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

ОТЧЕТ

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

Студент гр. 6307 _____ Мишанов А. А.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург
2020

Цель работы

Познакомиться со средствами предобработки данных библиотеки Scikit Learn.

Ход работы

Загрузка данных

1. Загружен датасет по ссылке. Данные загружены в датафрейм с исключением бинарных 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

Рисунок 1. Загруженный датасет

2. Построены гистограммы признаков.

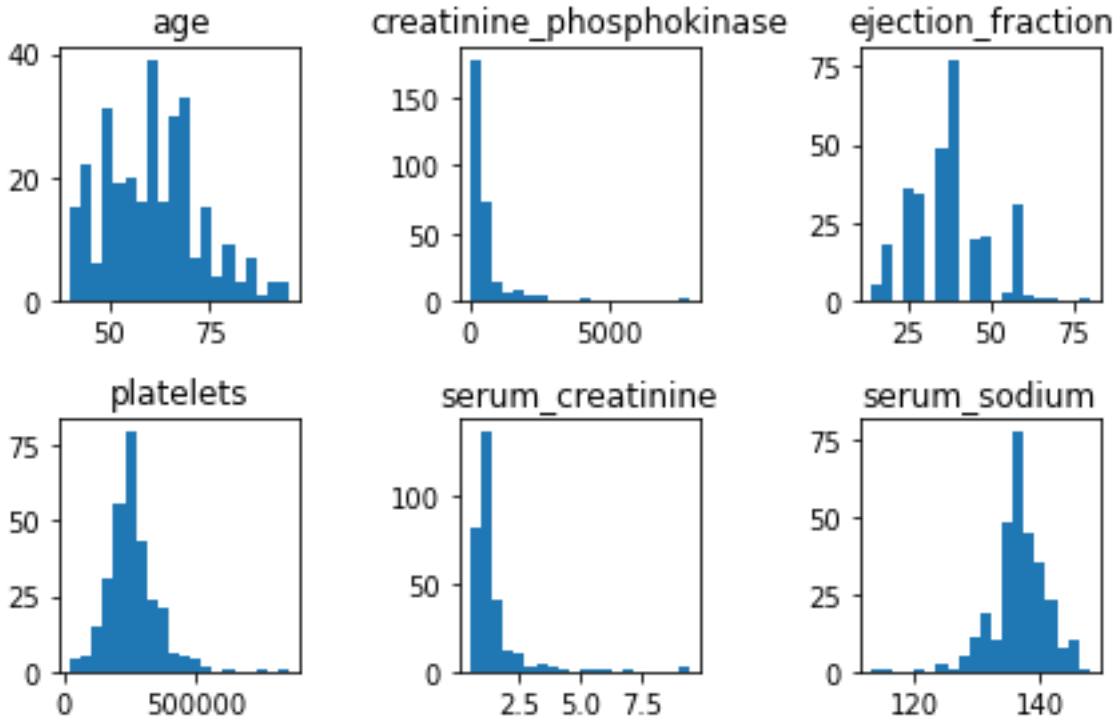

Рисунок 2. Гистограммы признаков

3. На основании гистограмм были определены диапазоны значений, а также возле какого значения лежит наибольшее количество наблюдений.

Признак	Диапазон	Наибольшее количество наблюдений
age	40-100	60
creatinine_phosphokinase	0-8000	200
ejection_fraction	10-80	38
platelets	0-875000	250000
serum_creatinine	0.1-9.75	1.2
serum_sodium	110-150	137

Стандартизация данных

1. Была проведена стандартизация на основе первых 150 наблюдений и на основе всех наблюдений.

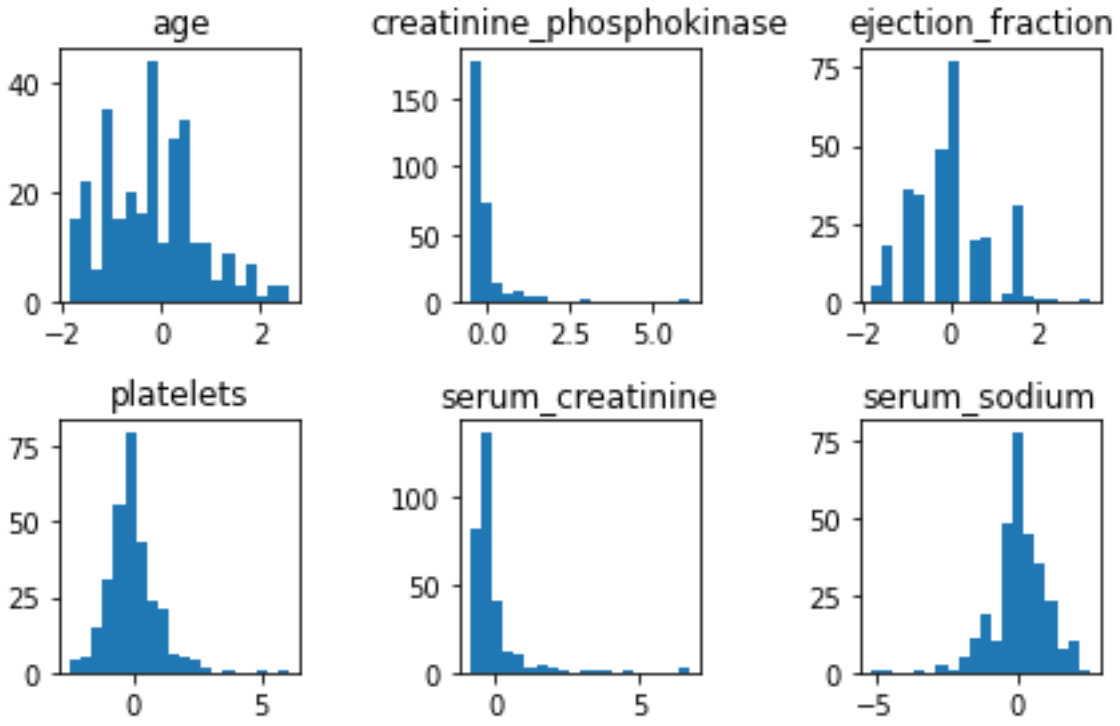

Рисунок 3. Гистограмма стандартизированных данных первых 150 наблюдений

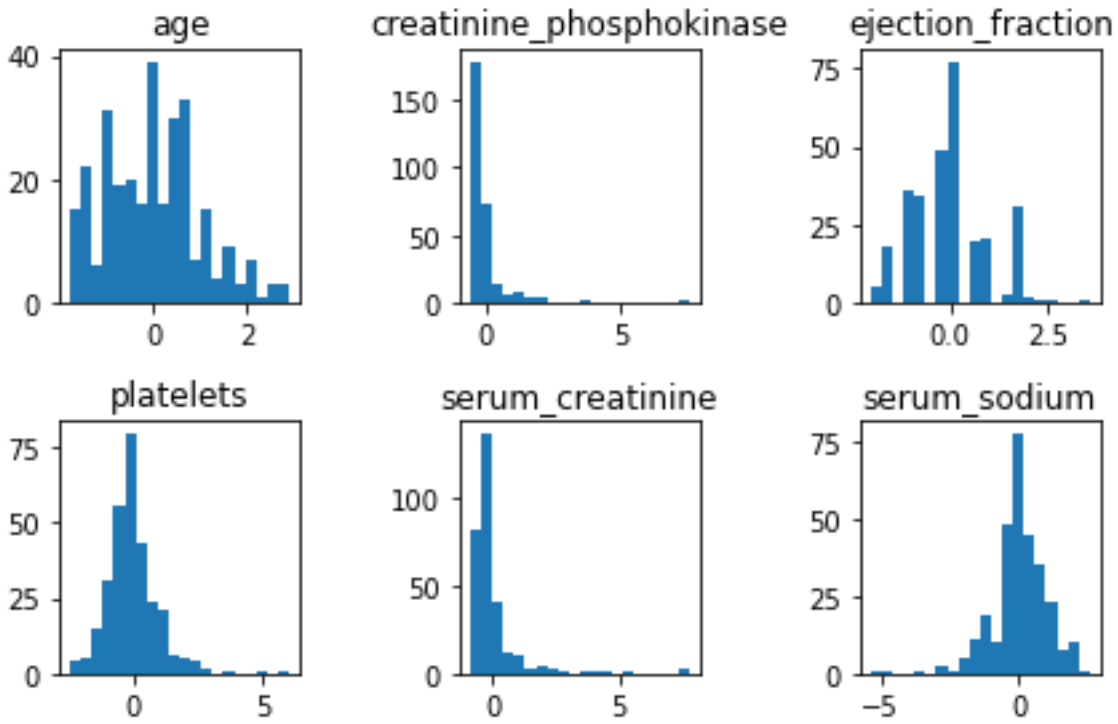

Рисунок 4. Гистограмма стандартизированных данных для всех наблюдений

2. Вычисленное мат. ожидание и СКО до и после стандартизации.

- До стандартизации.
    - Мат. ожидание: [6.0830000e+01, 5.8184000e+02, 3.8080000e+01, 2.6335803e+05,1.3900000e+00, 1.3663000e+02]
    - СКО: [1.18749014e+01, 9.68663967e+02, 1.18150335e+01, 9.76405477e+04, 1.03277867e+00, 4.40509238e+00]
- Стандартизация на основе 150 наблюдений.
    - Мат. ожидание: [-0.16970362, -0.02127675, 0.01050249, -0.03522879, -0.1086408, 0.0379076]
    - СКО: [0.95382379, 0.81417905, 0.90610822, 1.01506113, 0.88542887, 0.9703736]
- Полная стандартизация.
    - Мат. ожидание: [5.70335306e-16, 0.00000000e+00, -3.26754603e-17, 7.72329061e-17, 1.42583827e-16, -8.67384945e-16]
    - СКО: [1., 1., 1., 1., 1., 1.]

3. На основании этих значений можно вывести формулу, по которым они стандартизировались.

$$Y = \frac{X - mean(X)}{std(X)}$$

4. Значения мат. ожидания и дисперсии соответствуют с полями *mean_* и *var_* объекта *Scaler*.

Приведение к диапазону

1. Приведение к диапазону [0, 1] с помощью *MinMaxScaler*. Гистограмма представлена на рисунке 5.

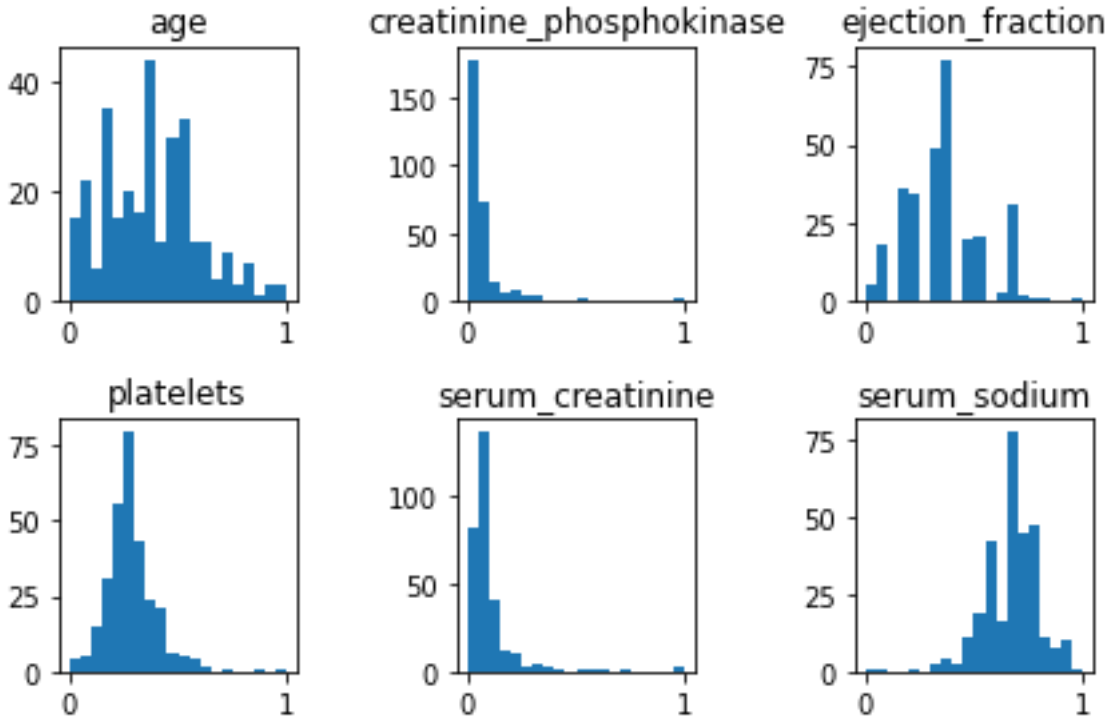

Рисунок 5. Приведение к диапазону с помощью *MinMaxScaler*

2. Через параметры *MixMaxScaler* были определены минимальное и максимальное значение в данных для каждого признака.
   - Минимальные значения: [4.00e+01, 2.30e+01, 1.40e+01, 2.51e+04, 5.00e-01, 1.13e+02]
   - Максимальные значения: [9.500e+01, 7.861e+03, 8.000e+01, 8.500e+05, 9.400e+00, 1.480e+02]

3. Аналогично были трансформированы данные с использованием *MaxAbsScaler* и *RobustScaler*. Гистограммы представлены на рисунке 6 и рисунке 7 соответственно.

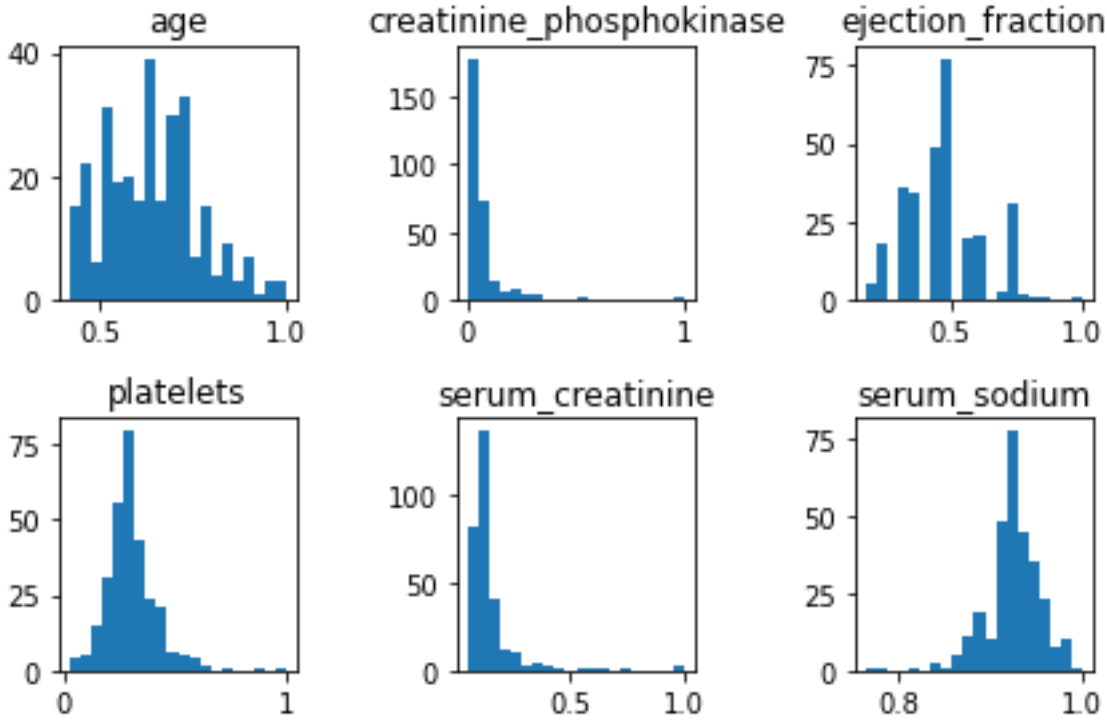

Рисунок 6. Приведение к диапазону с помощью *MaxAbsScaler*

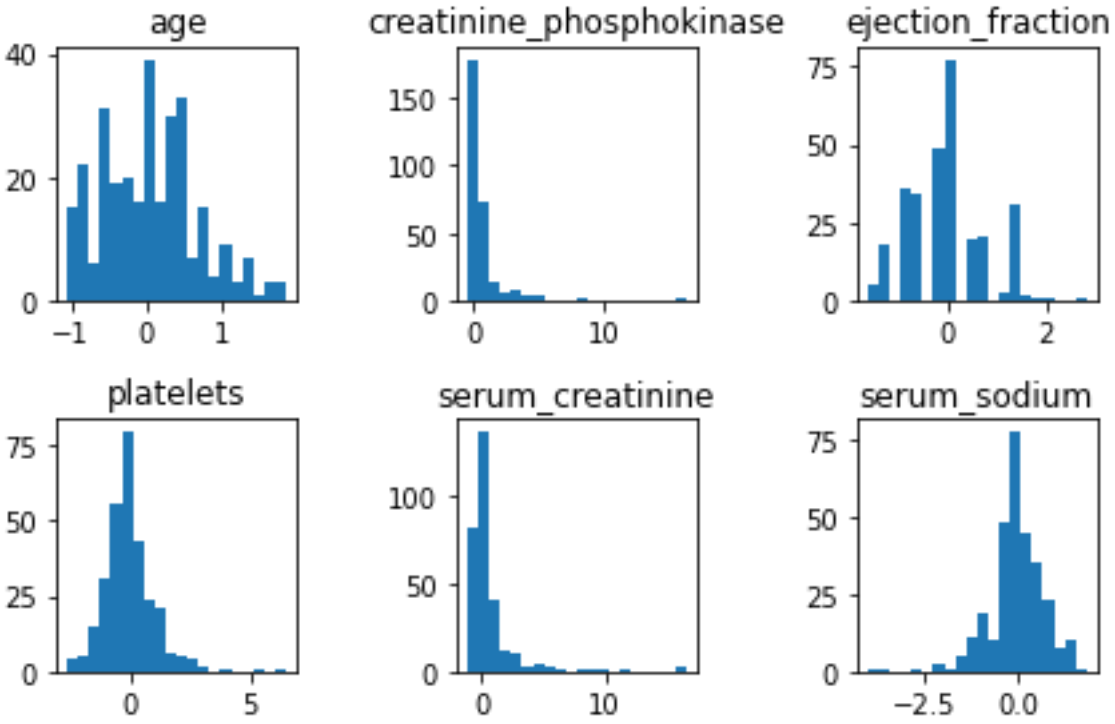

Рисунок 7. Приведение к диапазону с помощью *RobustScaler*

4. *MaxAbsScaler* масштабирует данные таким, что максимальное по модулю значения равно 1. RobustScaler центрирует значения по медиане и масштабируют их по межквартильному размаху.
5. Была написана функция, которая приводит данные к диапазону [-5, 10].

$$Y = 15 * \frac{X - \min(X)}{\max(X) - \min(X)} - 5$$

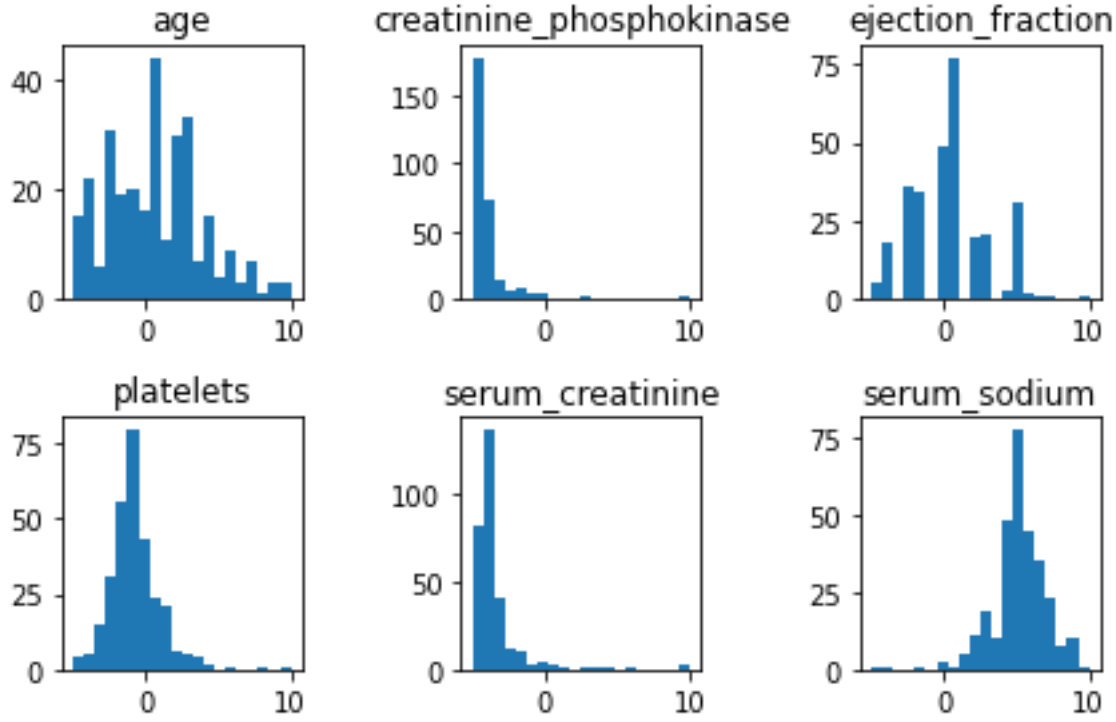

Рисунок 8. Приведение к диапазону [-5 10]

Нелинейные преобразования

1. Приведение данных к равномерному распределению с помощью QuantileTransformer. Гистограмма представлена на рисунке 9. Чем больше количество квантилей (параметр *n_quantiles*), тем лучше приближение к требуемому распределению.

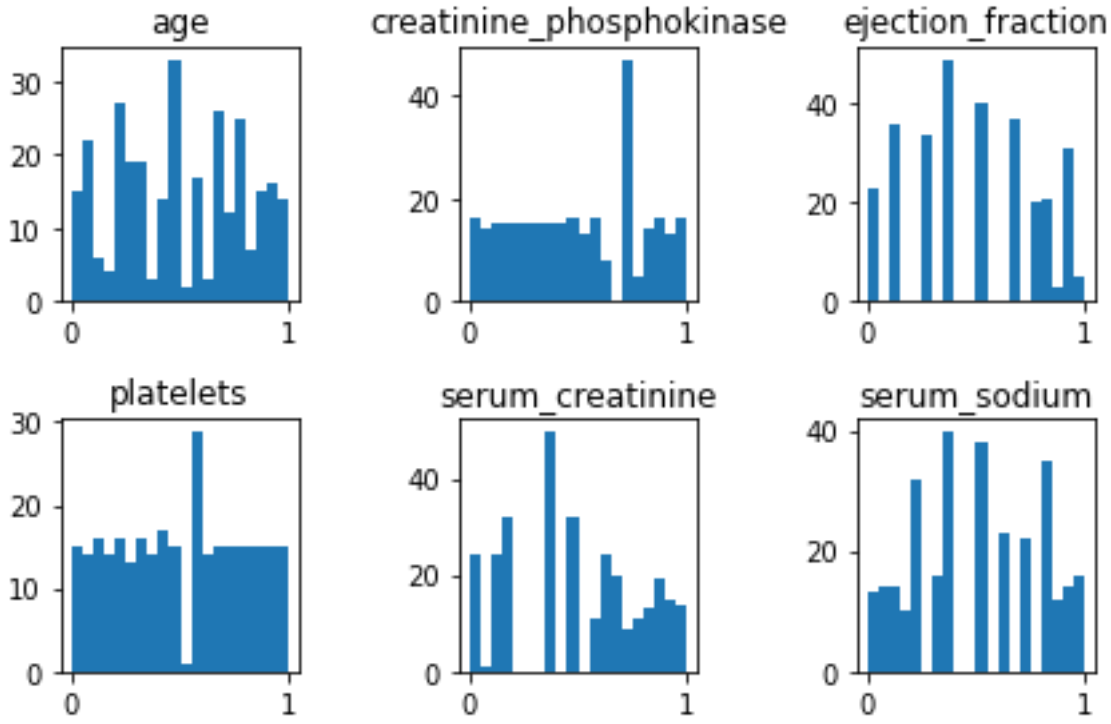

Рисунок 9. Приведение к равномерному распределению с помощью *QuantileTransformer*

2. Были приведены данные к нормальному распределению с использованием параметра *output_distribution='normal'*. Также построены гистограммы для нормального распределения.

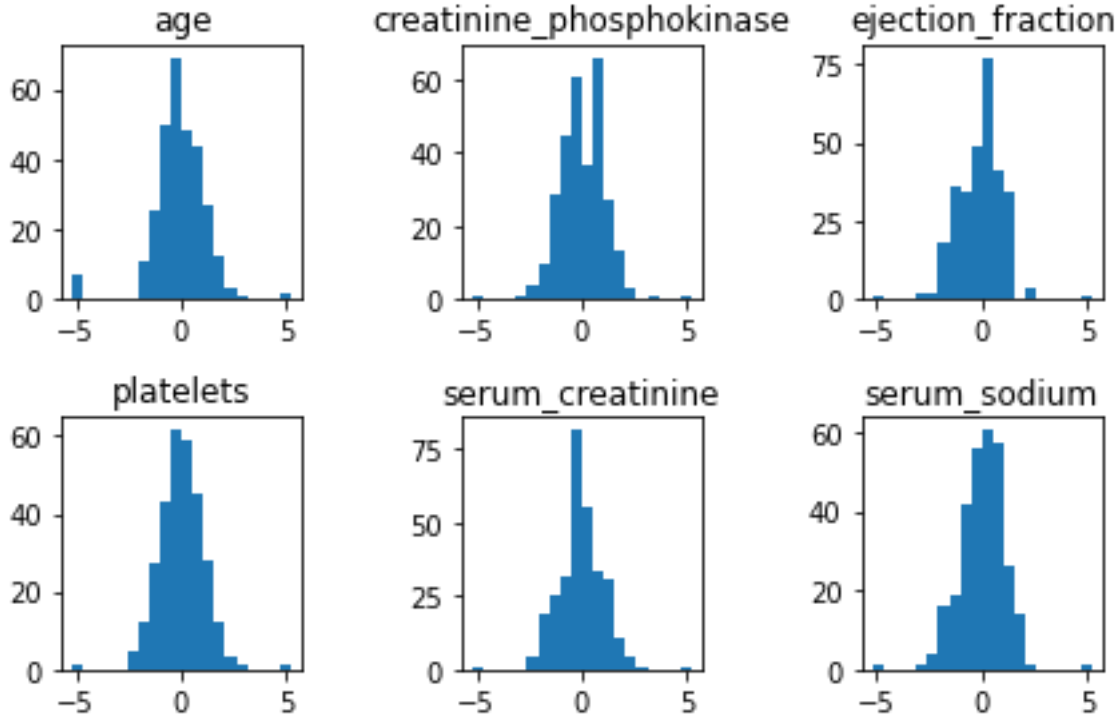

Рисунок 10. Приведение к нормальному распределению с помощью *QuantileTransformer*

3. Также с помощью объекта *PowerTransformer* данные были приведены к нормальному распределению.

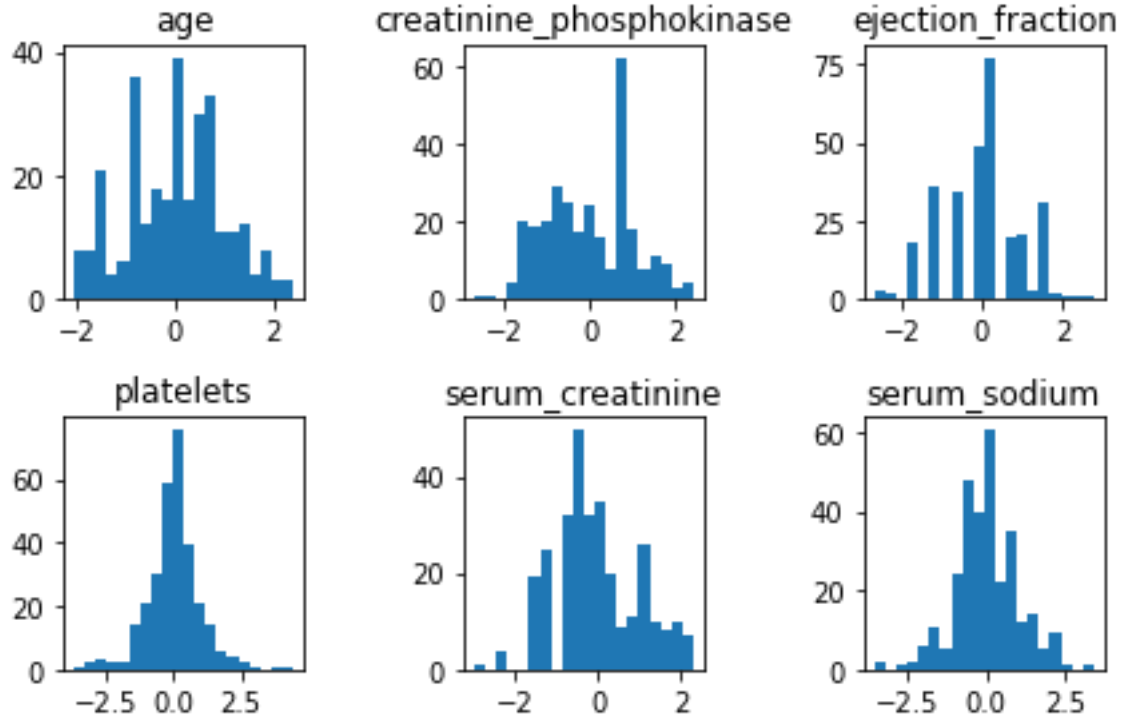

Рисунок 11. Приведение к нормальному распределению с помощью *PowerTransformer*

Дискретизация признаков

1. Была проведена дискретизация признаков с помощью KBinsDiscretizer. Гистограммы представлены на рисунке 12.

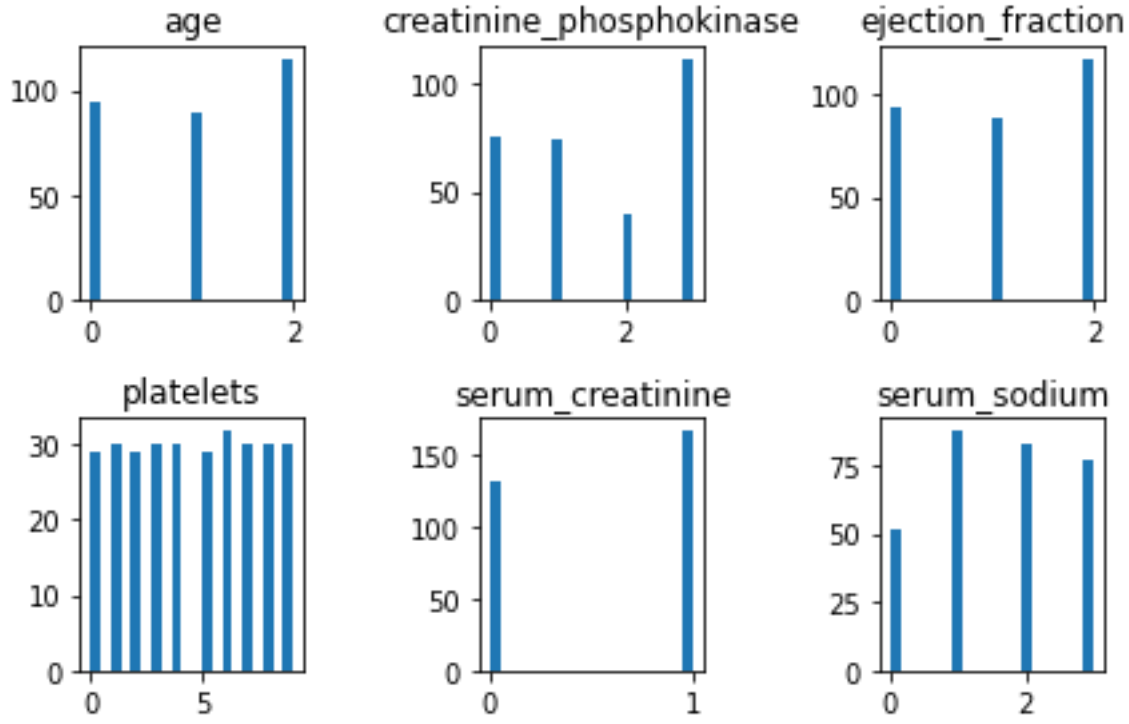

Рисунок 12. Дискретизация признаков с помощью KBinsDiscretizer

2. Диапазоны для каждого интервала.

Признак	Диапазон интервалов
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.]

Вывод

В результате работы были получены практические навыки с методами предобработки данных с помощью библиотеки Scikit Learn.