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

ОТЧЁТ

по лабораторной работе №1
по дисциплине «Машинное обучение»
Тема: Предобработка

Студент гр. 6304 _____ Корытов П.В.
Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург
2020

1. Цель работы

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

2. Ход работы

2.1. Загрузка данных

1. Загружен указанный набор данных, проведены преобразования данных (рис. 1)

	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

Рисунок 1 – Вывод набора данных

2. Построены гистограммы для выделенных признаков (рис. 2)

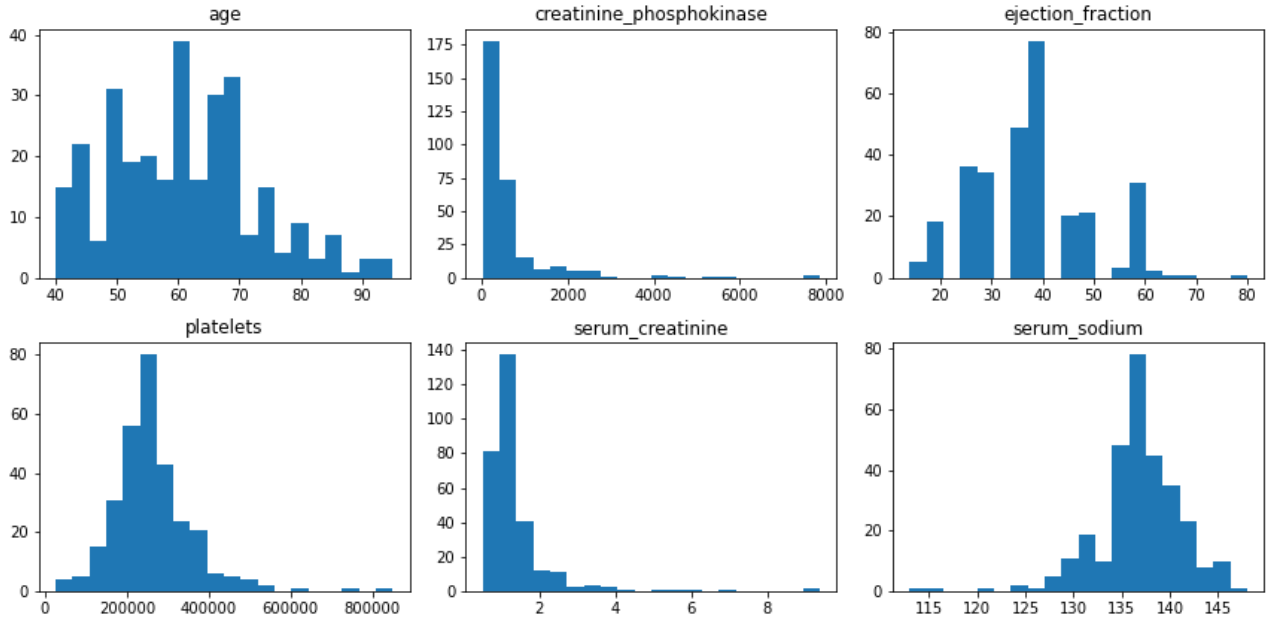

Рисунок 2 – Гистограммы призаков

2.2. Стандартизация данных

1. Проведена нормализация данных с помощью `StandardScaler` на основе первых 150 наблюдений. Гистограммы стандартизованных данных представлены на рис. 3.

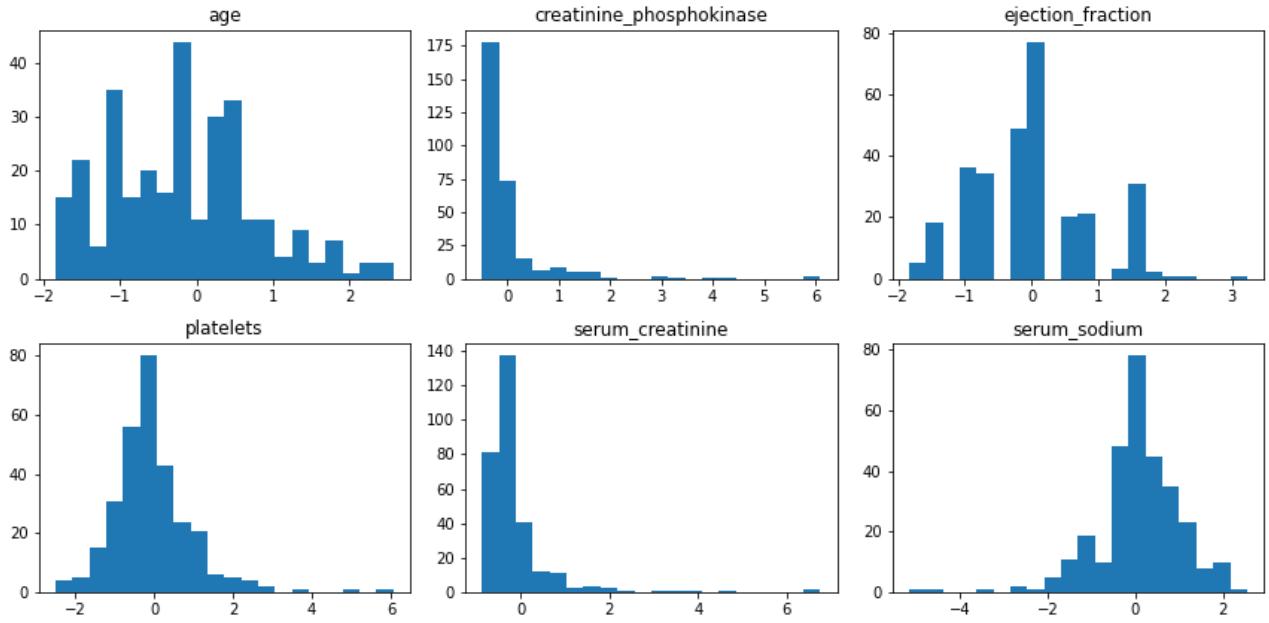

Рисунок 3 – Гистограммы нормализованных признаков (на осн. первых 150)

2. Проведена та же процедура на всем датасете. Результаты на рис. 4

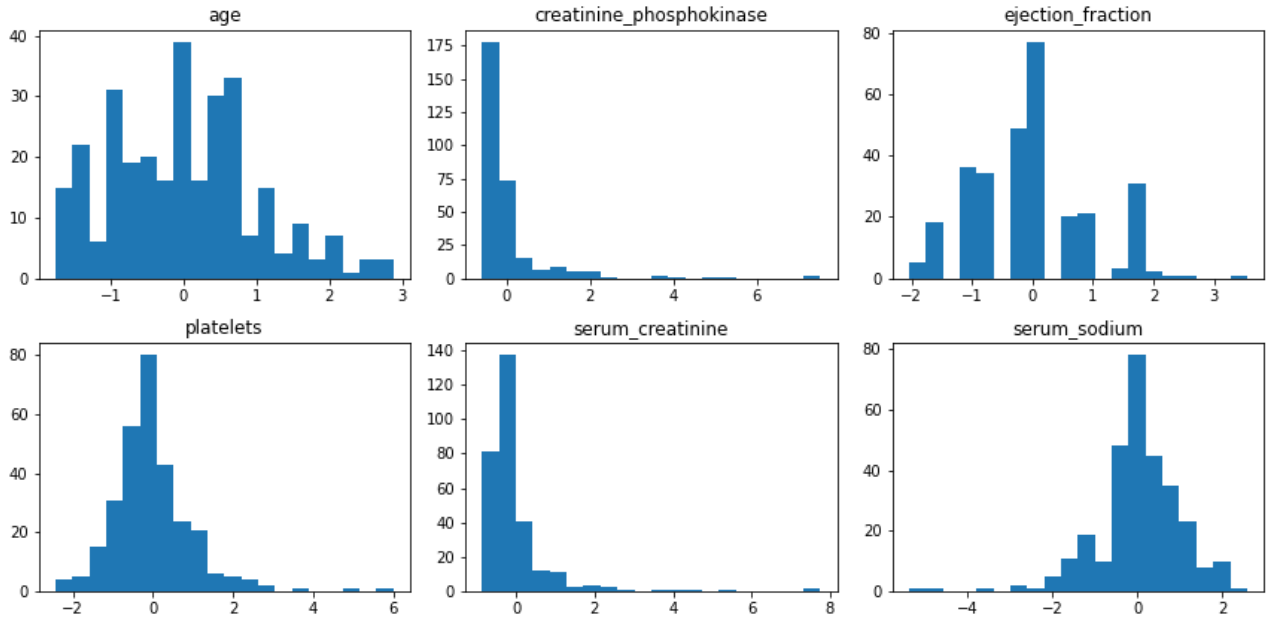

Рисунок 4 – Гистограммы нормализованных признаков

4. Вычислено мат. ожидание и СКО для исходных данных и обоих порций нормализованных данных. Результаты в таблице 1.

Таблица 1. Признаки

Признак	age	creatinine...	ejection_f...	platelets	serum_crea...	serum_sodi...
Среднее (исх.)	60.8339	581.8395	38.0836	263358.0293	1.3939	136.6254
Среднее (стандарт. 150)	-0.1697	-0.0213	0.0105	-0.0352	-0.1086	0.0379
Среднее (стандарт. 150 scaler)	62.9467	607.1533	37.9467	266746.7495	1.5206	136.4533
Среднее (стандарт. полн.)	0.0000	0.0000	-0.0000	0.0000	0.0000	-0.0000
Среднее (стандарт. полн. scaler)	60.8339	581.8395	38.0836	263358.0293	1.3939	136.6254
СКО (исх)	11.8749	968.6640	11.8150	97640.5477	1.0328	4.4051
СКО (стандарт. 150)	0.9538	0.8142	0.9061	1.0151	0.8854	0.9704
СКО (стандарт. 150 scaler)	12.4498	1189.7432	13.0393	96191.7902	1.1664	4.5396
СКО (стандарт. полн.)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
СКО (стандарт. полн. scaler)	11.8749	968.6640	11.8150	97640.5477	1.0328	4.4051

Судя по результатам в таблице 1 и рис. 4, `StandardScaler` центрирует данные относительно среднего и масштабирует относительно дисперсии. Предположительная формула:

$$Y_i = \frac{X_i - M[X]}{D[X]}, \tag{2.1}$$

где X — исходные данные, Y — преобразованные данные.

В объекте scaler записывается дисперсия и мат. ожидание исходных данных. При ограничении размера выборки для настройки результаты нормализуются менее качественно, т.е. мат. ожидание и СКО отличаются от 0 и 1.

2.3. Приведение к диапазону

1. Данные приведены к диапазону в помощью `MinMaxScaler`. Гистограммы приведены на рис. 5.

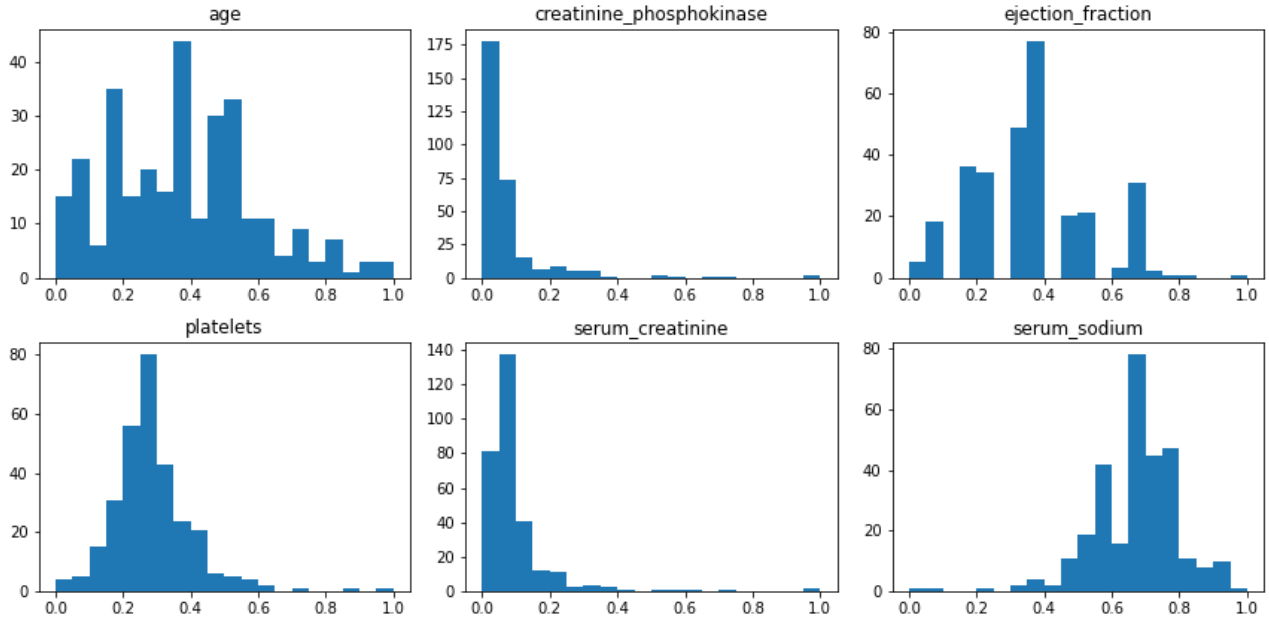

Рисунок 5 – Гистограммы данных после `MinMaxScaler`

Исходя из гистограмм, `MinMaxScaler` масштабирует данные к промежутку $[0, 1]$.

2. Из атрибутов scaler-а получены значения атрибутов. Результаты приведены в таблице 2.

Таблица 2. Минимальные и максимальные значения признаков

Признак	Минимум	Максимум
age	40	95
creatinine_phosphokinase	23	7861
ejection_fraction	14	80
platelets	25100	850000
serum_creatinine	0.5	9.4
serum_sodium	113	148

3. Проведено приведение с помощью `MaxAbsScaler` и `RobustScaler`. Гистограммы данных приведены на рис. 6 и рис. 7.

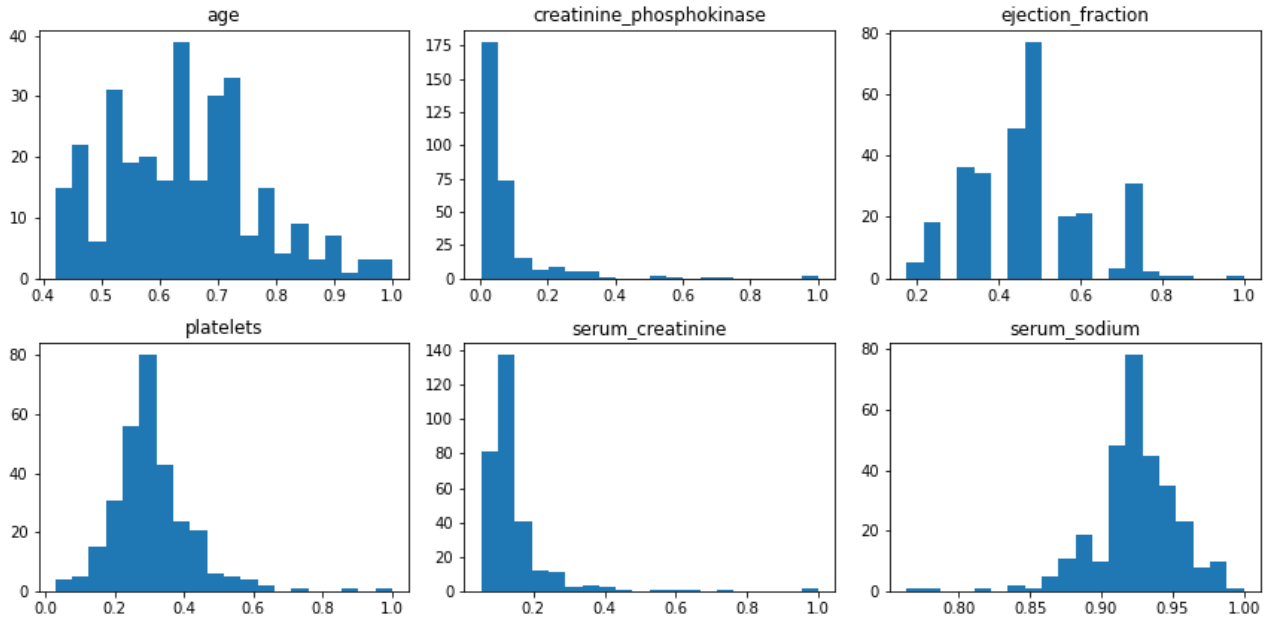

Рисунок 6 – Гистограммы данных после `MaxAbsScaler`

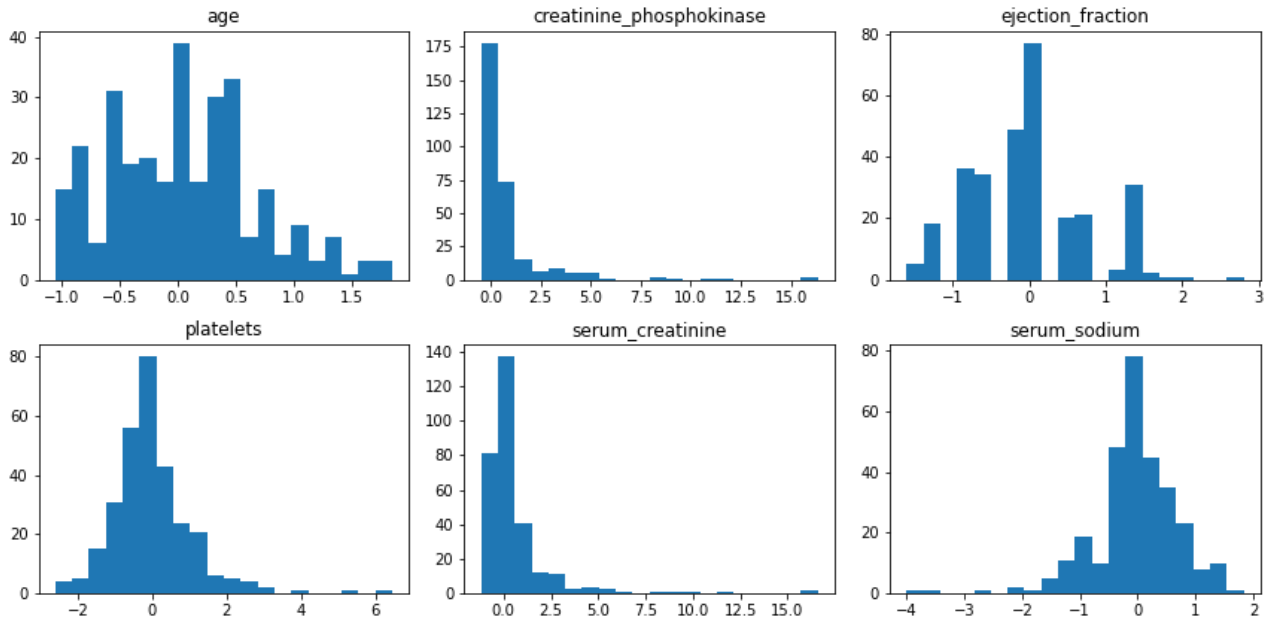

Рисунок 7 – Гистограммы данных после `RobustScaler`

`MaxAbsScaler` изменяет данные таким образом, чтобы максимальное значение по модулю было равно 1. `RobustScaler` центрирует по медиане и масштабирует данные относительно диапазона между 25-м и 75-м процентилем.

4. Написана функция для приведения данных к диапазону $[-5, 10]$. Результат приведен в листинге 1.

Листинг 1. Функция для приведения данных к диапазону $[-5, 10]$

```
1  def fit_5_10(data):
2      data = data.copy()
3      for col in range(data.shape[1]):
4          min_, max_ = np.min(data[:, col]), np.max(data[:, col])
5          data[:, col] = [(x - min_) / (max_ - min_) * 15 - 5 for x
           ↪  in data[:, col]]
6      return data
```

Гистограммы для этого преобразования представлены на рис. 8.

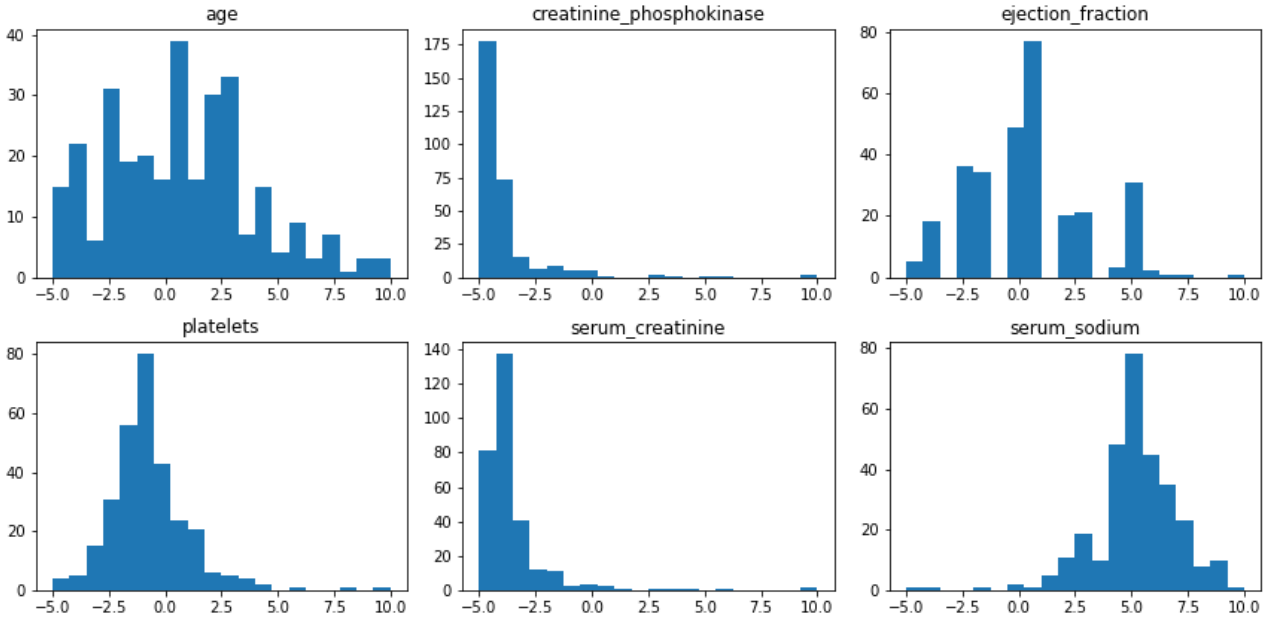

Рисунок 8 – Гистограммы после `fit_5_10`

2.4. Нелинейные преобразования

1. С помощью `QuantileTransform` данные приведены к равномерному и нормальному распределению. Результаты представлены на рис. 9, 10.

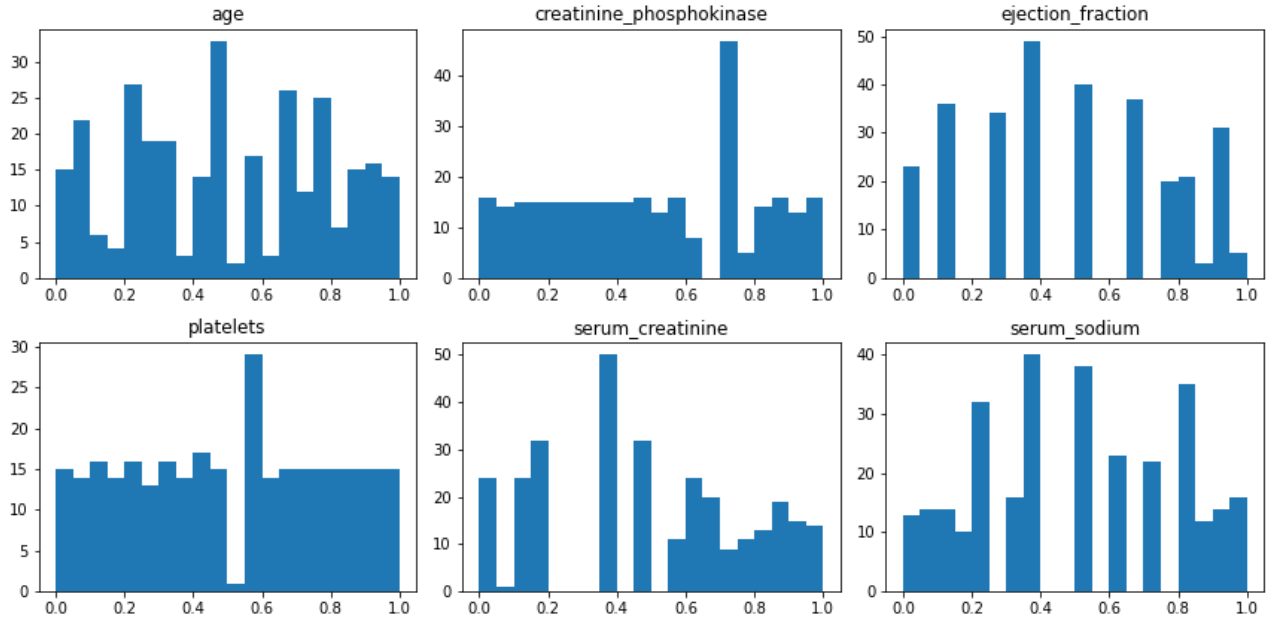

Рисунок 9 – Гистограммы после `QuantileTransform` с равномерным распределением

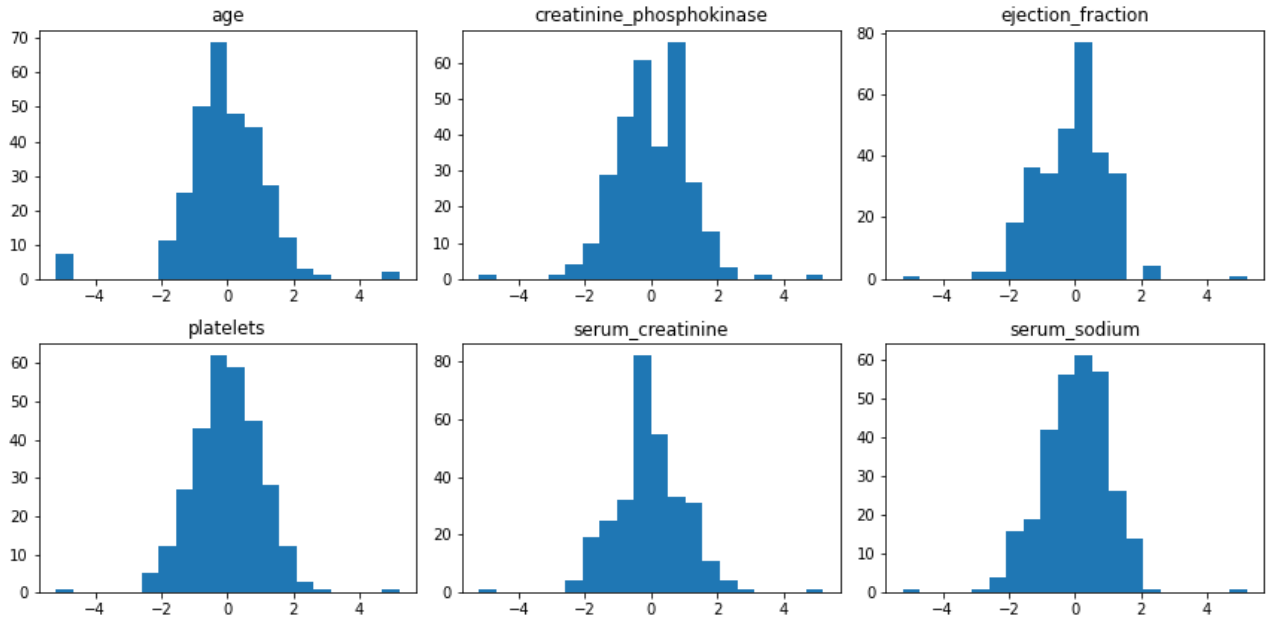

Рисунок 10 – Гистограммы после `QuantileTransform` с нормальным распределением

Параметр `n_quantiles` определяет количество вычисляемых процентилей в ходе настройки. Увеличение повышает частоту дискретизации функции распределения.

2. Данные приведены к нормальному распределению через `PowerTransformer`. Результаты на рис. 11.

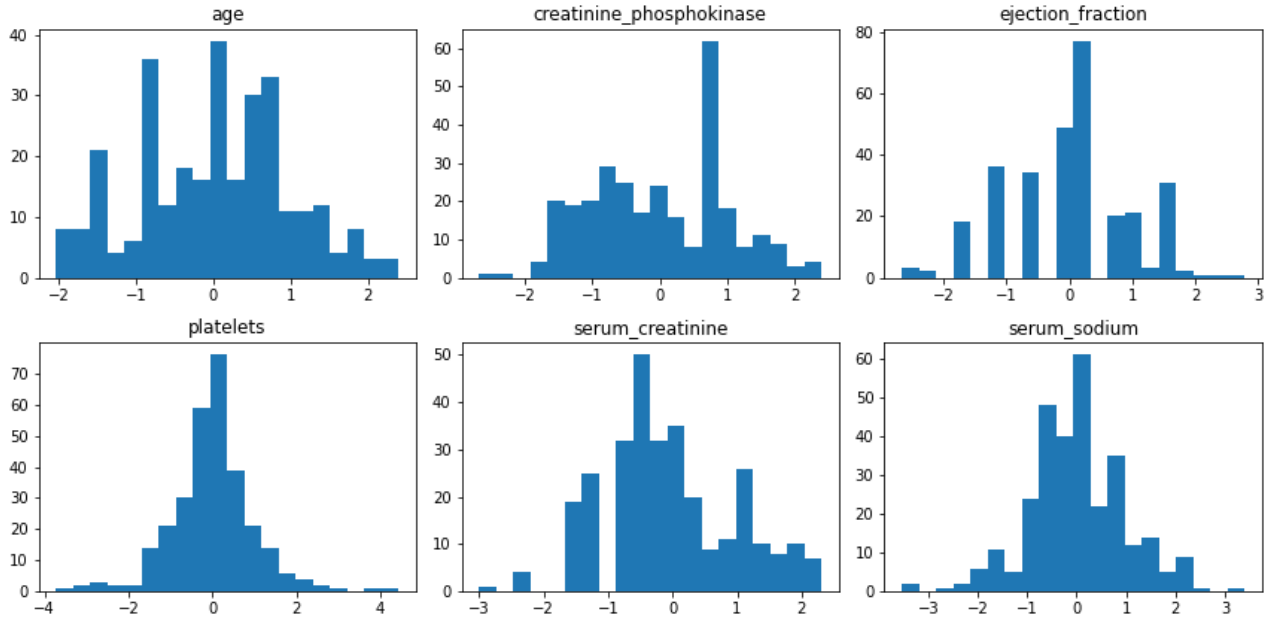

Рисунок 11 – Гистограммы после `PowerTransformer`

2.5. Дискретизация признаков

1. Проведена дискретизация признаков с помощью `KBinsDiscretizer`. Гистограммы дискретизованных данных приведены на рис 12.

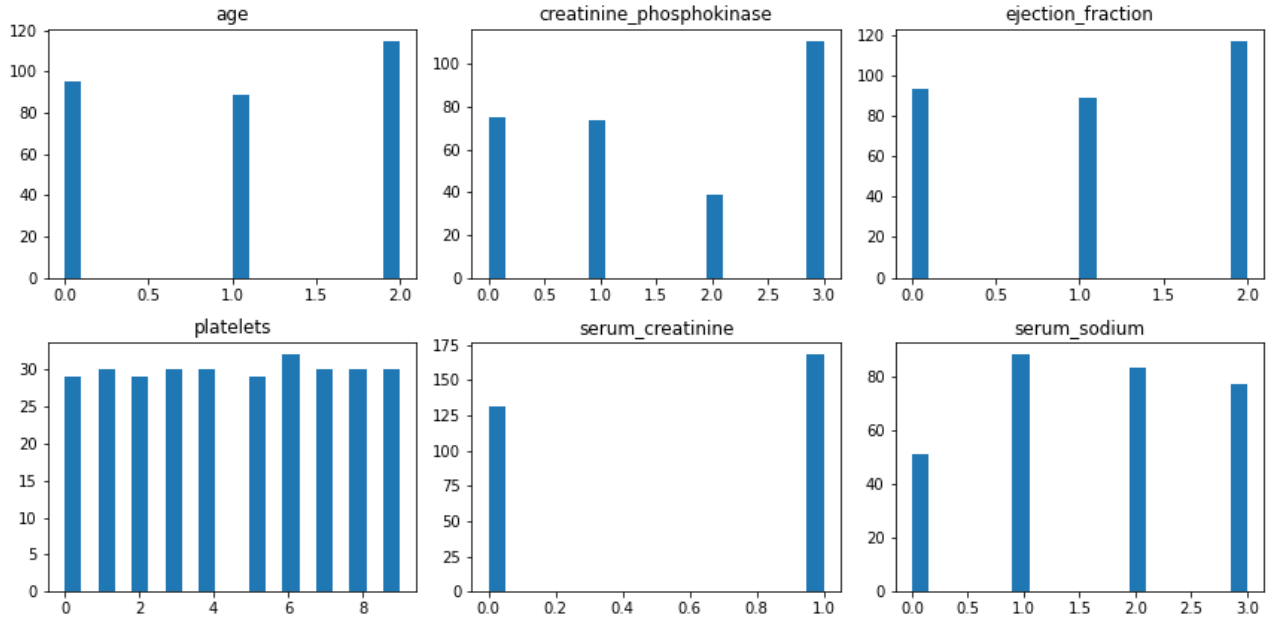

Рисунок 12 – Гистограммы после `KBinsDiscretizer`

Поскольку значения по оси ординат являются числовыми идентификаторами дискретных значений, построение гистограммы не имеет смысла.

3. Выводы

Произведено знакомство с методами предобработки данных библиотеки *Scikit Learn*.

Проведена стандартизация данных; установлено, что стандартизация с учетом неполного набора данных снижает качество выходных данных.

Проведено приведение данных к диапазону. Гистограммы данных, приведенных к диапазону, схожи с гистограммами стандартизованных данных.

Также проведены нелинейные преобразования данных. Предположительно, использование `QuantileTransform` может иметь смысл при наличии в данных выбросов; однако в этом случае искажается структура данных.

Также проведена дискретизация данных.