к равномерному распределению используя QuantileTransformer. Постройте гистограммы и сравните с исходными данными

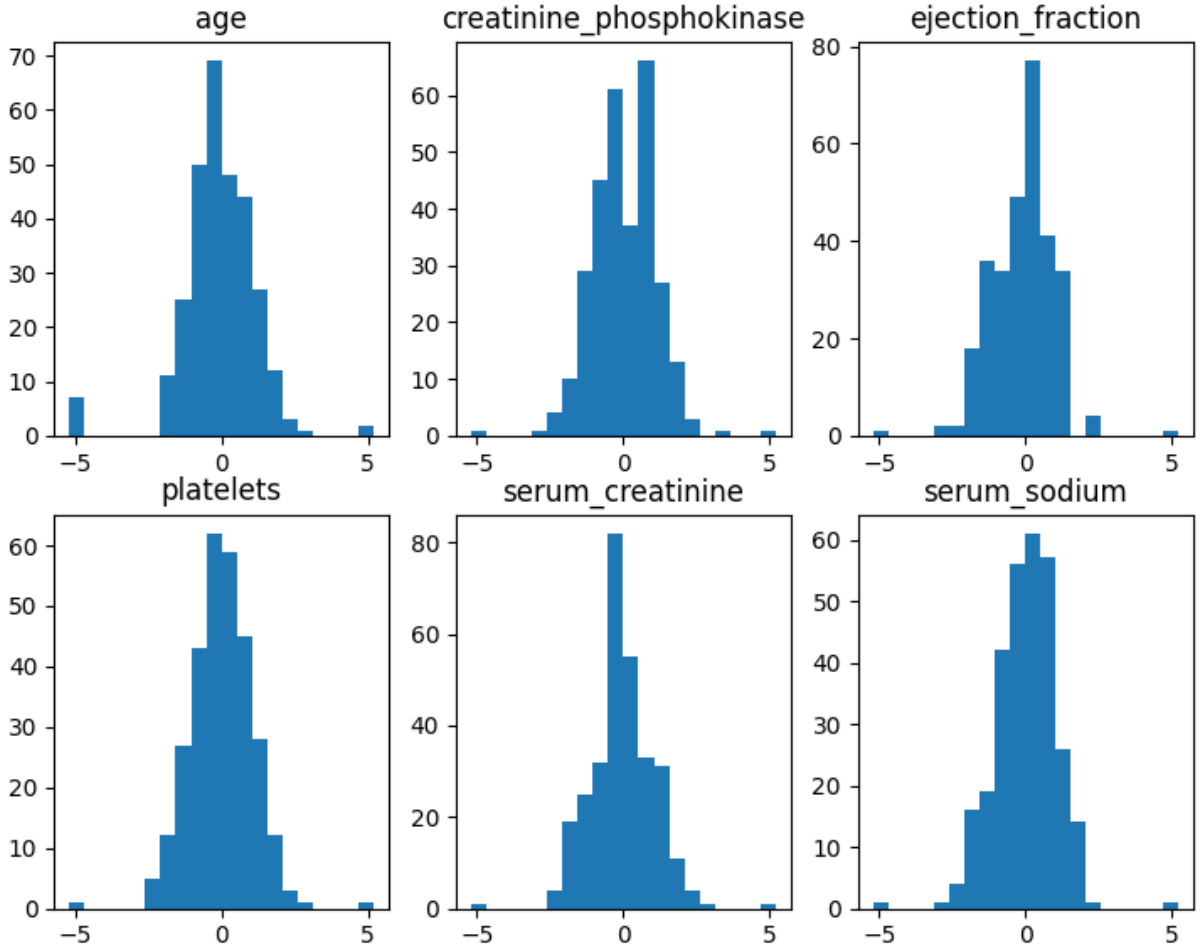


2. Определите, как и на что влияет значение параметра n_quantiles При 10:

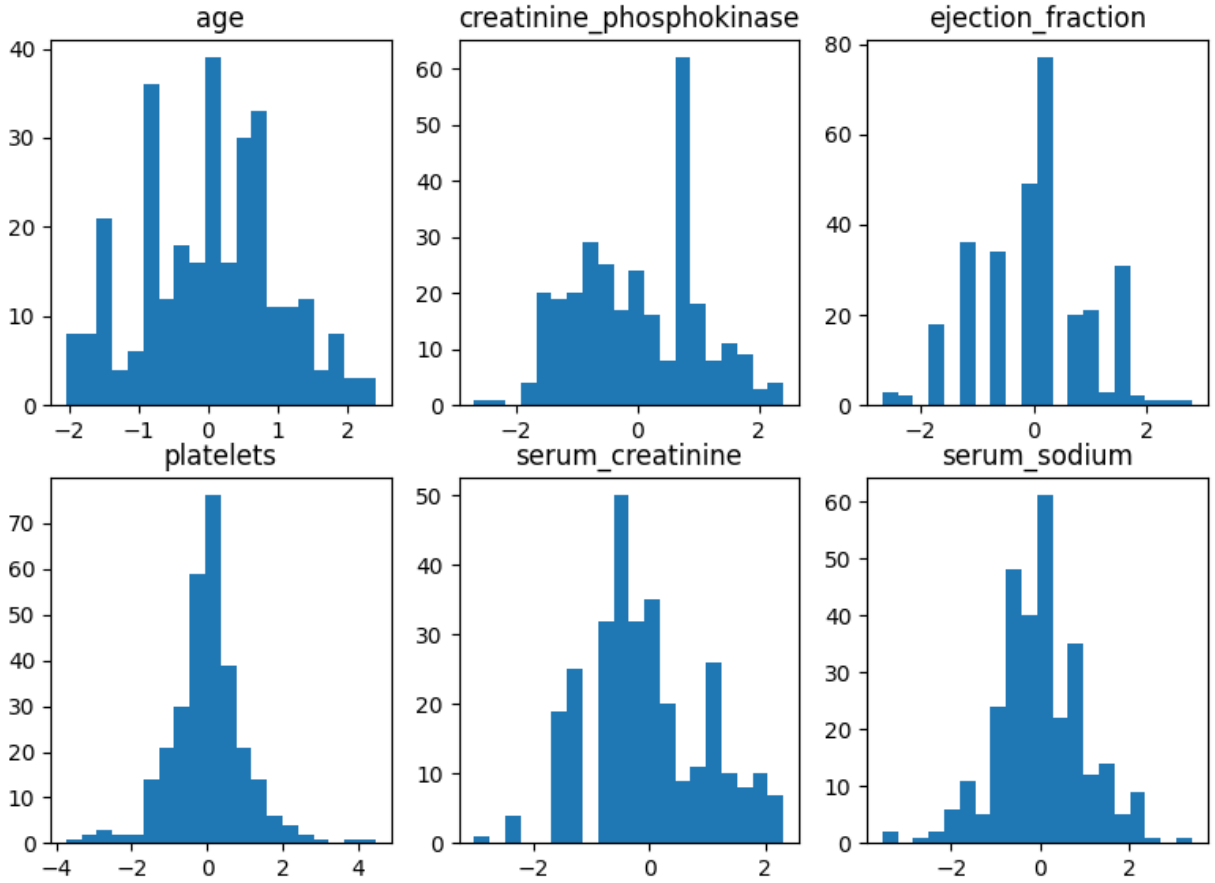


Число вычисляемых квантилей. Чем больше число, тем точнее апроксимация.

3. Приведите данные к нормальному распределению передав в QuantileTransformer параметр output_distribution='normal'

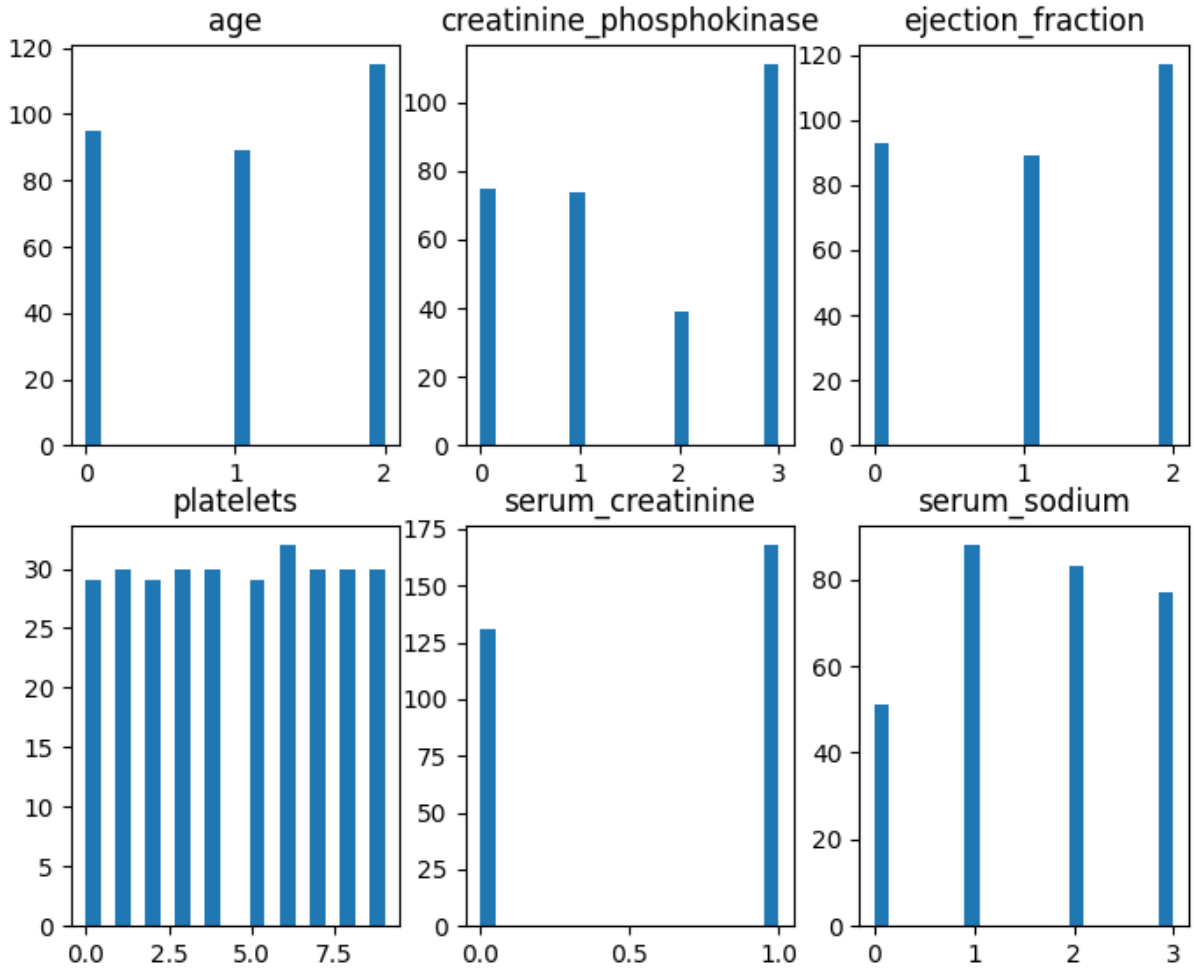


4. Самостоятельно приведите данные к нормальному распределению используя PowerTransformer



Дискретизация признаков

1. Проведите дискретизацию признаков, используя KBinsDiscretizer, на следующее количество диапазонов: age - 3 creatinine_phosphokinase - 4 ejection_fraction - 3 platelets - 10 serum_creatinine - 2 serum_sodium - 4 Постройте гистограммы. Объясните полученные результаты.



Данная функция разбивает данные на указанное количество интервалов. Ось x – номер интервала.

2. Через параметр bin_edges_ выведите диапазоны каждого интервала для каждого признака

[array([40., 55., 65., 95.])
array([23. , 116.5, 250. , 582. , 7861.])
array([14., 35., 40., 80.])
array([25100., 153000., 196000., 221000., 237000., 262000., 265000.,
285200., 319800., 374600., 850000.])
array([0.5, 1.1, 9.4])
array([113., 134., 137., 140., 148.])]

Вывод.

В ходе работы была произведена предобработка данных с помощью библиотеки Scikit Learn различными методами. Были произведены: стандартизация данных, приведение к диапазону. Данные методы используются для приведения разнородных данных к единому формату, что может понадобиться для дальнейших вычислений. Также были произведены нелинейные преобразования и дискретизация признаков.