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. 6304 _____ Ястребков А. С.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург
2020

Цель работы:

Ознакомиться с методами предобработки данных из библиотеки Scikit Learn.

Ход работы

1. Загрузка данных. Загружен требуемый набор данных, csv-файл загружен в скрипт с помощью инстурментов модуля pandas. Из датасета удалены бинарные признаки. Фаргмент получившегося датасета приведён на 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
5	90.0	47	40	204000.00	2.1	132
6	75.0	246	15	127000.00	1.2	137

Рис. 1. Фрагмент исходного датасета.

Для датасета построены гистограммы, приведённые на рис. 2.

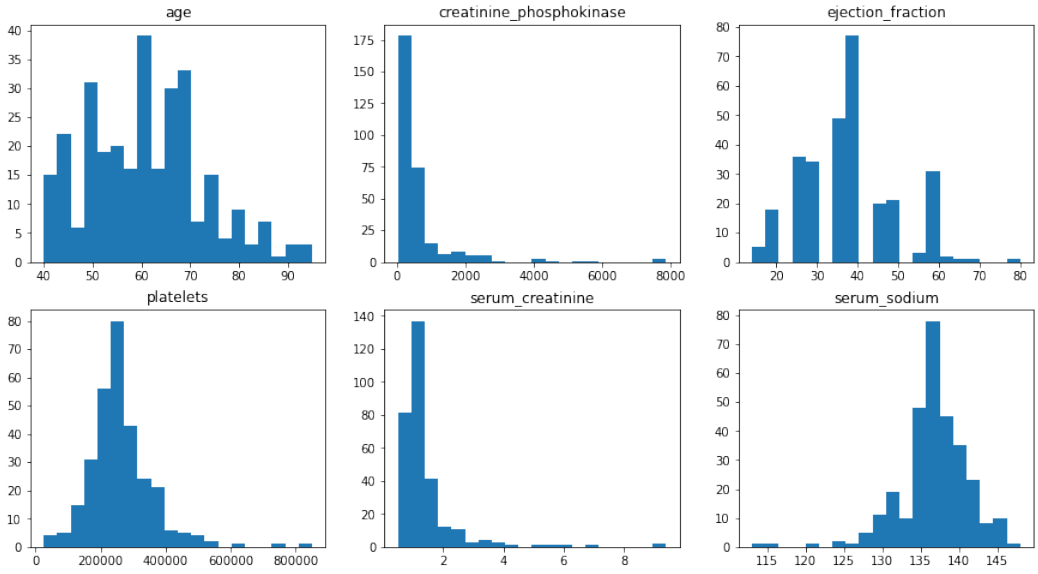

Рис. 2. Гистограммы признаков.

По гистограммам были приблизительно установлены диапазоны значений признаков и значения, которым принадлежит наибольшее количество наблюдений. Данные сведены в таблицу 1.

Табилца 1. Оценки диапазона и моды признаков.

признак	минимальное	максимальное	мода
age	40	92	61
creatinine_phosphokinase	0	7800	0-400
ejection_fraction	14	80	40
platelets	0	850000	250000
serum_creatinine	0	10	1.5
serum_sodium	114	143	137

2. Стандартизация данных. Для стандартизации используется StandardScaler модуля pandas. Стандартизация проводится на первых 150 наблюдений (рис. 3) и всей выборки (рис. 4).

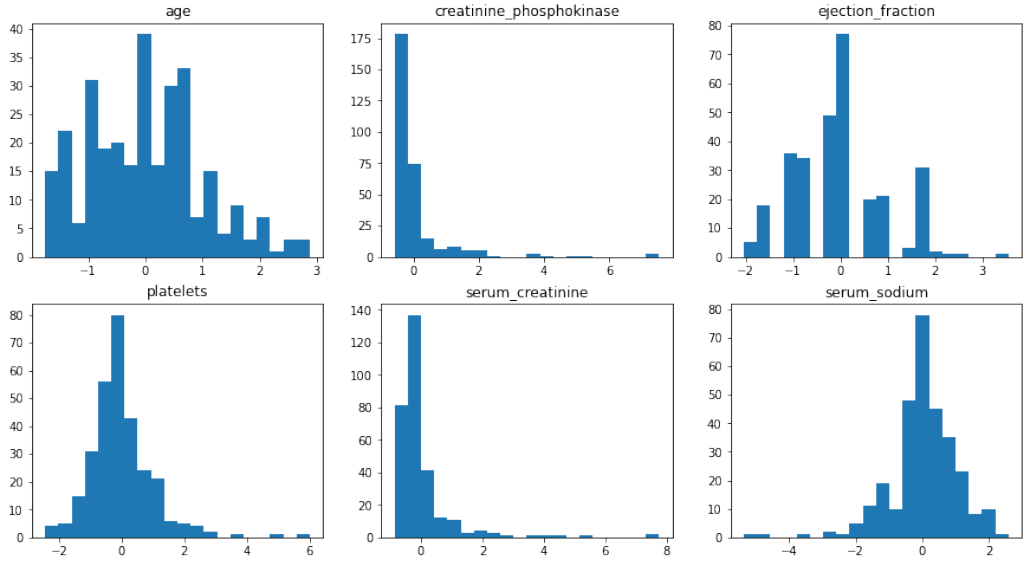

Рис. 3. Гистограмма данных после стандартизации по первым 150 наблюдениям.

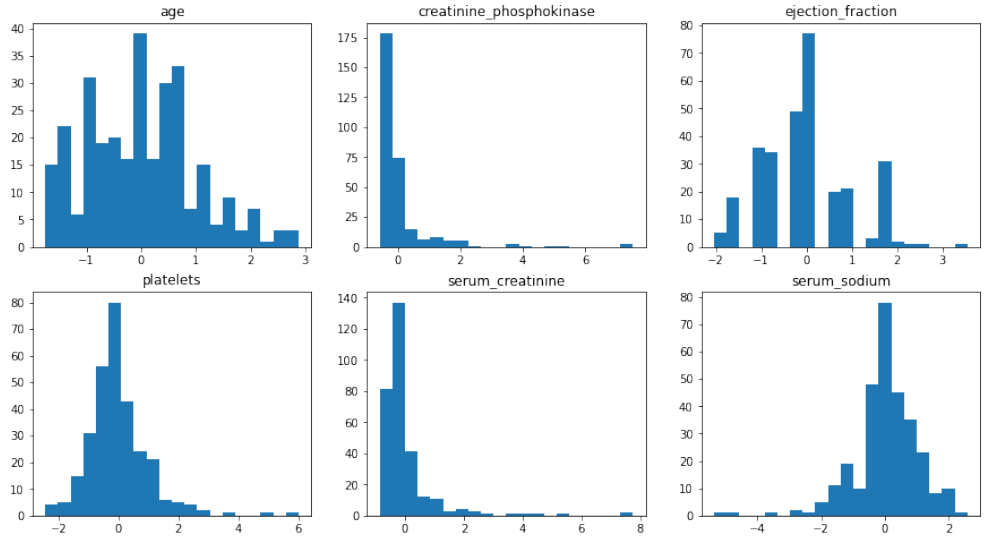

Рис. 4. Гистограмма после стандартизации по полной выборке.

В таблице 2 сведены данные по стандартному отклонению и матожиданию признаков для исходных и стандартизированных данных.

Таблица 2. Матожидание и стандартное отклонение признаков до и после стандартизации.

признак	M_{before}	*scaler.mean_*	M_{150}	M_{full}	σ_{before}	*scaler.var_*	σ_{150}	σ_{full}
age	60.834	62.947	-0.17	0	11.875	154	0.954	1
creatinine_phosphokinase	581.84	607.15	-0.021	0	968.66	1415489	0.814	1
ejection_fraction	38.084	37.95	0.011	0	11.815	170.02	0.906	1
platelets	263358	266747	-0.035	0	97641	9252860500	1.015	1
serum_creatinine	1.394	1.521	-0.107	0	1.033	1.36	0.885	1
serum_sodium	136.625	136.453	0.038	0	4.405	20.607	0.97	1

По приведённым в таблице 2 данным и гистограммам можно предположить, что StandardScaler приводит матожидание к нулю, а стандартное отклонение и дисперсию у единице. При этом, если брать только первые 150 наблюдений, при подготовке к стандартизации значения матожидания и дисперсии вычисляются не точно, и итоговые значения матожидания и стандартного отклонения отличны от желаемых. Приблизительная формула работы StandardScaler:

$$X'_i = \frac{X_i - M[X]}{\sqrt{D[X]}},$$

где X_i— значение до преобразования, X'_i— преобразованное значение.

3. Приведение к диапазону.

MinMaxScaler. Гистограммы для данных, преобразованных с помощью MinMaxScaler, приведены на рис. 5. По сравнению с исходными признаками, поменялся диапазон: для каждого признака он был приведён к интервалу [0; 1]. Исходные максимальные и минимальные значения признаков можно получить из свойств data_min_ и data_max_ объекта MinMaxScaler соответственно (таблица 3).

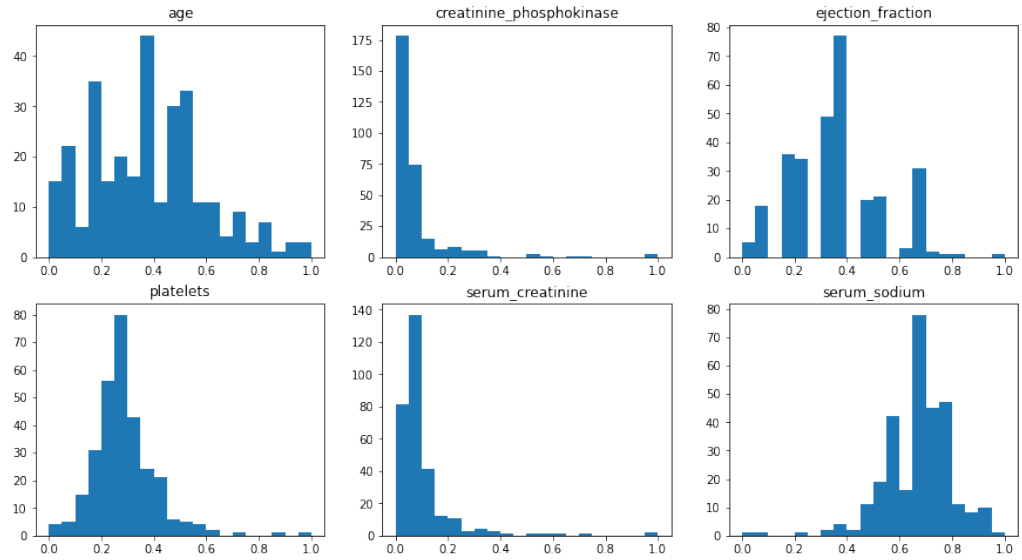

Рис. 5. Гистограммы признаков после обработки MinMaxScaler.

Таблица 3. Минимальыне и максимальные значения признаков.

признак	**минимальное**	**максимальное**
age	40	95
creatinine_phosphokinase	23	7861
ejection_fraction	14	80
platelets	2510	850000
serum_creatinine	0.5	9.4
serum_sodium	113	148

MaxAbsScaler и **RobustScaler**. Были построены гистограммы данных с помощью приведения к диапазону с помощью MaxAbsScaler (рис. 6) и RobustScaler (рис. 7). Первый преобразовывает данные так, чтобы максимальное значение было равно единице по модулю. Второй приводит медианное значение к нулю и масштабирует данные по межквартильному диапазону (по умолчанию, используются 25-й и 75-й перцентиль).

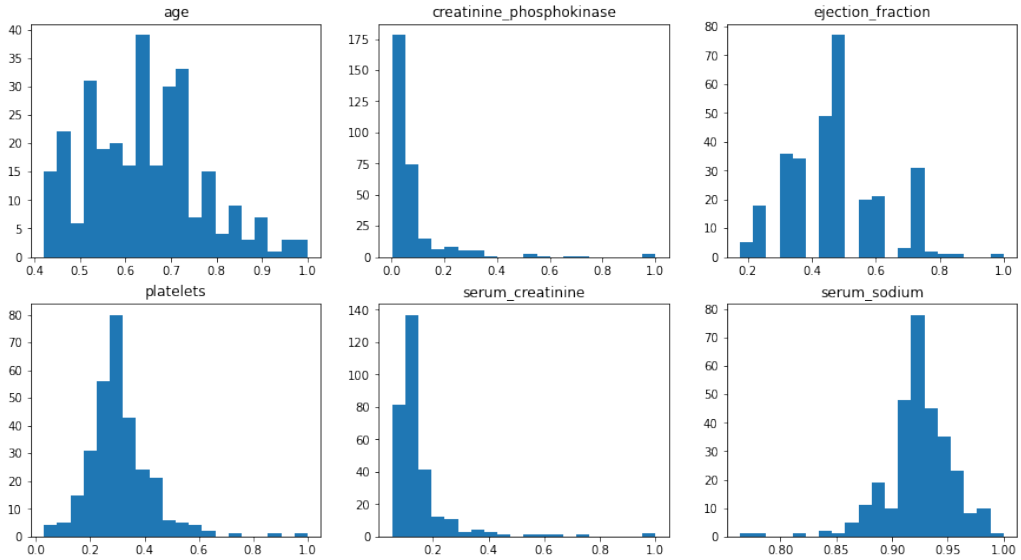

Рис. 6. Гистограммы признаков после преобразования MaxAbsScaler.

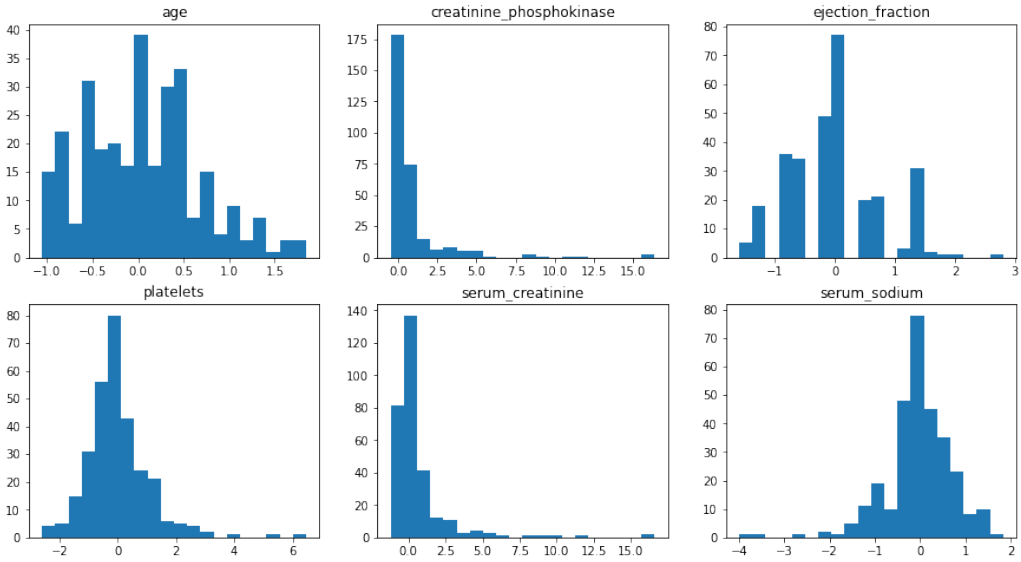

Рис. 7. Гистограммы признаков после преобразования RobustScaler.

Приведение к диапазону [-5; 10]. Функция приведения к указанному диапазону представлена ниже, гистограммы после преобразования приведены на рис. 8.

```
def fit_range(arr, min_val=-5, max_val=10):
    if type(arr) is not np.ndarray:
        raise ValueError('Numpy array is expected!')

    scaled = np.empty(arr.shape)
    rng = max_val - min_val
    for i in range(arr.shape[1]):
        min_, max_ = np.min(arr[:, i]), np.max(arr[:, i])
        scaled[:, i] = [(x - min_) / (max_ - min_) * rng + min_val
for x in arr[:, i]]

    return scaled
```

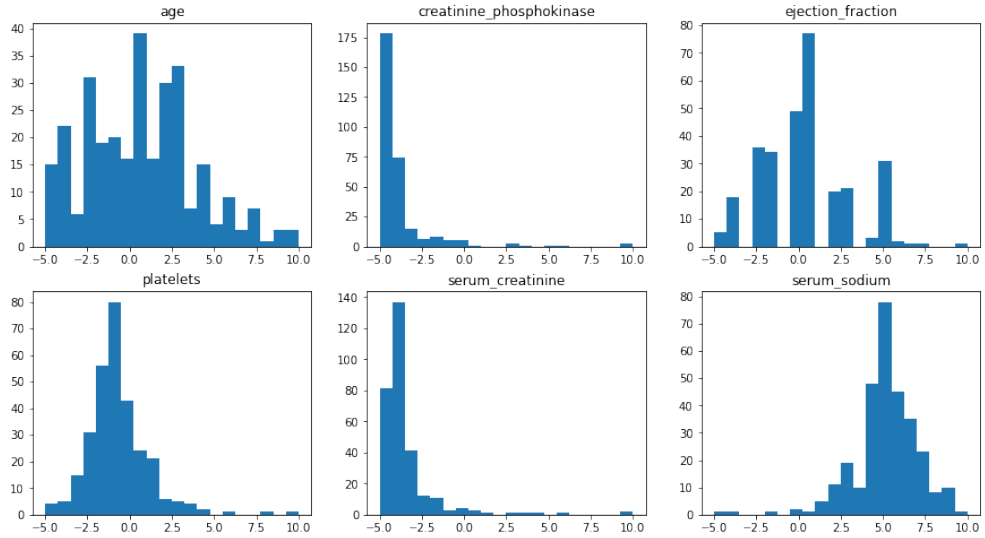

Рис. 8. Гистограммы признаков после приведения к диапазону [-5; 10].

4. Нелинейные преобразования. Для преобразования к равномерному и нормальному распределению был использован QuantileTransformer, гистограммы признаков после преобразования пресдставлены на рис. 9-10. Параметр n_quantiles определяет количество квантилей, используемых для преобразования, чем больше квантилей, тем выше частота дискретизации фукнции распределения.

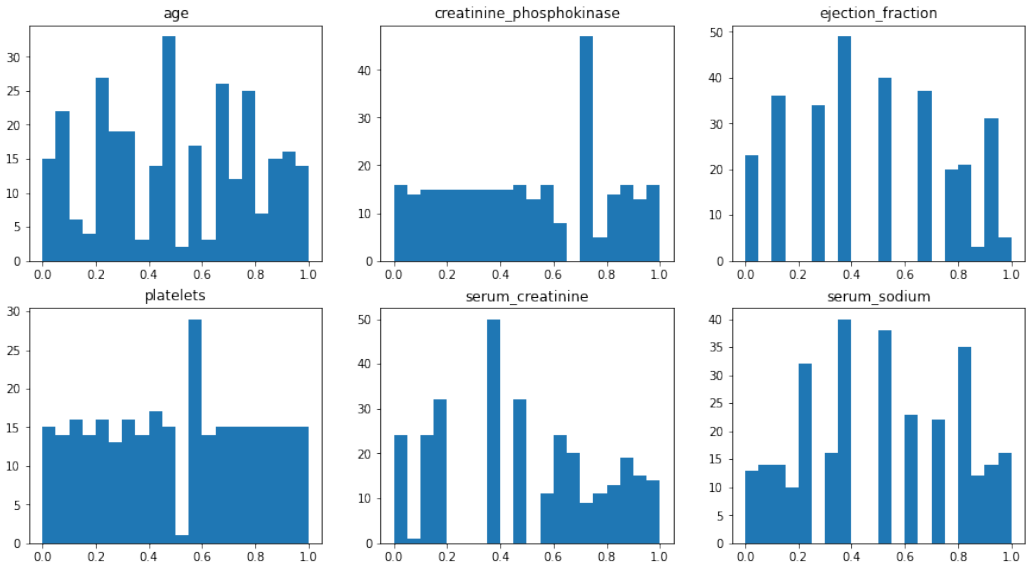

Рис. 9. Гистограммы признаков после преобразования QuantileTransformer в равномерное распределение.

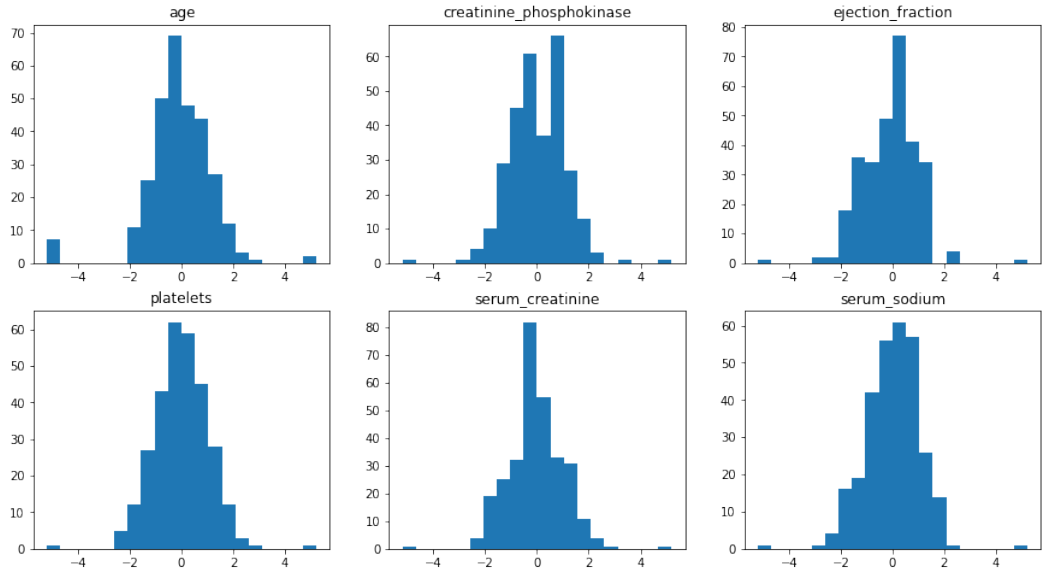

Рис. 10. Гистограммы признаков после преобразования QuantileTransformer в нормальное распределение.

Преобразование к нормальному распределению можно выполнить с помощью PowerTransformer, результат представлен на рис. 11.

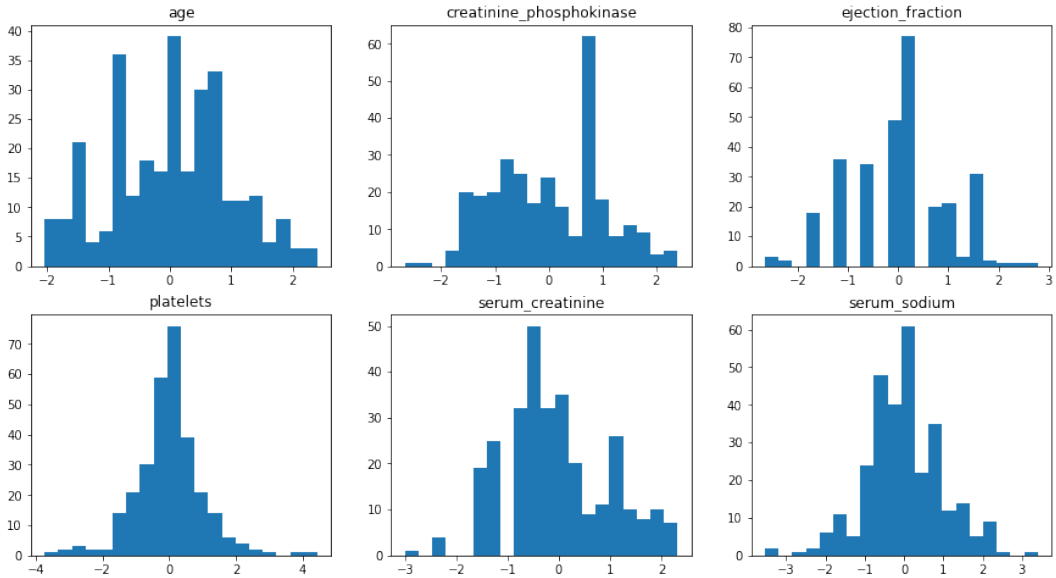

Рис. 11. Гистограммы признаков после преобразования PowerTransformer.

5. Дискретизация признаков. Для дискретизации признаков был использован KBinsDescretizer. Гистограмма показана на рис. 12. Поскольку при использованной дискретизации значения — это числовые обозначения классов, данные гистограммы не имеют особого смысла.

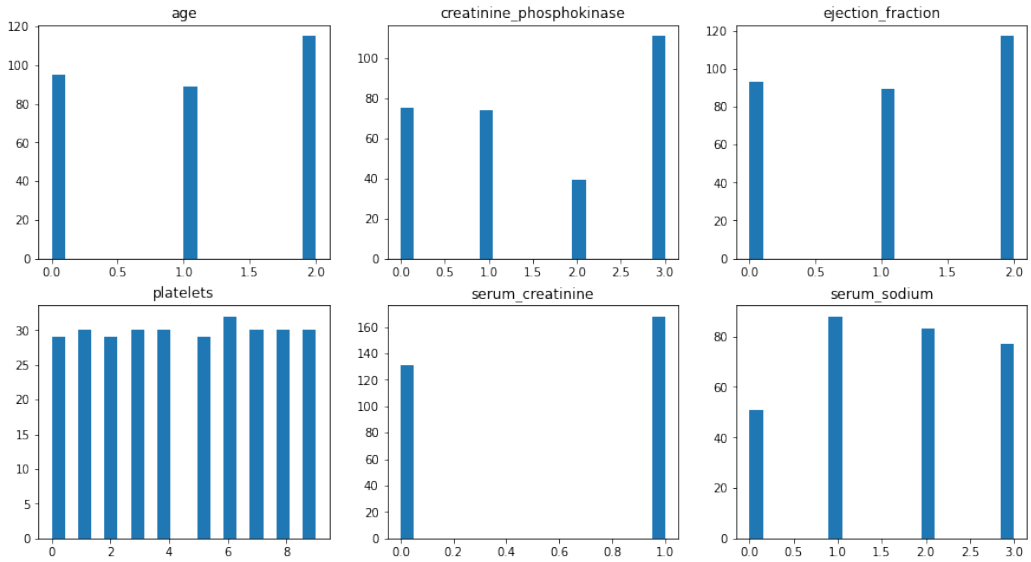

Рис. 12. Гистограммы признаков после дискретизации.

Параметр bin_edges_ объекта KBinsDiscretizer позволяет получить границы полученных после дискретизации диапазонов:

```
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 результате выполнения лабораторной работы были изучены различные методы предобработки данных библиотеки Skikit-Learn. В результате стандартизации данных было установлено, что стандартизация по неполной выборке снижает качество стандартизации. Приведение данных к диапазону позволило изменить границы данных без изменения формы распределения, нелинейные же преобразования данных, напротив, позволяют изменить форму распределния данных, подогнав их к равномерному или нормальному распределению. Дискретизация данных позволяет разбить данные на классы.