МИНОБРНАУКИ РОССИИ

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ

ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)

Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №1

по дисциплине «Машинное обучение»

Тема: Предобработка данных

Студент гр. 6307 _____ Ходос А.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами предобработки данных из библиотеки Scikit Learn.

Ход работы

1. Загрузка данных

Загруженный датасет, состоящий из 299 наблюдения и 6 признаков, представле в таблице 1.

Таблица 1 — Загруженный датасет

	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 × 6 columns

Гистограммы признаков представлены на рисунке 1. На их основании можем получить примерный диапазон значений для каждого из признаков, а также возле какого значения лежит наибольшее количество наблюдений (сделаем это в пункте 3).

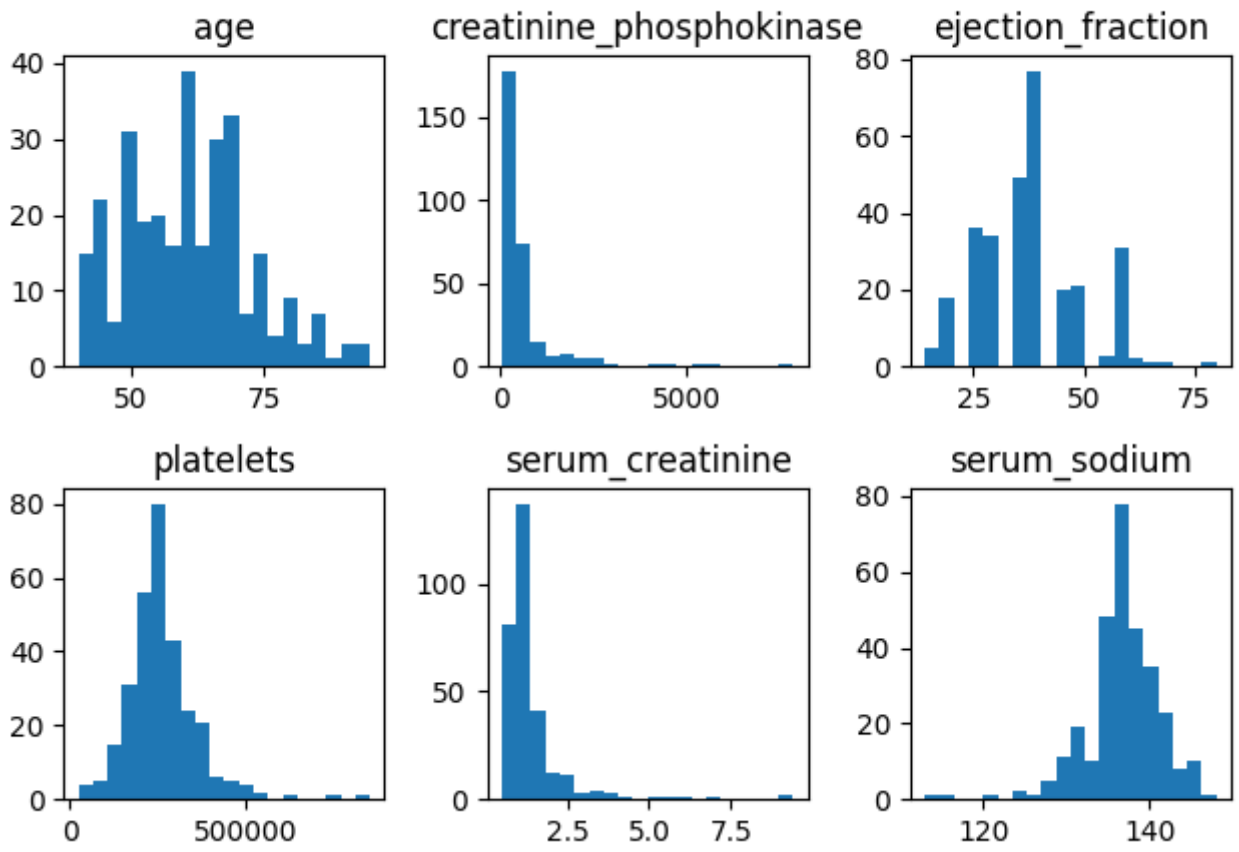

Рисунок 1 — Гистограммы признаков

2. Стандартизация

На рисунке 2 представлены данные после стандартизации с помощью StandardScaler. Как видно, шкала X была приведена к новой шкале (далее будет описано по какому принципу).

Рассчитанное мат. ожидание *(np.mean(data, axis=0))*:

[6.08338930e+01 5.81839465e+02 3.80836120e+01 2.63358029e+05 1.39387960e+00 1.36625418e+02]

Значение *scaler.mean_*:

[6.08338930e+01 5.81839465e+02 3.80836120e+01 2.63358029e+05 1.39387960e+00 1.36625418e+02]

Расчитанное СКО *(np.std(data, axis=0))*:

[1.18749014e+01 9.68663967e+02 1.18150335e+01, 9.76405477e+04 1.03277867e+00 4.40509238e+00]

Значение *scaler.var_(1/2)*:**

[1.24497854e+01 1.18974318e+03 1.30393183e+01 9.61917902e+04 1.16641630e+00 4.53958393e+00]

Рассчитанные значения мат. ожидания и дисперсии совпадают с полями mean_ и var_ объекта scaler.

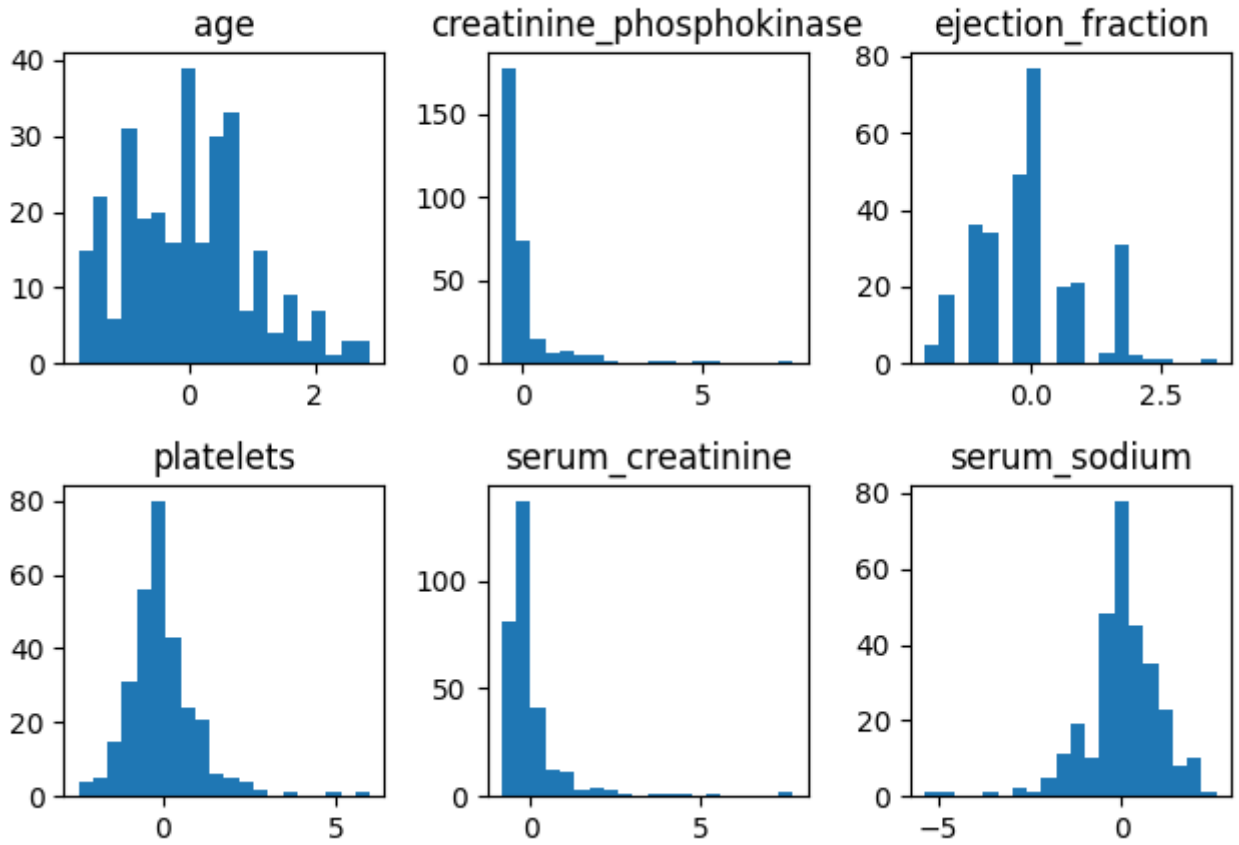

Рисунок 2 — Данные после стандартизации

Получим формулы для каждого признака, по которым производилась стандартизация:

zi = (xi — M) / o

1) zi = (xi - 60.83389297658862) / 11.874901429842655

2) zi = (xi - 581.8394648829432) / 968.6639668032415

3) zi = (xi - 38.08361204013378) / 11.815033462318585

4) zi = (xi - 263358.02926421404) / 97640.54765451424

5) zi = (xi - 1.3938795986622072) / 1.0327786652795918

6) zi = (xi - 136.62541806020067) / 4.405092379513557

Мат. ожидание после стандартизации: [5.70335306e-16 0.00000000e+00 -3.26754603e-17 7.72329061e-17 1.42583827e-16 -8.67384945e-16]

СКО после стандартизации: [1. 1. 1. 1. 1. 1.]

Видно, что данные после стандартизации имеют мат. ожидание 0, а СКО 1.

Мат. ожидание первых 150 элементов: [6.29466667e+01, 6.07153333e+02, 3.79466667e+01, 2.66746749e+05, 1.52060000e+00, 1.36453333e+02]

СКО первых 150 элементов: [1.24497854e+01, 1.18974318e+03, 1.30393183e+01, 9.61917902e+04, 1.16641630e+00, 4.53958393e+00]

3. Приведение к диапазону

Приведение к диапазону позволяет привести шкалу X к новой. Данные, приведенные к диапазону [0, 1], представлены на рисунке 3.

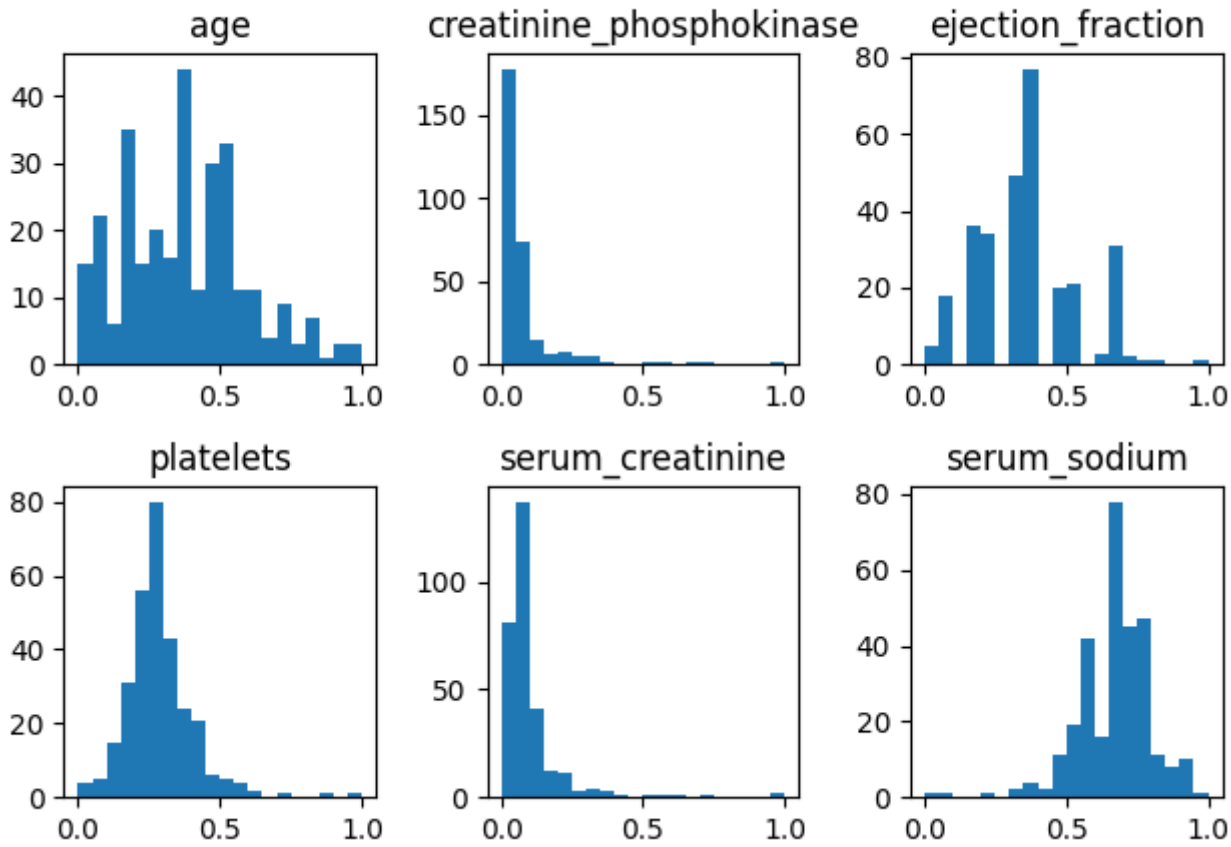

Рисунок 3 — Данные, приведенные к диапазону [0, 1]

Минимальные значения каждого признака *(scaler.data_min_)*: [4.00e+01 2.30e+01 1.40e+01 2.51e+04 5.00e-01 1.13e+02]

Максимальные значения каждого признака *(scaler.data_max_)*: [9.500e+01 7.861e+03 8.000e+01 8.500e+05 9.400e+00 1.480e+02]

Аналогично приведем данных к масштабируемому по максимальному абсолютному значению с помощью MaxAbsScaler (переводит максимальное абсолютное значение в 1) и к надежному (центрирует по медиане и масштабирует данные относительно межквартильного диапазона) диапазонам с помощью RobustScaler, гистограммы представлены на рисунках 4 и 5, соответственно.

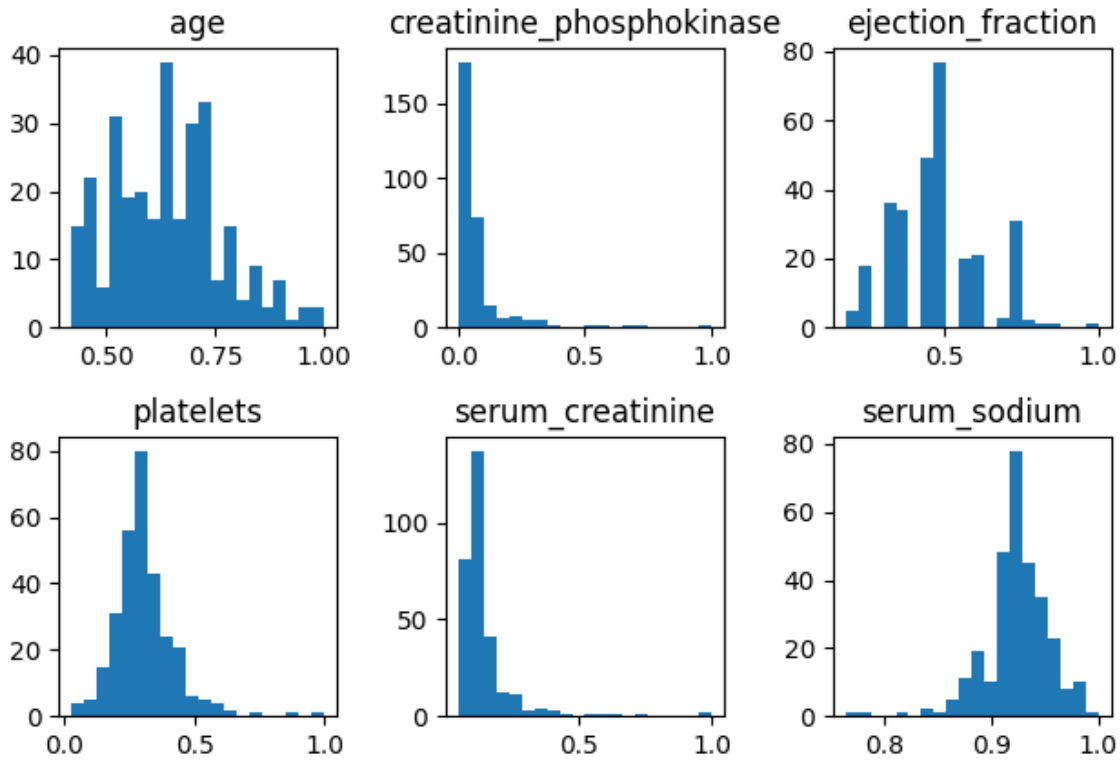

Рисунок 4 — Данные, приведенные к диапазону, масштабируемому по максимальному абсолютному значению

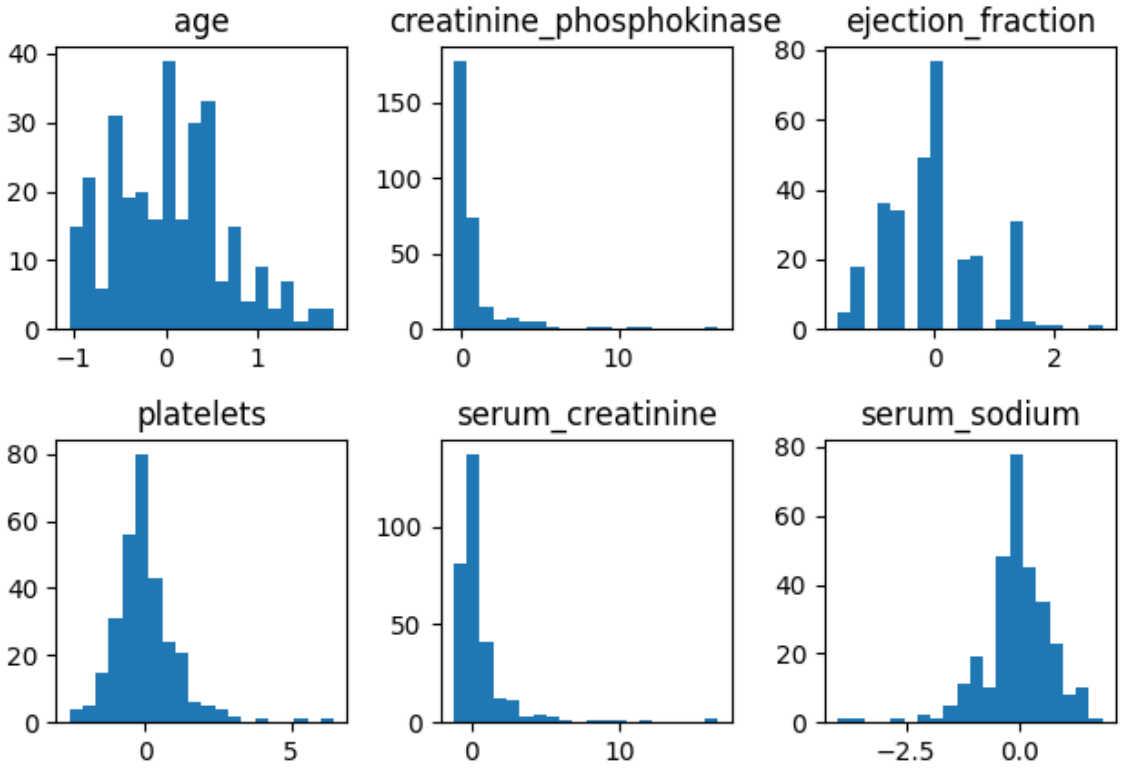

Рисунок 5 — Данные приведенные к надежному диапазону

Функция, приводящая данные к диапазону [-5, 10]:

$$zi = (xi — xmin) / (xmax — xmin) * (10 + 5) - 5$$

$$zi = 15 * (xi — xmin) / (xmax - xmin) - 5$$

Гистограммы данных, приведенных к диапазону [-5, 10], полученных при помощи MinMaxScaler приведены на рисунке 6.

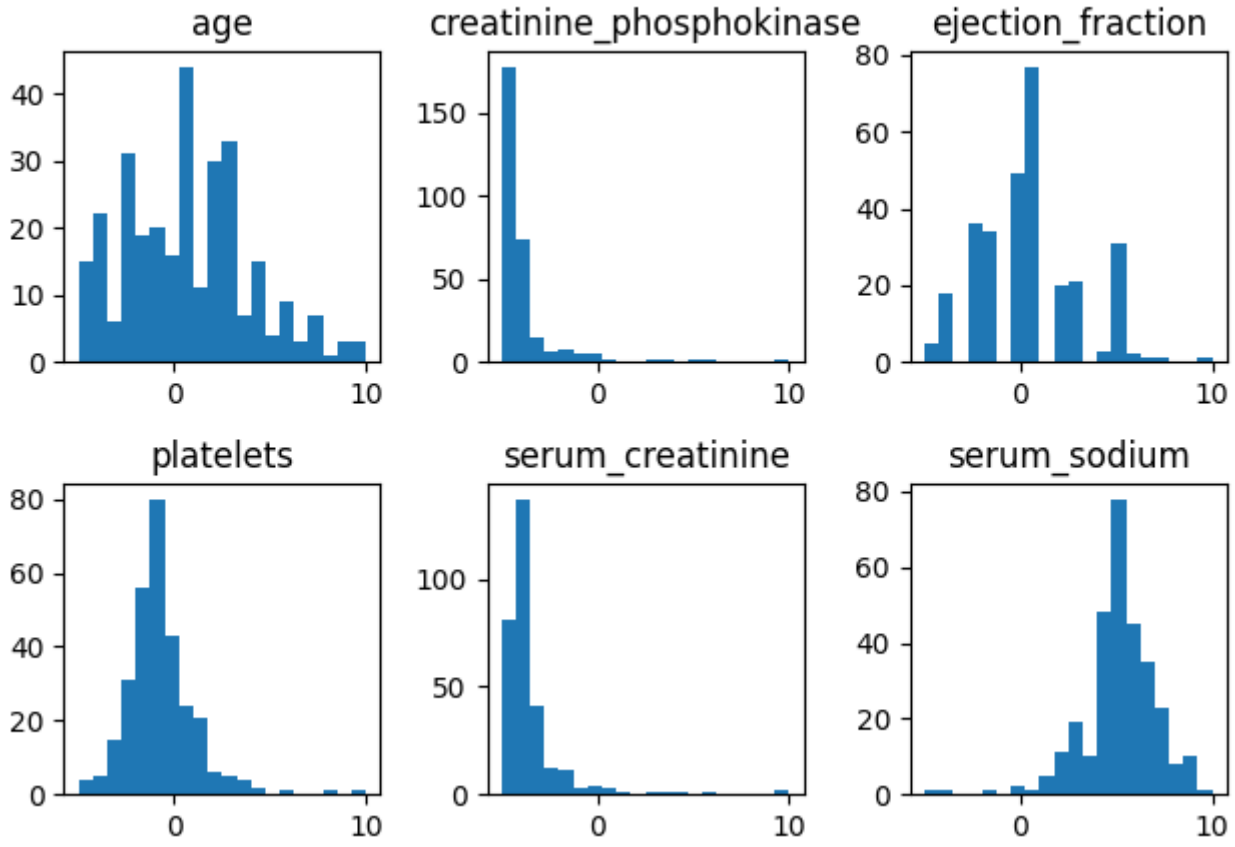

Рисунок 6 — Данные, приведенные к диапазону [-5, 10]

4. Нелинейные преобразования

Данные, приведенные к равномерному распределению в диапазоне [0, 1] с помощью QuantileTransformer, представлены на рисунке 7. После преобразования видно, что плотность вероятности стала более равномерной.

Чем выше число квантилей n_quantiles, тем лучше приближение к равномерному распределению.

Данные, преобразованные к нормальному распределению с помощью QuantileTransformer и PowerTransformer представлены на рисунках 8 и 9, соответственно.

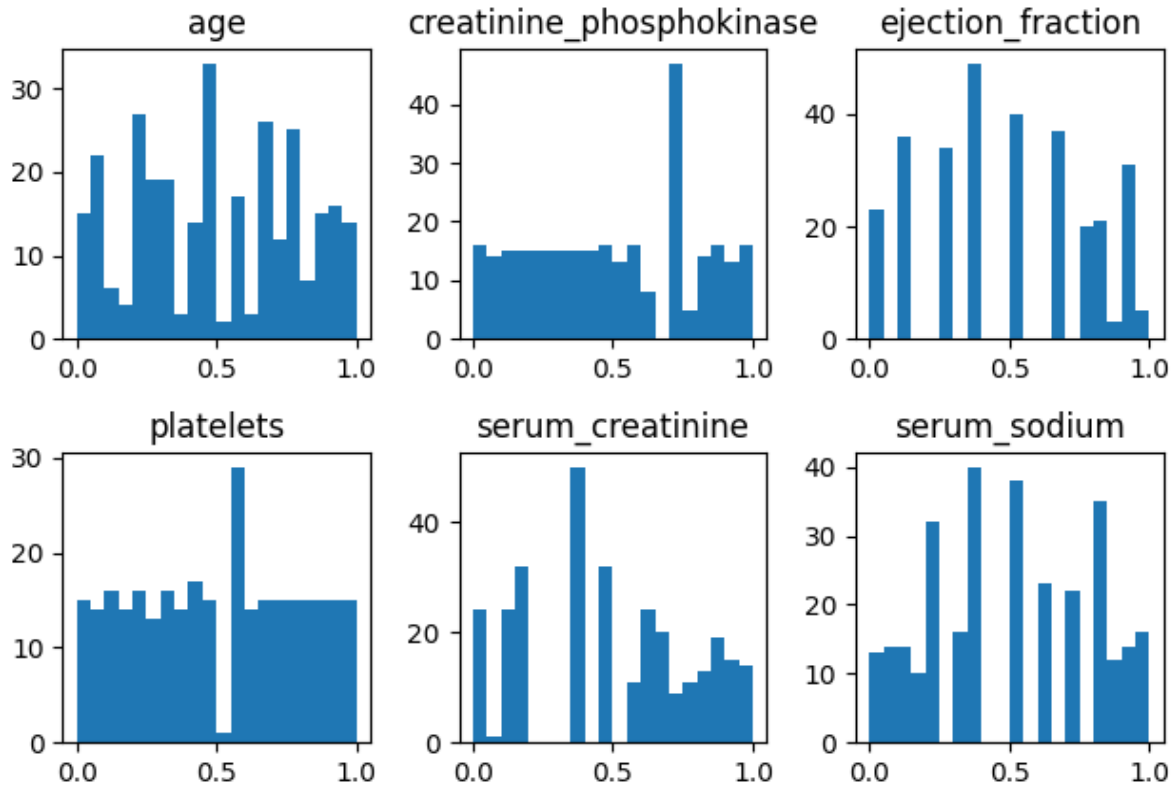

Рисунок 7 - данные, приведенные к равномерному распределению в диапазоне [0, 1] с помощью QuantileTransformer

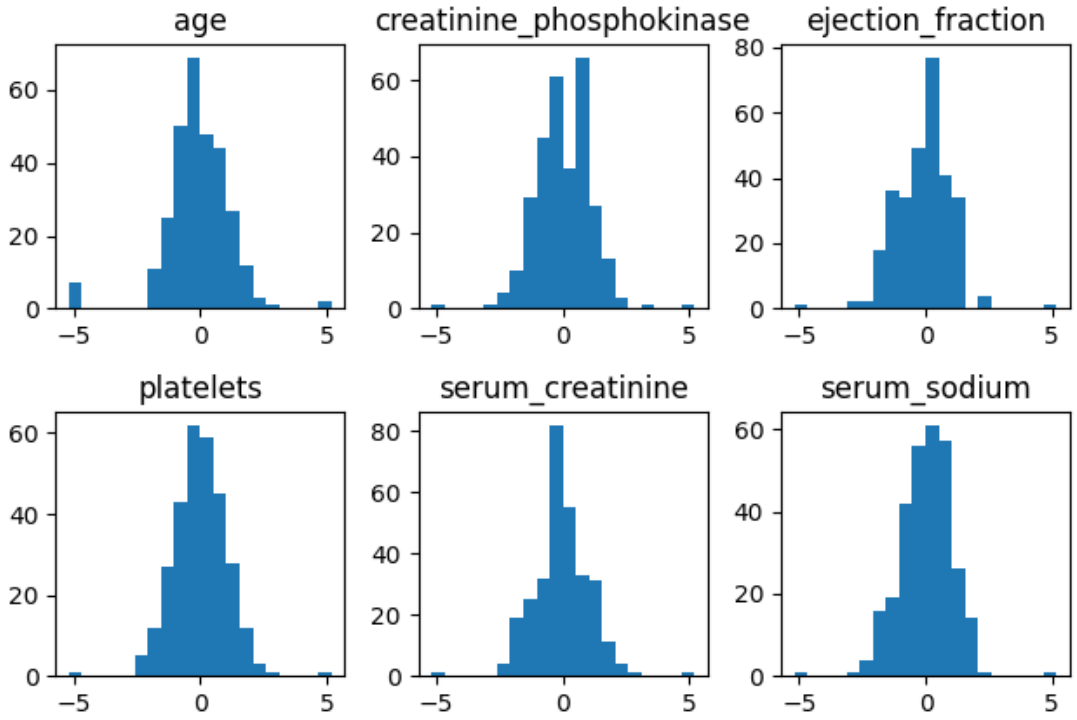

Рисунок 8 — Данные, преобразованные к нормальному распределению с помощью QuantileTransformer

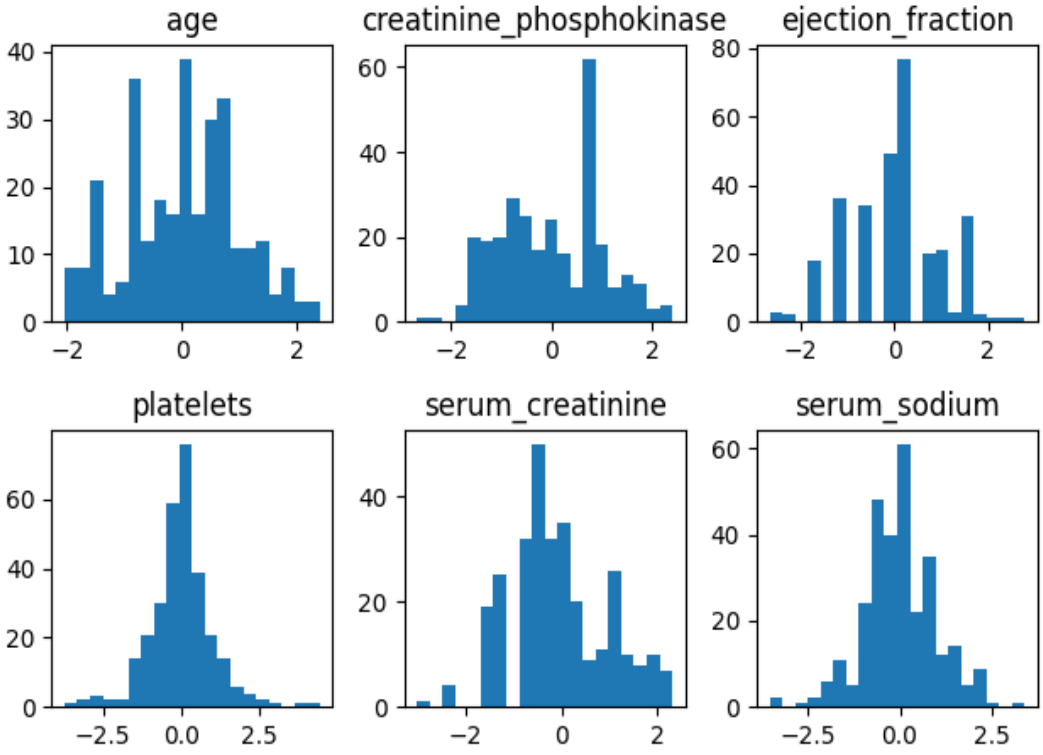

Рисунок 9 — Данные, преобразованные к нормальному распределению с помощью PowerTransformer

5. Дискретизация признаков

Дискретизированные данные с заданным количеством диапазонов представлена на рисунке 10.

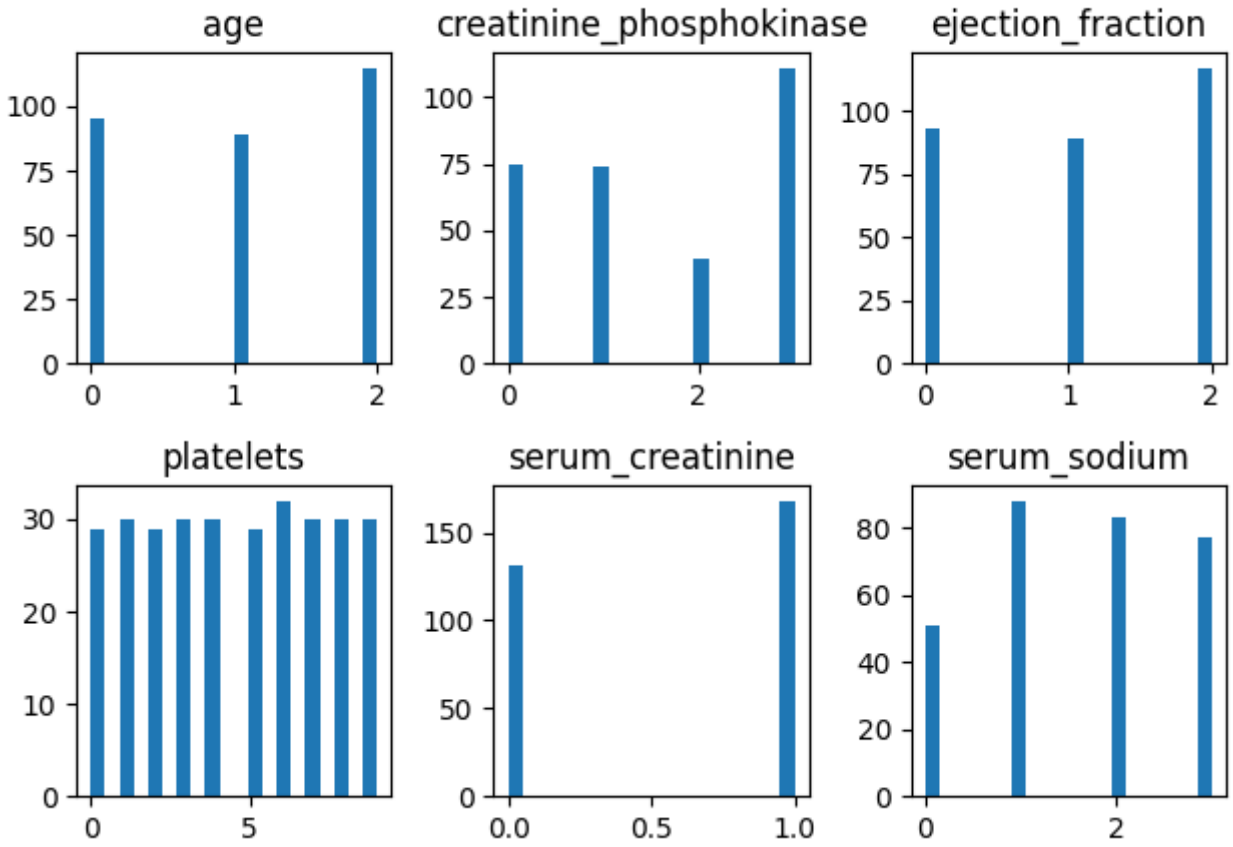

Рисунок 10 — Дискретизированные данные с заданным количеством диапазонов

Значения полученных диапазонов (параметр descritizer.bin_edges_):

[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.])]

Выводы

Были получены навыки работы с методами предобработки данных из библиотеки Scikit Learn, позволяющих выполнить стандартизацию, приведение к диапазонам, нелинейные преобразования и дискретизацию данных.