

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

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

Студенты гр. 6304

Преподаватель

Тимофеев А.А.

Жангиров Т.Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами предобработки данных из библиотеки Scikit Learn

Ход работы

Загрузка данных

1. Был создан датафрейм Pandas на основе загруженного датасета (<https://www.kaggle.com/andrewmvd/heart-failure-clinical-data>).
2. Из датафрейма были исключены следующие признаки: anaemia, diabetes, high_blood_pressure, sex, smoking, time, DEATH_EVENT.

	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium
count	299.000000	299.000000	299.000000	299.000000	299.000000	299.000000
mean	60.833893	581.839465	38.083612	263358.029264	1.39388	136.625418
std	11.894809	970.287881	11.834841	97804.236869	1.03451	4.412477
min	40.000000	23.000000	14.000000	25100.000000	0.500000	113.000000
25%	51.000000	116.500000	30.000000	212500.000000	0.900000	134.000000
50%	60.000000	250.000000	38.000000	262000.000000	1.100000	137.000000
75%	70.000000	582.000000	45.000000	303500.000000	1.400000	140.000000
max	95.000000	7861.000000	80.000000	850000.000000	9.400000	148.000000

Рисунок 1 – Описание полученного датафрейма

3. Были построены гистограммы признаков, определены диапазоны значений для каждого признака, а также максимальные значения и промежутки, в которых они наблюдались.

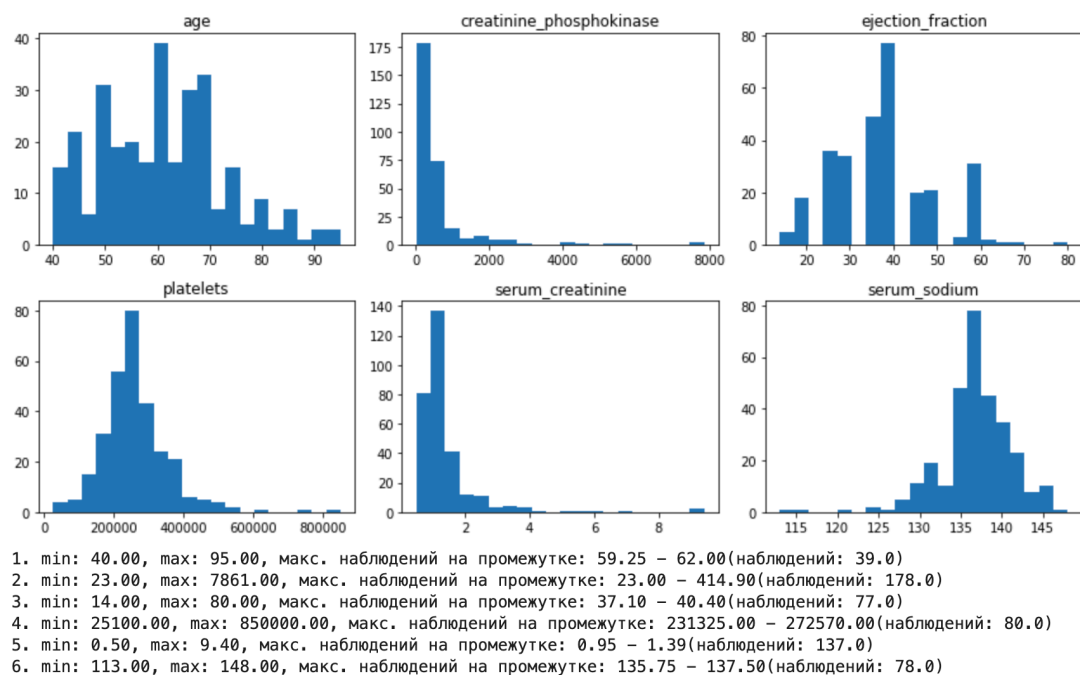


Рисунок 2 – Гистограммы признаков

Стандартизация данных

1. Была произведена стандартизация на 150 значениях, построены гистограммы полученных данных.

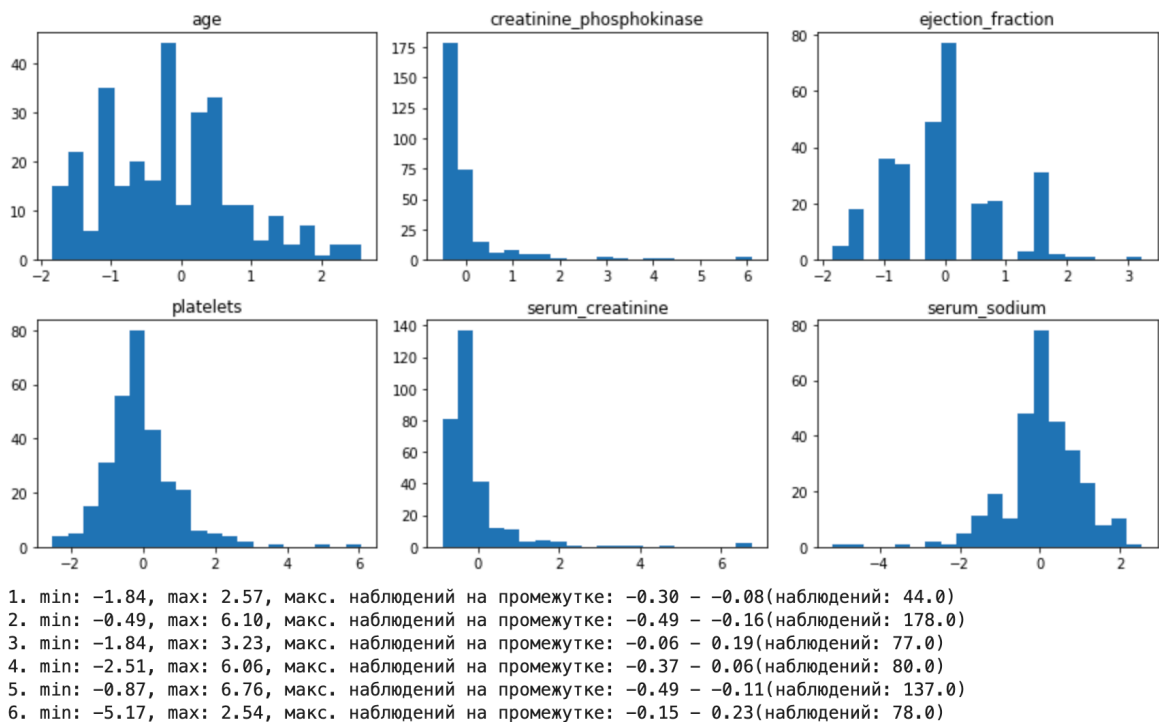


Рисунок 3 – Гистограмма данных, стандартизированных на 150 значениях

2. Было выполнено сравнение математического ожидания и СКО для исходных данных, данных, стандартизированных на 150 значений, а также данных, стандартизированных на всей выборке. В сравнении были рассмотрены значения из атрибутов скейлера mean_ и scale_. Результаты сравнения представлены в таблице 1.

Таблица 1 – Сравнение мат. Ожидания и СКО

	age	creatinine_phospho kinase	ejection_fraction	platelets	serum_creatinine	serum_sodium
Среднее	60.83	581.84	38.08	263358.03	1.39	136.63
СКО	11.87	968.66	11.82	97640.55	1.03	4.41
Среднее (ст. 150)	-0.17	-0.02	0.01	-0.04	-0.11	0.04
СКО (ст. 150)	0.95	0.81	0.91	1.02	0.89	0.97

Среднее (ст. 150 mean_)	62.95	607.15	37.95	266746.75	1.52	136.45
СКО (ст. 150 scale_)	12.45	1189.74	13.04	96191.79	1.17	4.54
Среднее (ст. все)	0.00	0.00	-0.00	0.00	0.00	-0.00
СКО (ст. все)	1.00	1.00	1.00	1.00	1.00	1.00
Среднее (ст. все mean_)	60.83	581.84	38.08	263358.03	1.39	136.63
СКО (ст. все scale_)	11.87	968.66	11.82	97640.55	1.03	4.41

На основании полученных данных можно сделать вывод, что StandardScaler приводит исходную выборку к выборке с нулевым мат. ожиданием и единичной дисперсией, по формуле:

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

где X_i – значения исходной выборки, а X_i' – результат. В атрибутах mean_ и scale_ скейлера хранятся значения мат. ожидания и СКО исходной выборки, которые используются при стандартизации.

Приведение к диапазону

1. Было выполнено приведение к диапазону при помощи MinMaxScaler, построены гистограммы признаков.

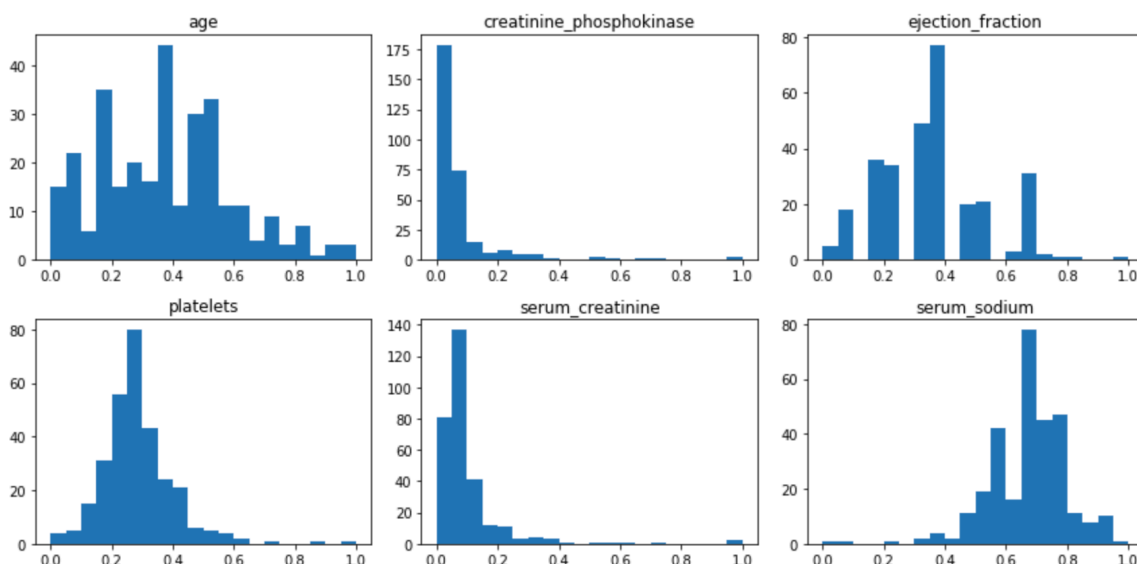


Рисунок 4 – Гистограммы после использования MinMaxScaler

Данные были приведены к диапазону [0, 1] по формуле:

$$X_i' = \frac{X_i - \min(X)}{\max(X) - \min(X)}$$

где X_i – значения исходной выборки, а X_i' – результат.

- Из атрибутов `data_min_` и `data_max_` были получены минимальные и максимальные значения признаков. Результаты представлены в таблице 2.

Таблица 2 – Минимальные и максимальные значения признаков.

	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium
Min	40	23	14	25100	0.50	113
Max	95	7861	80	850000	9.40	148

- Были выполнены трансформации с помощью `MaxAbsScaler` и `RobustScaler`.

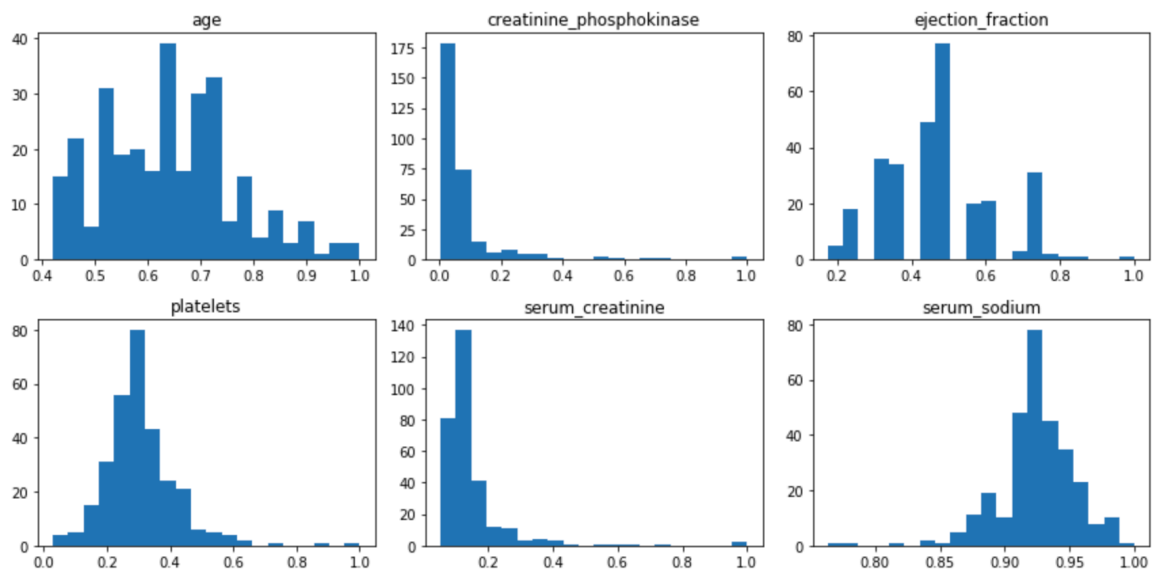


Рисунок 5 – Гистограммы после использования `MaxAbsScaler`

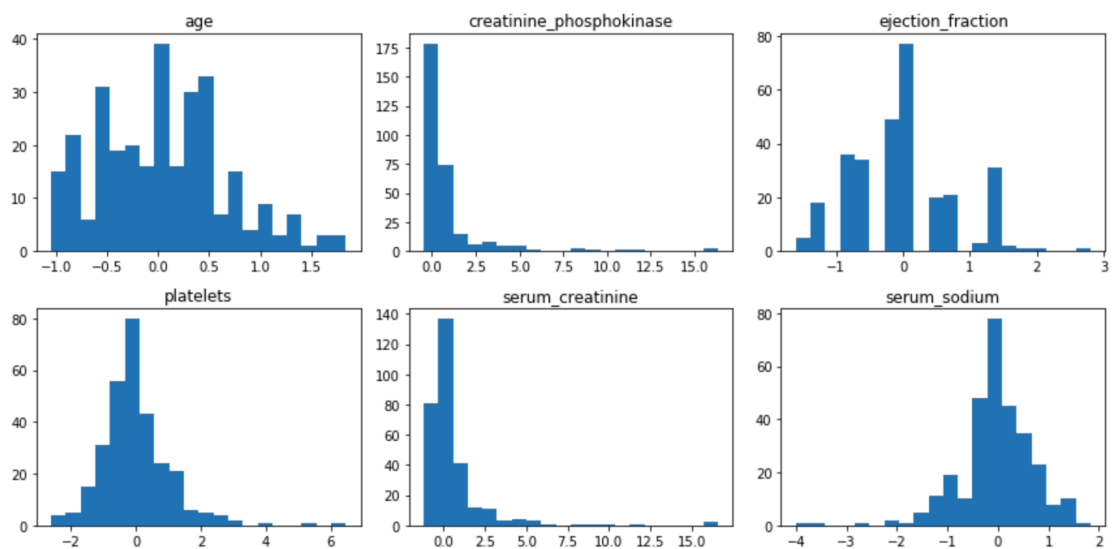


Рисунок 6 – Гистограммы после использования RobustScaler

MaxAbsScaler приводит к диапазону таки образом, чтобы максимальное абсолютное значение выборки было равно единице. RobustScaler приводит медиану выборки к 0, а также масштабирует выборку относительно межквартильного диапазона.

4. Была написана функция, приводящая выборку к диапазону [-5, 10]:

```
def fit_5_10(data):
    scaler = preprocessing.MinMaxScaler(feature_range=(-5, 10)).fit(data)
    return scaler.transform(data)
```

Нелинейные преобразования

1. Было выполнено приведение данных к равномерному и нормальному распределениям при помощи QuantileTransformer.

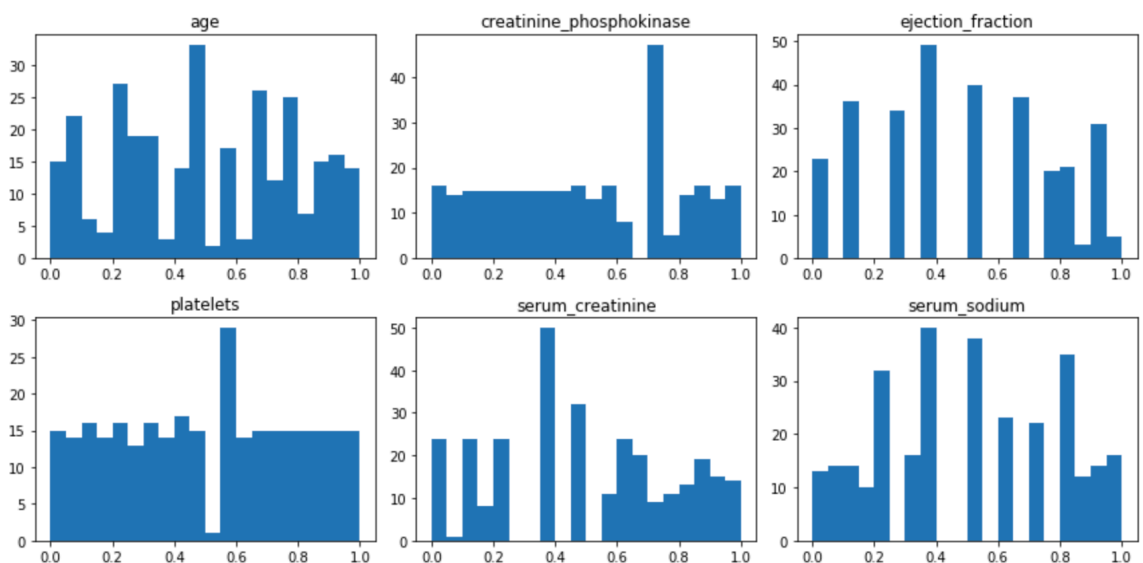


Рисунок 7 – Гистограммы после использования QuantileTransformer

(равн.)

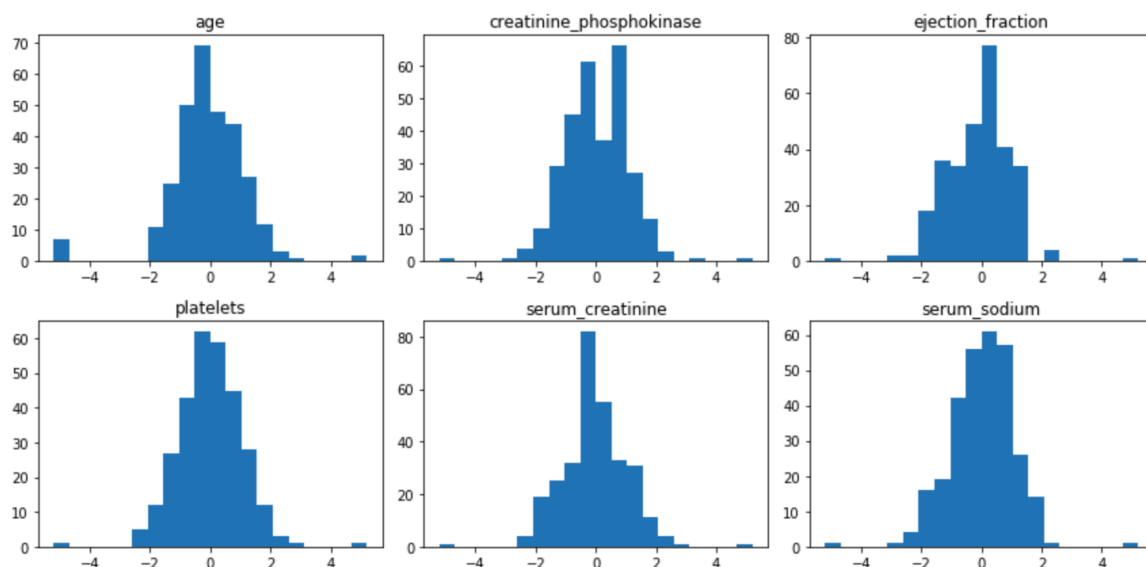


Рисунок 8 – Гистограммы после использования QuantileTransformer
(норм.)

Задаваемое количество квантилей определяет их число при дискретизации функции распределения. Чем больше значение, тем распределение результирующей выборки ближе к требуемому.

2. Было выполнено приведение данных к нормальному распределению при помощи PowerTransformer.

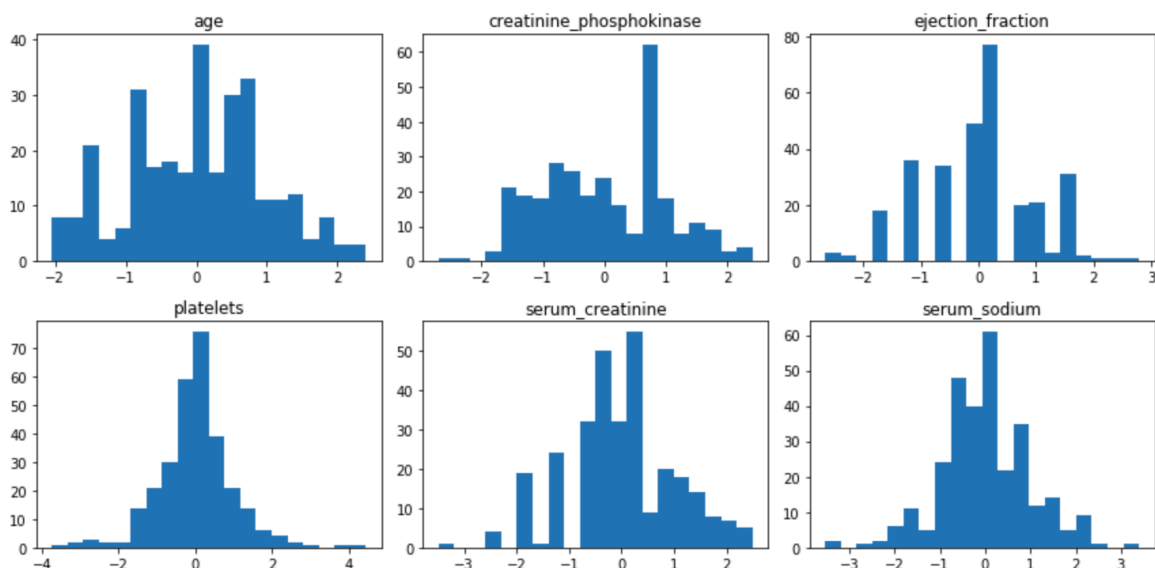


Рисунок 9 – Гистограммы после использования PowerTransformer

Дискретизация признаков

1. Была проведена дискретизация признаков при помощи KBinsDiscretizer.

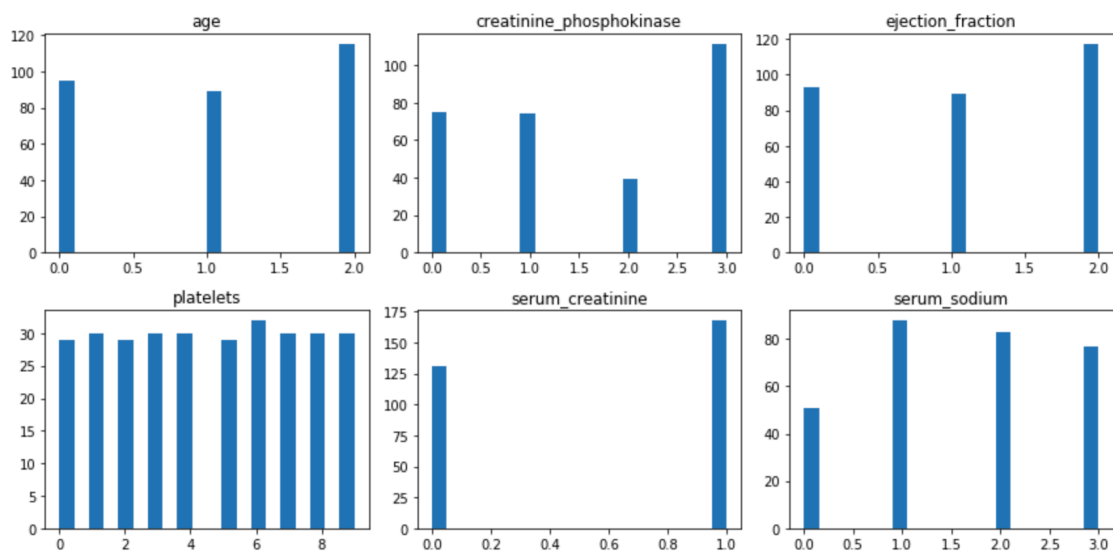


Рисунок 9 – Гистограммы после дискретизации

Через атрибут `bin_edges_` были получены диапазоны интервалов для каждого признака. Результаты представлены в таблице 3.

Таблица 3 – Диапазоны интервалов признаков

Признак	Диапазоны
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_creatinine	0.5, 1.1, 9.4
serum_sodium	113., 134., 137., 140., 148.

Выводы

Было выполнено знакомство с методами предобработки данных из библиотеки Scikit Learn.

Полученные результаты показали:

- Количество значений в выборке влияет на качество стандартизации
- Приведение к диапазону не изменяет форму выборки
- Нелинейные преобразования изменяют форму выборки в соответствии с выбранным распределением