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

Студентка гр. 6307 _____ Кичерова А. Д.

Преподаватель _____ Жангирова Т. Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами предобработки данных из библиотеки Scikit Learn.

Ход работы

1. Загрузка данных

Загружаем датасет в датафрейм и исключаем бинарные признаки и признаки времени. Фрагмент получившегося датасета приведен на рисунке 1.

	age	creatinine_phosphokinase	...	serum_creatinine	serum_sodium
0	75.0	582	...	1.9	130
1	55.0	7861	...	1.1	136
2	65.0	146	...	1.3	129
3	50.0	111	...	1.9	137
4	65.0	160	...	2.7	116
..	...	...	...	...	...
294	62.0	61	...	1.1	143
295	55.0	1820	...	1.2	139
296	45.0	2060	...	0.8	138
297	45.0	2413	...	1.4	140
298	50.0	196	...	1.6	136

[299 rows x 6 columns]

Рисунок 1. Фрагмент исходного датасета.

По исходному датасету построены гистограммы, приведенные на рисунке 2.

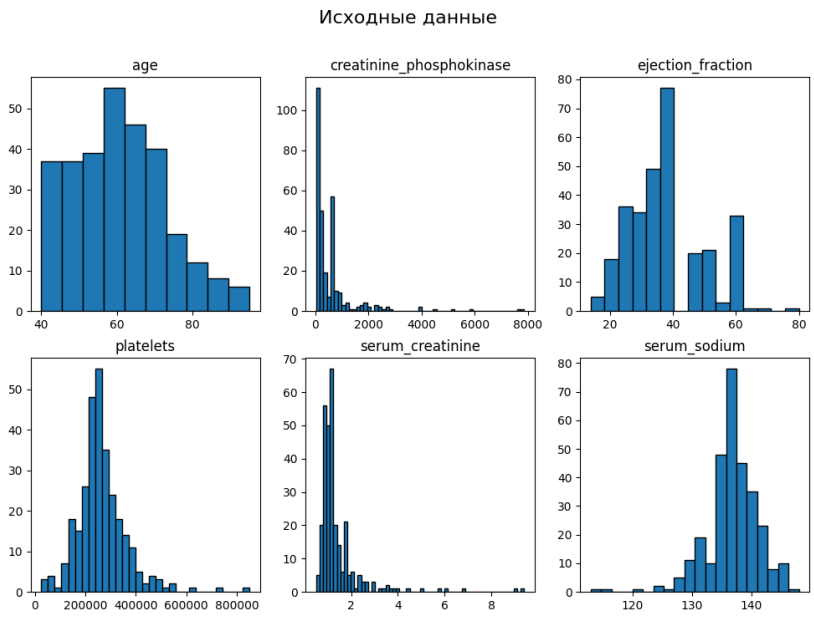

Рисунок 2. Гистограммы признаков

По гистограммам установились приблизительные диапазоны значений, а также значения, которым принадлежит наибольшее количество наблюдений. Установленные значения можно увидеть в таблице 1

Таблица 1. Оценки диапазона и мода признаков.

Признак	Диапазон	Мода
age	40 – 95	60
creatinine_phosphokinase	0 – 7861	20 - 415
ejection_fraction	14 – 80	38.5
platelets	25100 – 850000	250000
serum_creatinine	0.5 – 9	1.5
serum_sodium	113 – 148	136.5

2. Стандартизация данных

Первоначально данные стандартизируются на основе первых 150 наблюдений, затем стандартизация проходит по всем данным. Гистограммы стандартизированных данных представлены на рисунках 3 и 4.

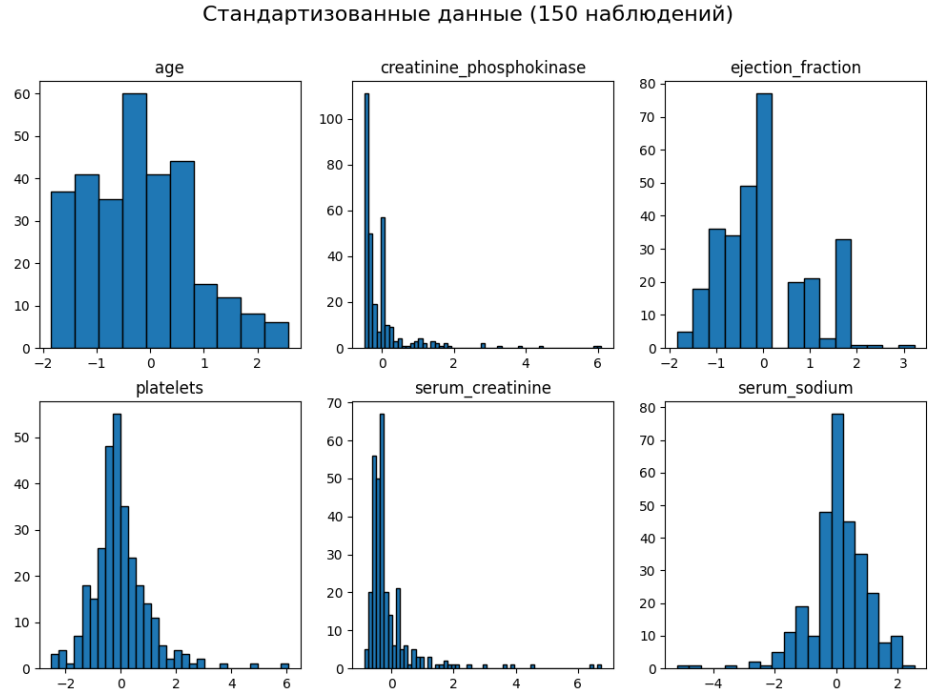

Рисунок 3. Гистограмма стандартизованных данных на основе 150 наблюдений.

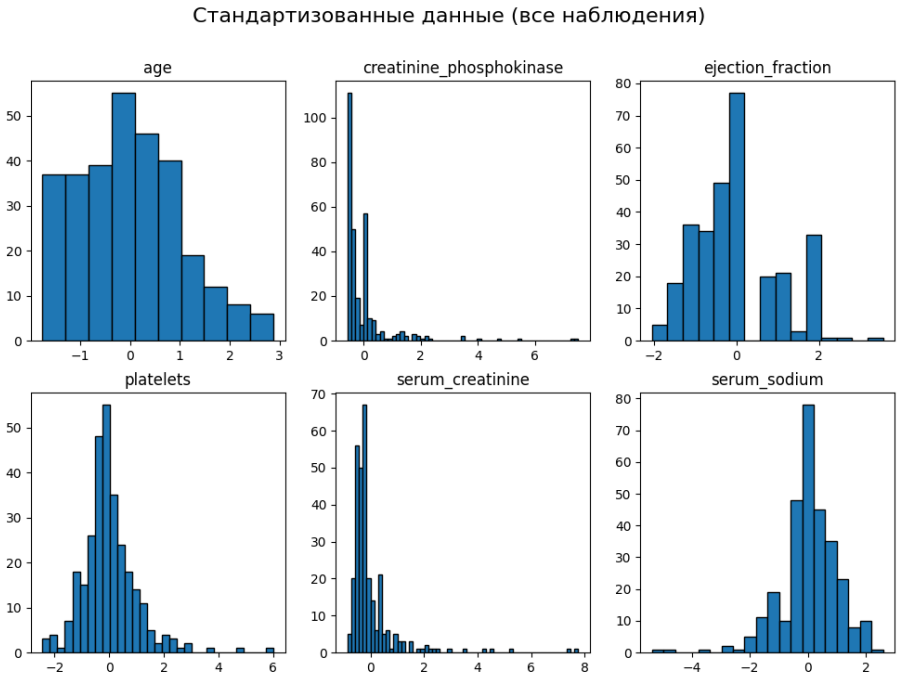

Рисунок 4. Гистограмма стандартизованных данных.

В таблице 2 представлены расчеты мат. Ожидания и СКО до и после стандартизации, а также сравнение с полями mean_ b и var_ объекта scaler.

Таблица 2. Сравнительная таблица до и после стандартизации.

Признак	mean original	mean scaled 150	scared.mean_ 150	mean scaled all	scared.mean_ all
age	60.83389297658862	-0.16970362369106984	62.946666666666665	5.703353062957326e-16	60.83389297658862
creatinine_phosphokinase	581.8394648829432	-0.021276750290383013	607.1533333333333	0.0	581.8394648829432
ejection_fraction	38.08361204013378	0.01050249484809085	37.946666666666665	-3.267546025652635e-17	38.08361204013378
platelets	263358.02926421404	-0.035228788194085287	266746.74946666666	7.723290606088045e-17	263358.02926421404
serum_creatinine	1.3938795986622072	-0.10864080163893569	1.5206000000000002	1.4258382657393315e-16	1.3938795986622072
serum_sodium	136.62541806020067	0.03790759894920013	136.45333333333335	-8.673849449914267e-16	136.62541806020067

Признак	std original	std scaled 150	scared.var_ 150	std scaled all	scared.var_ all
age	11.874901429842655	0.9538237876978354	154.99715555555557	0.9999999999999998	141.01328396847913
creatinine_phosphokinase	968.6639668032415	0.8141790488228113	1415488.8231555554	1.0	938309.8805829913
ejection_fraction	11.815033462318585	0.9061082161919123	170.02382222222224	1.0	139.5950157157079
platelets	97640.54765451424	1.0150611342848024	9252860499.078917	1.0	9533676546.273466
serum_creatinine	1.0327786652795918	0.8854288727548568	1.3605269733333336	1.0	1.066631771456695
serum_sodium	4.405092379513557	0.9703735961735016	20.607822222222225	0.9999999999999999	19.404838872048412

По гистограммам и таблице 2 можно предположить, что стандартизация приводит стандартное отклонение к 1, а математическое ожидание к 0. Приблизительная формула работы ($x_i -$ значение до преобразования; $x_i' -$ значение после преобразования

$$x_i' = \frac{x_i - M[x]}{\sqrt{Dx}}$$

3. Приведение к диапазону MinMaxScaler.

После приведения данных к диапазону с помощью MinMaxScaler диапазон для каждого признака был приведен к интервалу [0;1]. Гистограмма для данных, преобразованных с помощью MinMaxScaler приведена на рисунке 5.

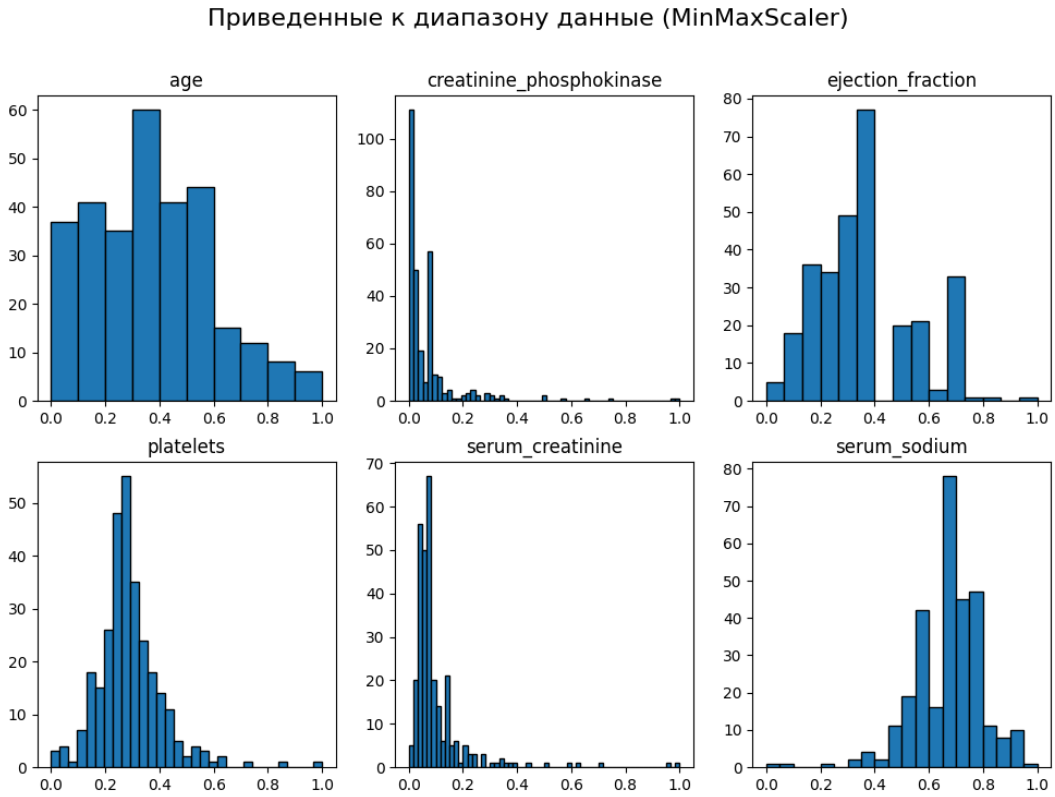

Рисунок 5. Гистограмма признаков после обработки MinMaxScaler

Исходные максимальные и минимальные значения признаков можно получить из свойств data_min_ и data_max_ объекта MinMaxScaler. Результат представлен на рисунке 6.

```
MinMaxScaler
Минимум:
40.0 23.0 14.0 25100.0 0.5 113.0

Максимум:
95.0 7861.0 80.0 850000.0 9.4 148.0
```

Рисунок 6. Минимальыне и максимальные значения признаков

MaxAbsScaler и RobustScaler.

Были построены гистограммы данных после приведения к диапазону с помощью MaxAbsScaler и RobustScaler. Результат на рисунках 7 и 8, соответственно.

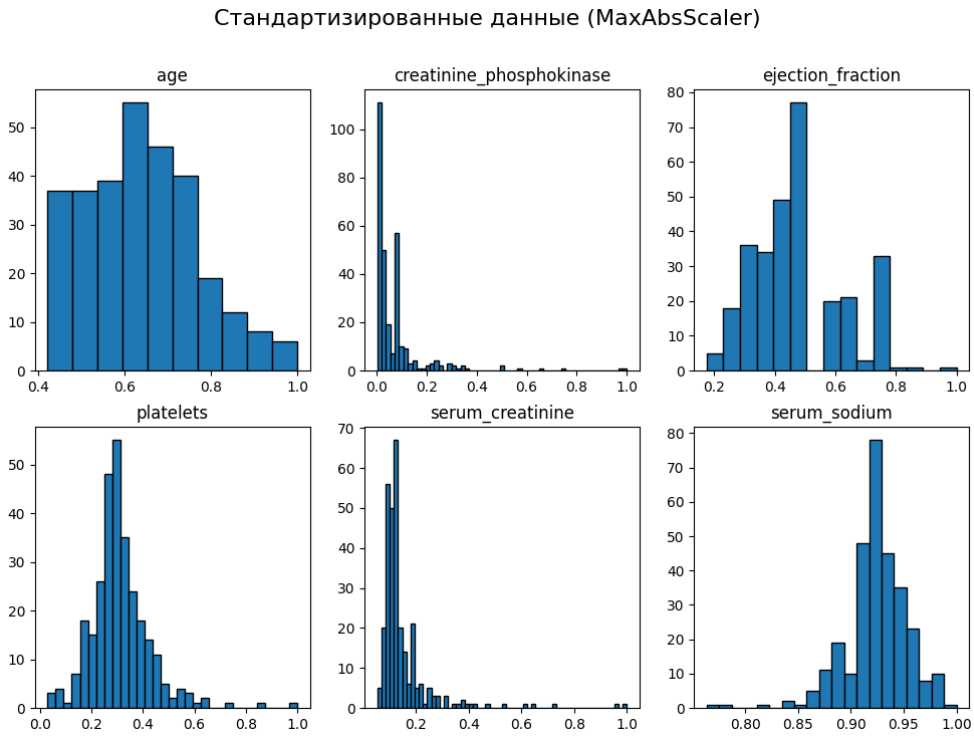

Рисунок 7. Гистограмма признаков после обработки MaxAbsScaler

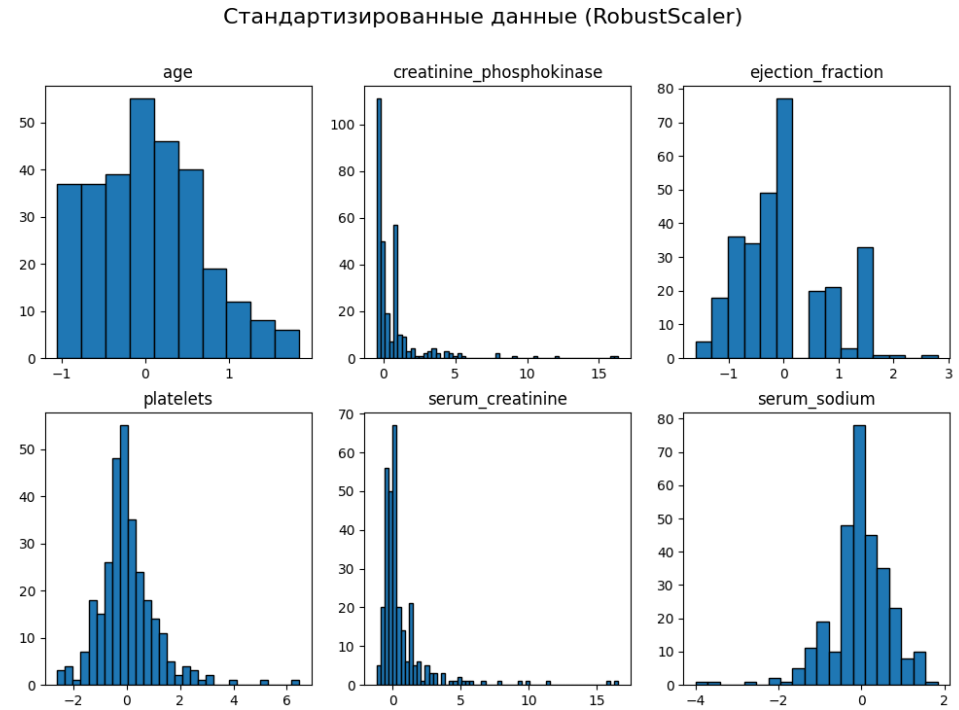

Рисунок 8. Гистограмма признаков после обработки RobustScaler

Различие их работы заключается в том, что MaxAbsScaler преобразует данные так, чтобы максимальное значение было равно 1. В свою очередь, RobustScaler приводит медианное значение к нулю и масштабирует данные в соответствии с квантильным диапазоном.

Приведение к диапазону [-5; 10]

Была написана функция, приводящая все данные к диапазону [-5; 10].

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

Гистограмма данных после приведения показана на рисунке 9.

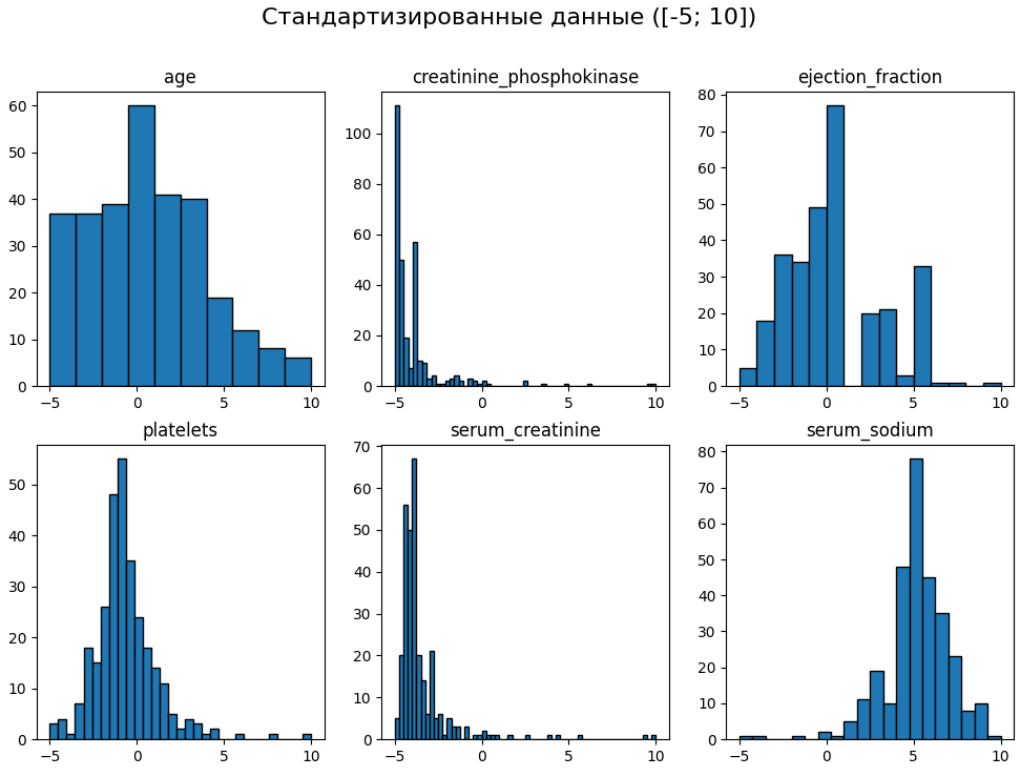

Рисунок 9. Гистограмма признаков после приведения к диапазону [-5; 10].

4. Нелинейные преобразования

Для преобразования к равномерному и нормальному распределению был использован QuantileTransformer, гистограммы признаков после преобразования представлены на рисунках 10, 11 и 12, 13 соответственно.

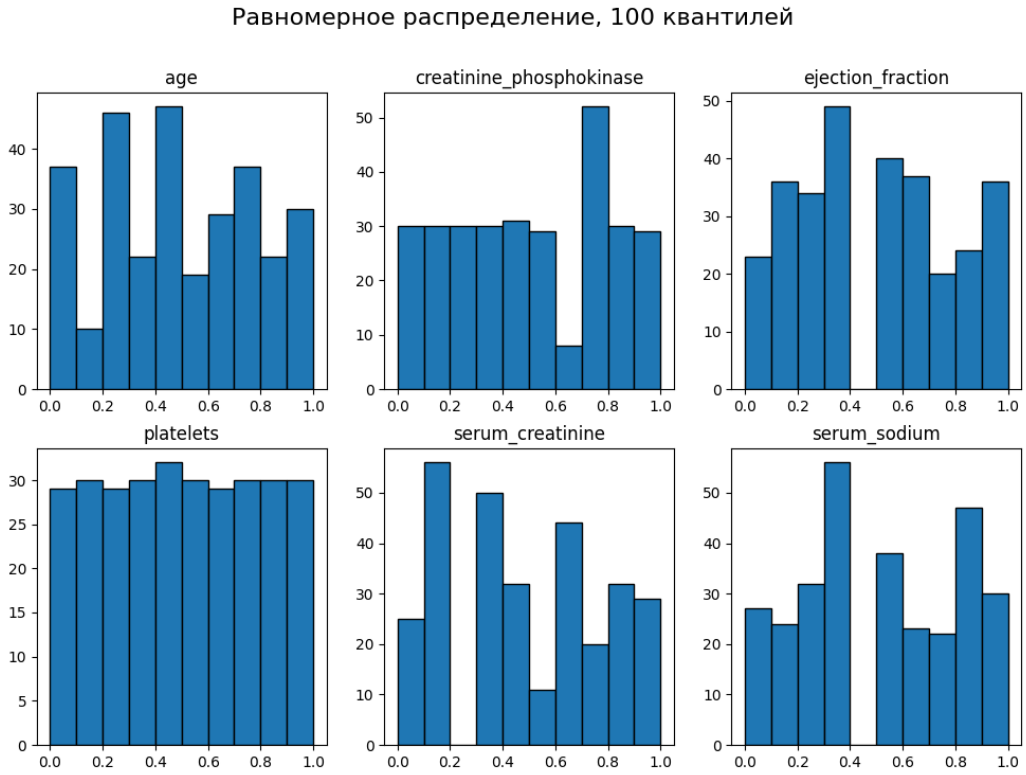

Рисунок 10

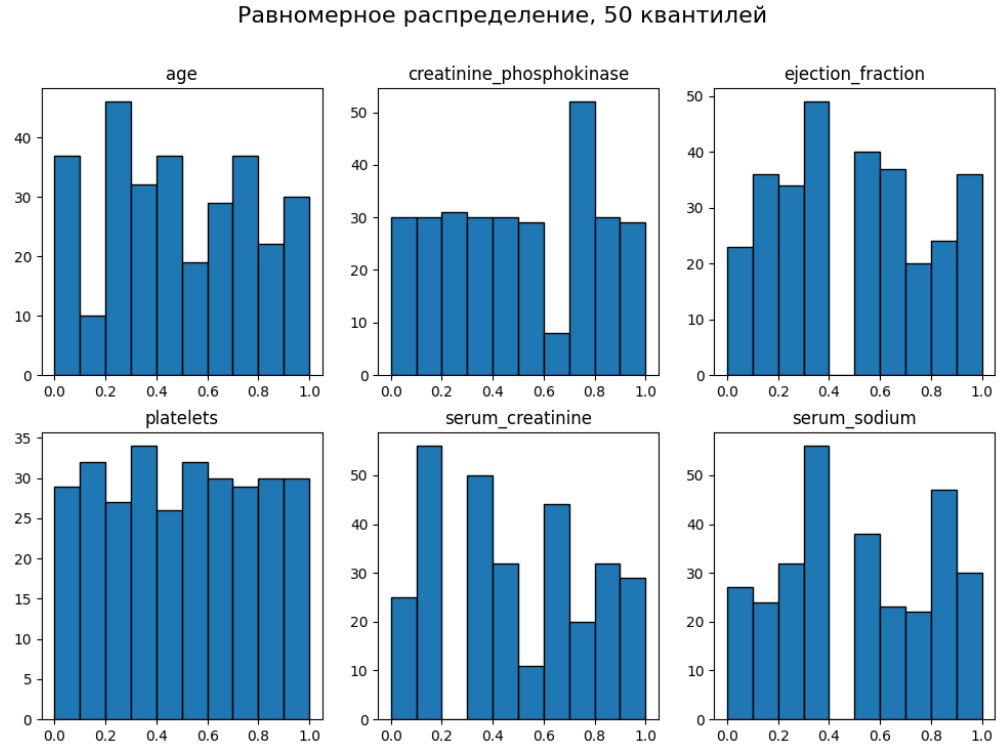

Рисунок 11

Параметр n_quantiles определяет количество квантилей, используемых для преобразования, чем больше квантилей, тем выше частота дискретизации функции распределения.

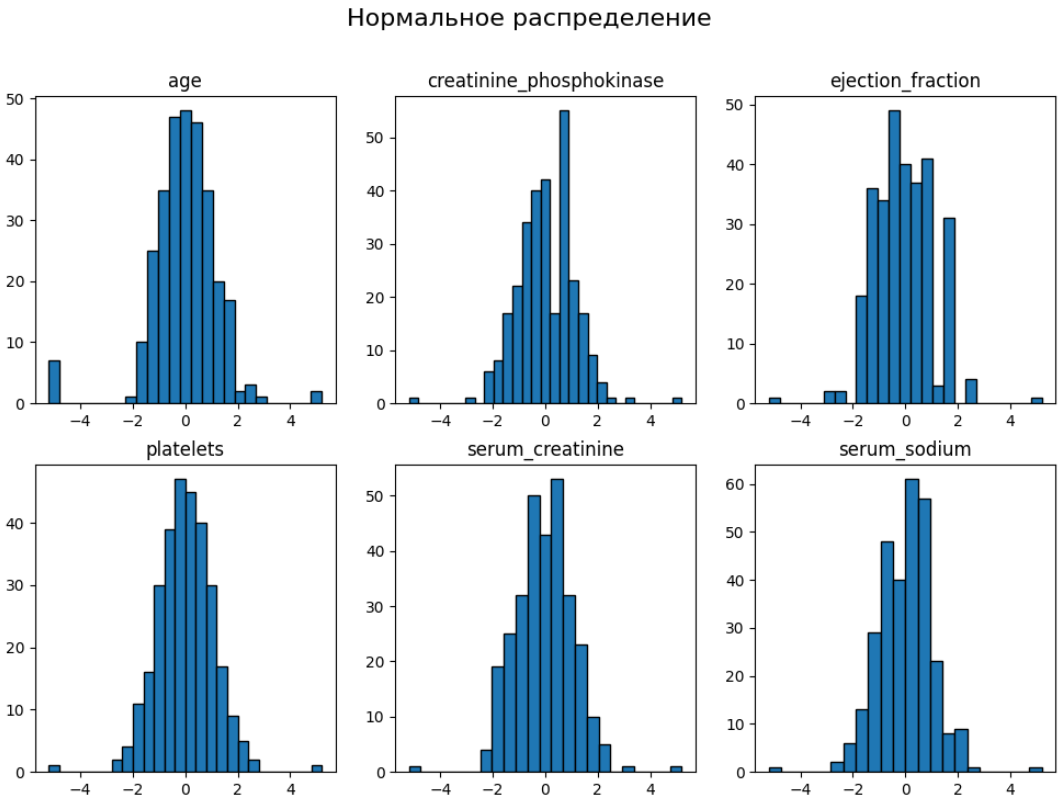

Рисунок 12.

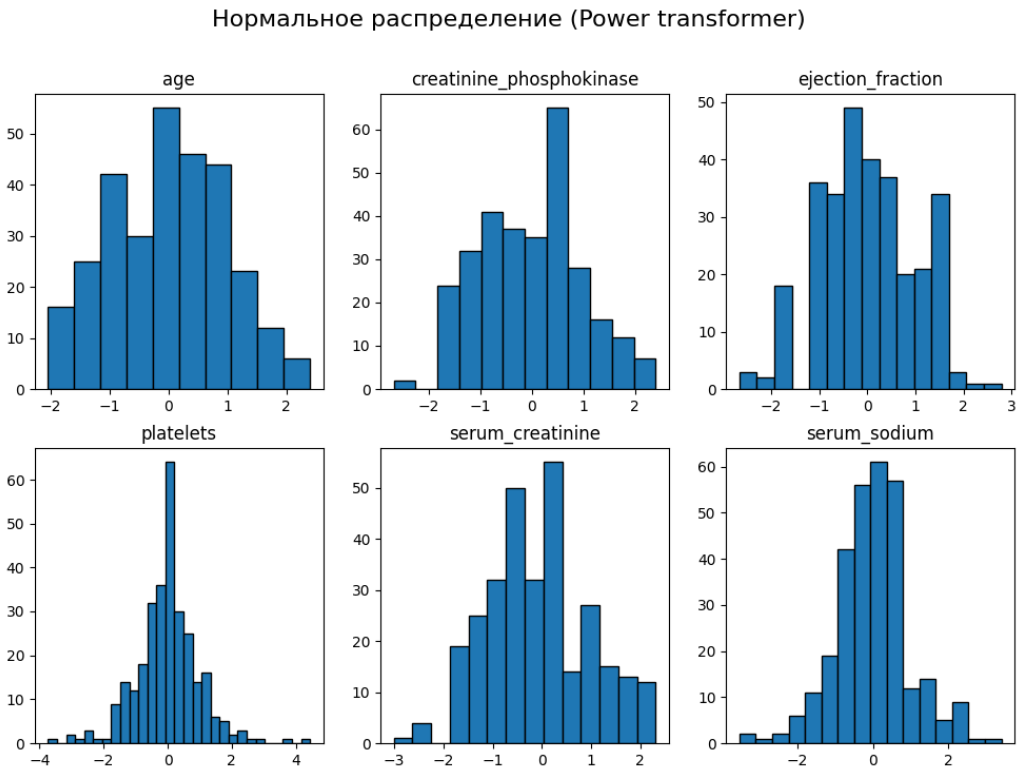

Рисунок 13

5. Дискретизация признаков

Для дискретизации признаков был использован KBinsDescretizer. Гистограмма полученных значений представлена на рисунке 14.

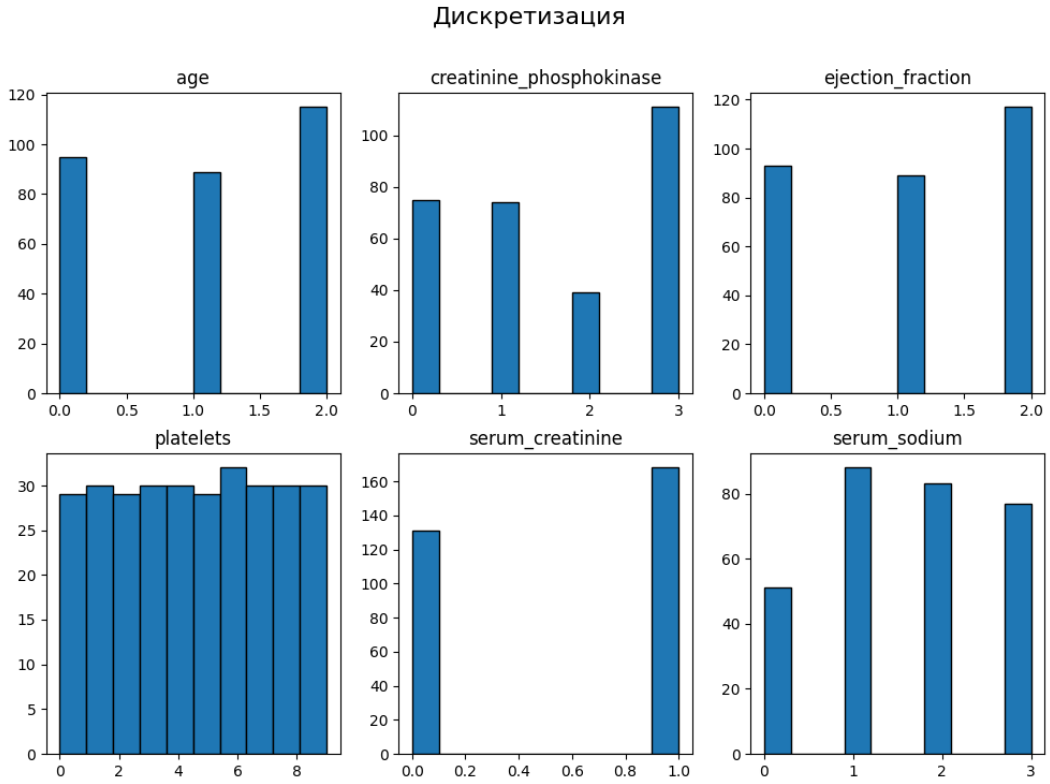

Рисунок 14.

Поскольку при использовании дискретизации значения – числовые обозначения классов, данные гистограммы не являются показательными, так не несут в себе какого-то смысла.

С помощью параметра bin_edges_ объекта KBinsDiscretizer получаем границы диапозонов:

```
Края диапазонов:
age: [ 40.0 55.0 65.0 95.0 ]
creatinine_phosphokinase: [ 23.0 116.5 250.0 582.0 7861.0 ]
ejection_fraction: [ 14.0 35.0 40.0 80.0 ]
platelets: [ 25100.0 153000.0 196000.0 221000.0 237000.0 262000.0 265000.0 285200.0 319800.0 374600.0 850000.0 ]
serum_creatinine: [ 0.5 1.1 9.4 ]
serum_sodium: [ 113.0 134.0 137.0 140.0 148.0 ]
```

Вывод

В результате выполнения рабораторной рабты были изучены различные методы предобработки данных библиотеки Scikit Learn. Поскольку python раньше мной не изучался, во время выполнения работы возникало множество сложностей, связанных с незнанием синтаксиса и основных функций. В ходе работы было установлено, что стандартизация по неполной выборке снижает качество стандартизации; приведение к диапазону позволяет изменять границы данных без изменения формы распределения и , наоборот, нелинейные преобразования позволяют изменить форму распределения данных.