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»

Студент гр. 6304 _____ Антонов С.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Цель работы:

Ознакомиться с методами обработки данных из библиотеки Scikit Learn.

Ход работы:

Загрузка данных

1. На данном этапе был скачан и загружен датасет в датафрейм, а также исключены его бинарные признаки и признаки времени.

```
df = pd.read_csv('heart_failure_clinical_records_dataset.csv')
df = df.drop(columns=['anaemia', 'diabetes', 'high_blood_pressure', 'sex',
'smoking', 'time', 'DEATH_EVENT'])
```

```
      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 Загруженный датасет

2. Были построены гистограммы признаков.

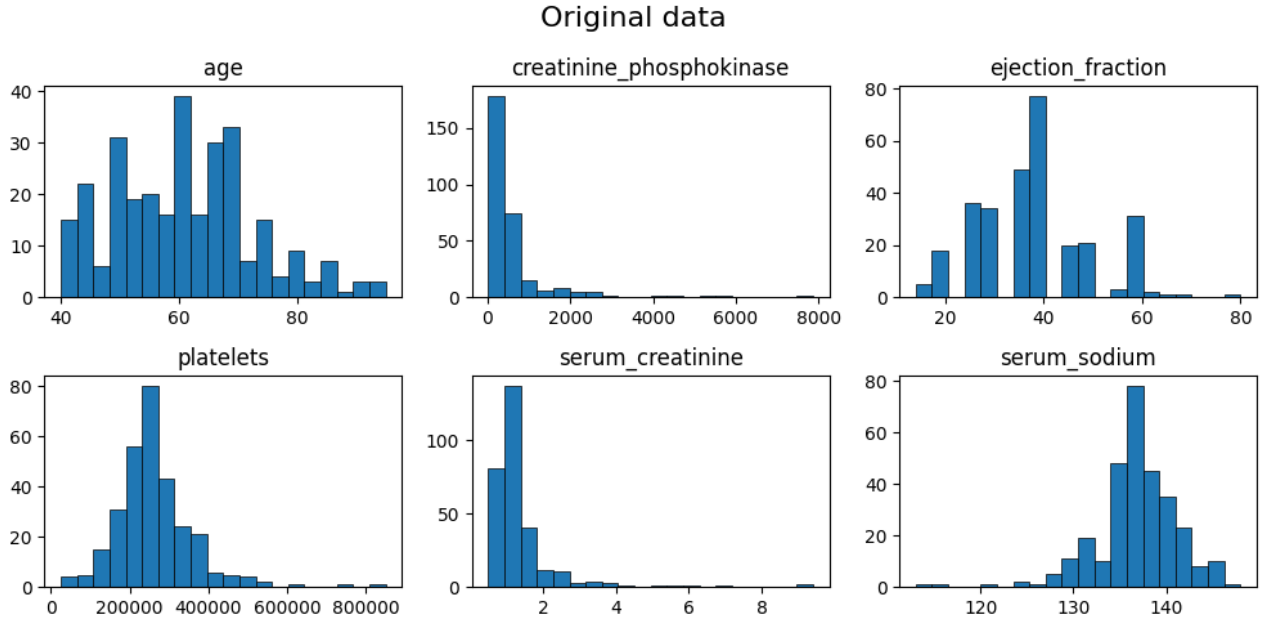

Рисунок 2 Гистограммы признаков

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

Признаки	Диапазон	Значение с наибольшим кол-вом наблюдений
Age	[40, 100]	60
Creatinine_phosphokinase	[0, 8000]	200
Ejection_faction	[10, 80]	38
Platelets	[0, 875] * 10^3	250 * 10^3
Serum_creatinine	[0.1, 9.75]	1.2
Serum_sodium	[110, 150]	137

4. Выполнено преобразование датафрейма к двумерному миассиву Numpy.

```
data = df.to_numpy(dtype='float')
```

Стандартизация данных

1. Был подключен модуль Sklearn и выполнена стандартизация данных с помощью StandartScaler на основе первых 150 наблюдений. Гистограммы стандартизированных данных представлены на рисунке 3.

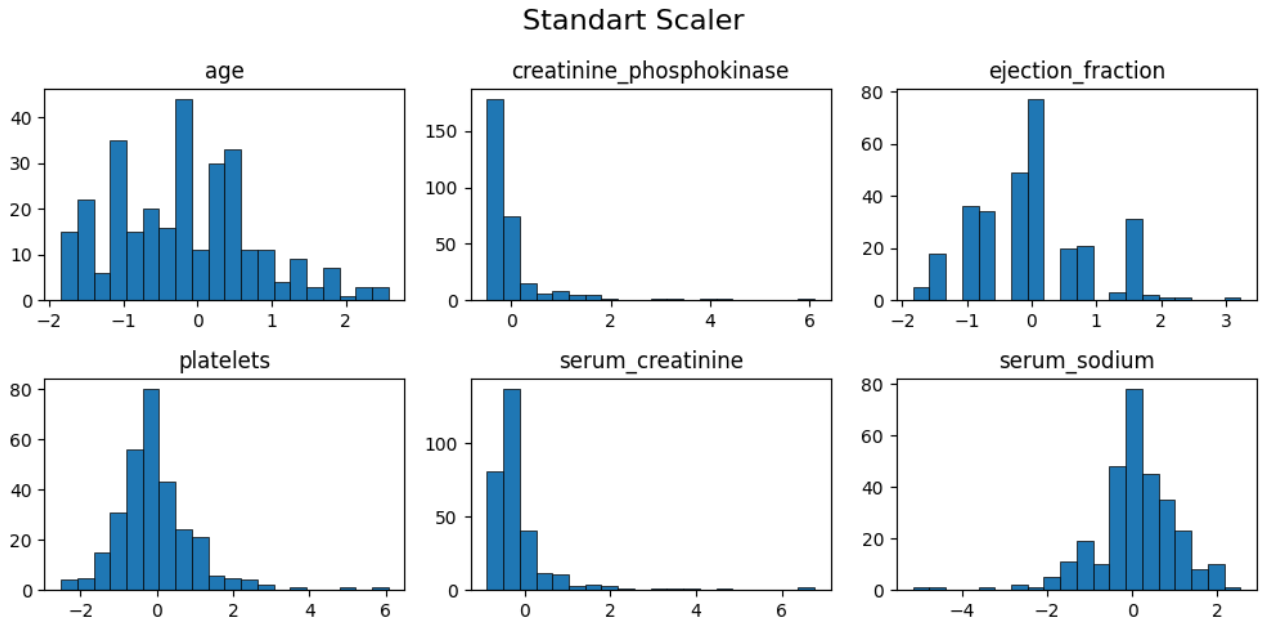

Рисунок 3 Гистограмма стандартизированных признаков

Проводя сравнение обычных и стандартизированных гистограмм, можно сделать вывод, что изменились диапазон и значение с наибольшим количеством наблюдений.

2. Была проведена стандартизация на полном наборе данных и вычислено Мат. Ожидание и СКО для каждого признака из 3 выборок.

Выборка	M / σ	Age	Creatinine_phosphokinase	Ejection_faction	Platelets	Serum_creatinine	Serum_sodium
Original	M	60.833	581.839	38.085	263e3	1.394	136.625
	σ	11.89	970.288	11.836	97e3	1.035	4.412
Standart 150	M	-0.170	-0.021	0.011	-0.036	-0.109	0.038
	σ	0.955	0.816	0.908	1.018	0.887	0.972
Standart full	M	5.703353e-16	0.0e+00	-3.269e-17	7.724e-17	1.426e-1	-8.674e-16
	σ	1.001676e+00	1.002e+00	1.002e+00	1.002e+00	1.002e+00	1.002e+00

На основании приведенных результатов можно сделать вывод, что StandartScaler центрирует результаты относительно дисперсии. Приведем формулу:

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

где X – исходные данные, Y – результат.

3. В поля mean_ и var_ объекта StandartScaler записываются мат. ожидание и дисперсия величин, на основе которых будет проводиться стандартизация.

Приведение к диапазону

1. Используя MinMaxScaler данные были приведены к диапазону.

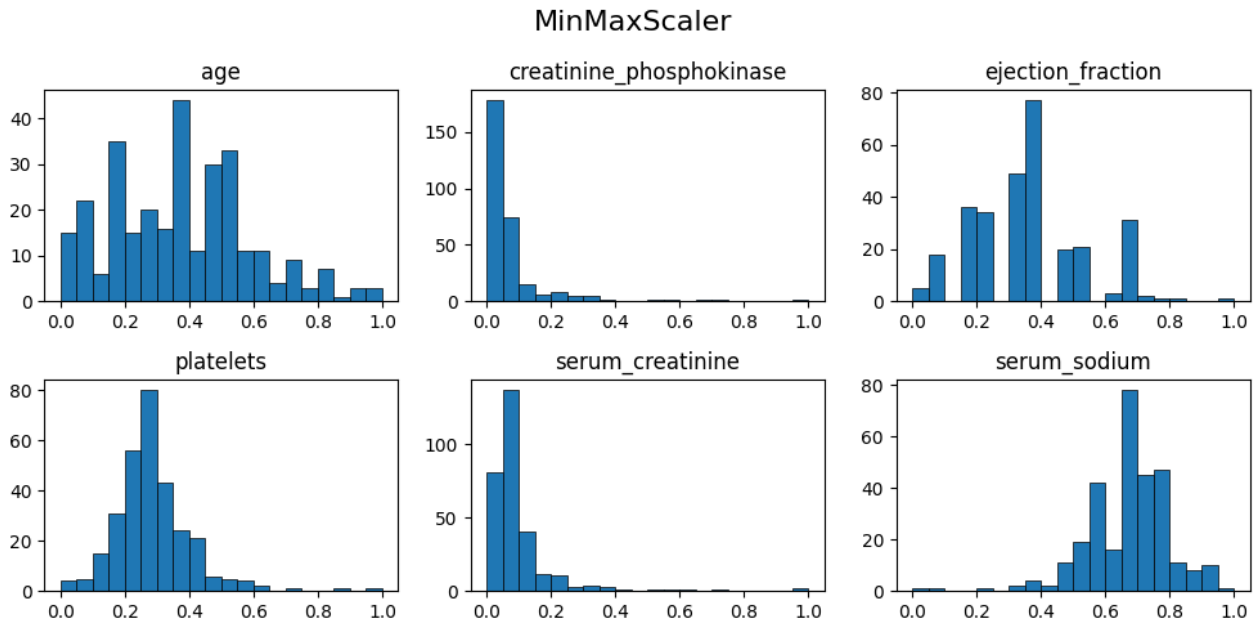

Рисунок 4 Гистограмма после приведения к диапазону

MinMaxScaler масштабирует исходные данные к промежутку [0, 1], применяя формулу:

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

2. Через параметры MinMaxScaler data_min_ и data_max_ были определены минимальные и максимальные значения для каждого признака.

	Age	Creatinine_phosphokinase	Ejection_fraction	Platelets	Serum_creatinine	Serum_sodium
min	4.00e+01	2.30e+01	1.40e+01	2.51e+04	5.00e-01	1.13e+02
max	9.500e+01	7.861e+03	8.000e+01	8.500e+05	9.400e+00	1.480e+02

3. Было выполнено приведение к диапазону с помощью MaxAbsScaler и RobustScaler.

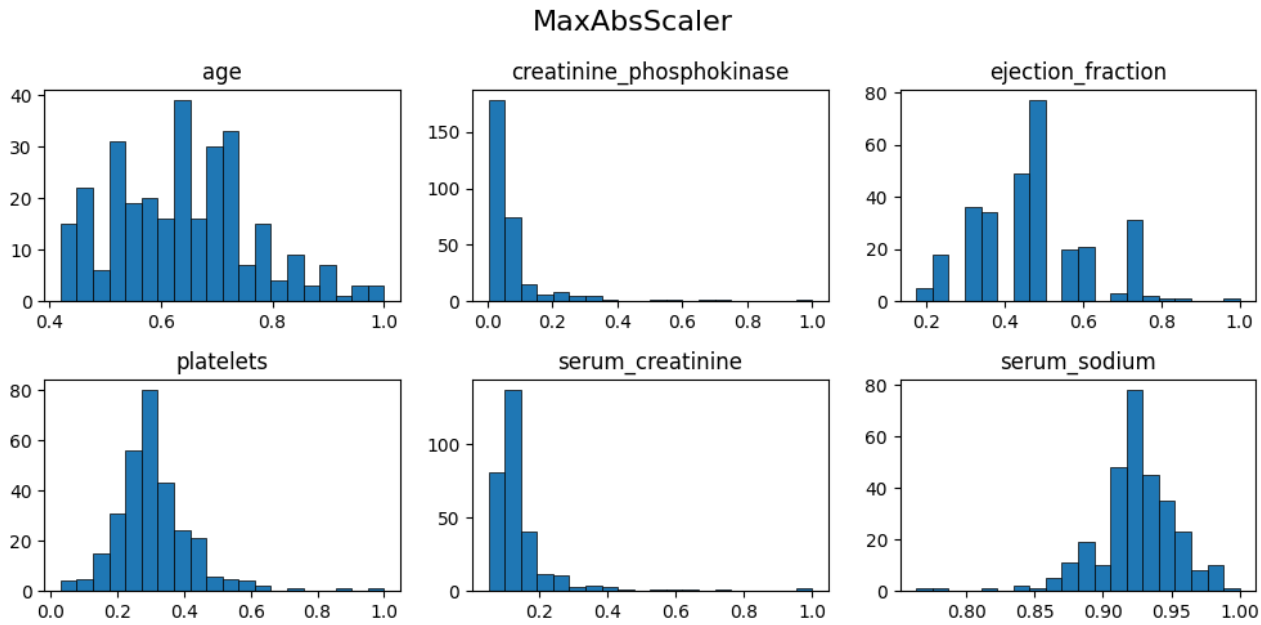

Рисунок 5 Гистограмма после MaxAbsScaler

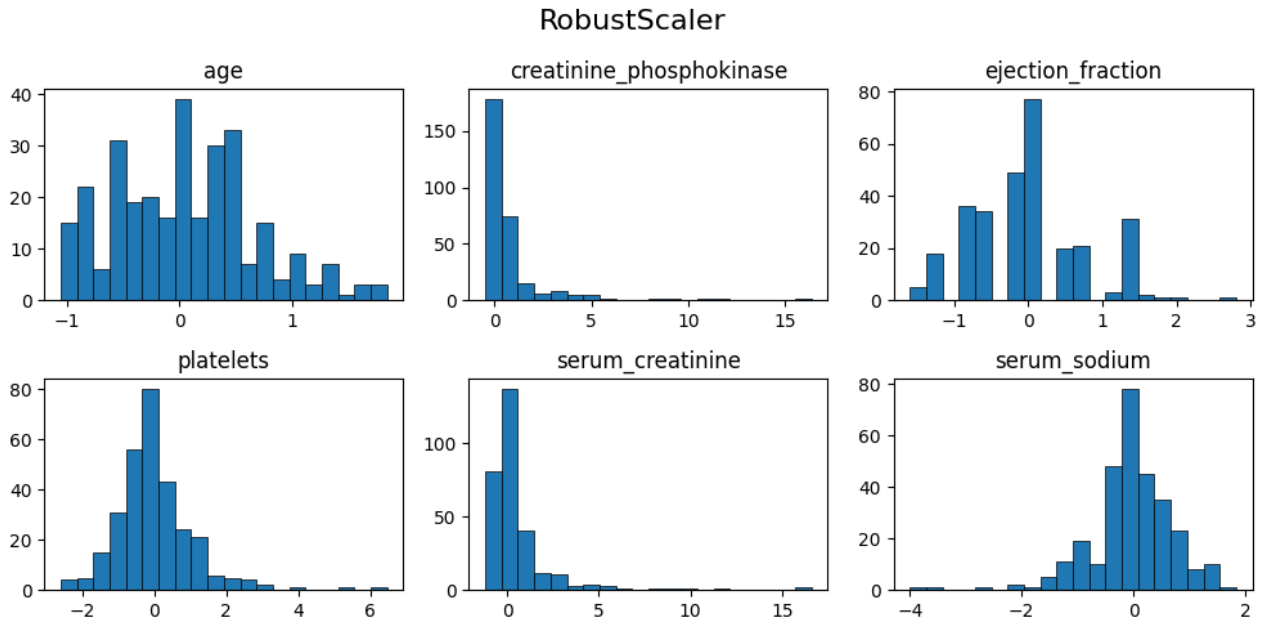

Рисунок 6 Гистограмма после RobustScaler

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

4. Была написана функция, которая приводит все данные к диапазону [-5 10].

```
def custom_range(data):
    custom_data = preprocessing.MinMaxScaler().fit_transform(data)*15-5
    return custom_data
```

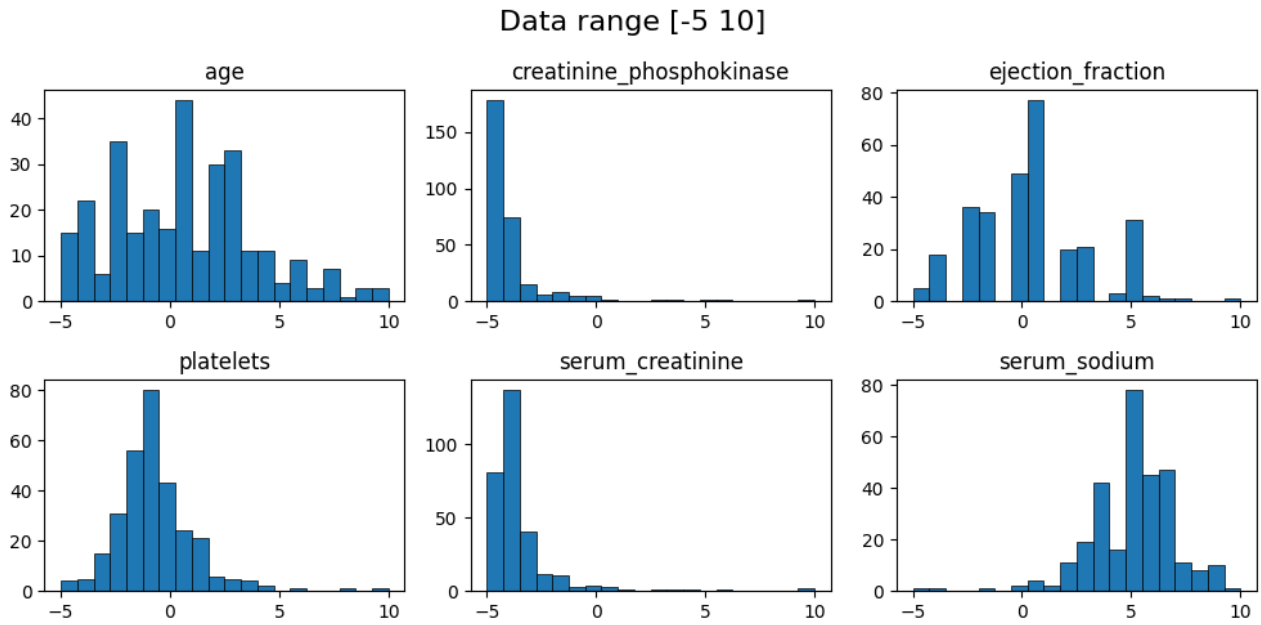

Рисунок 7 Гистограмма после применения собственной функции

Нелинейные преобразования

1. С помощью QuantileTransformer данные были приведены к равномерному и нормальному распределениям.

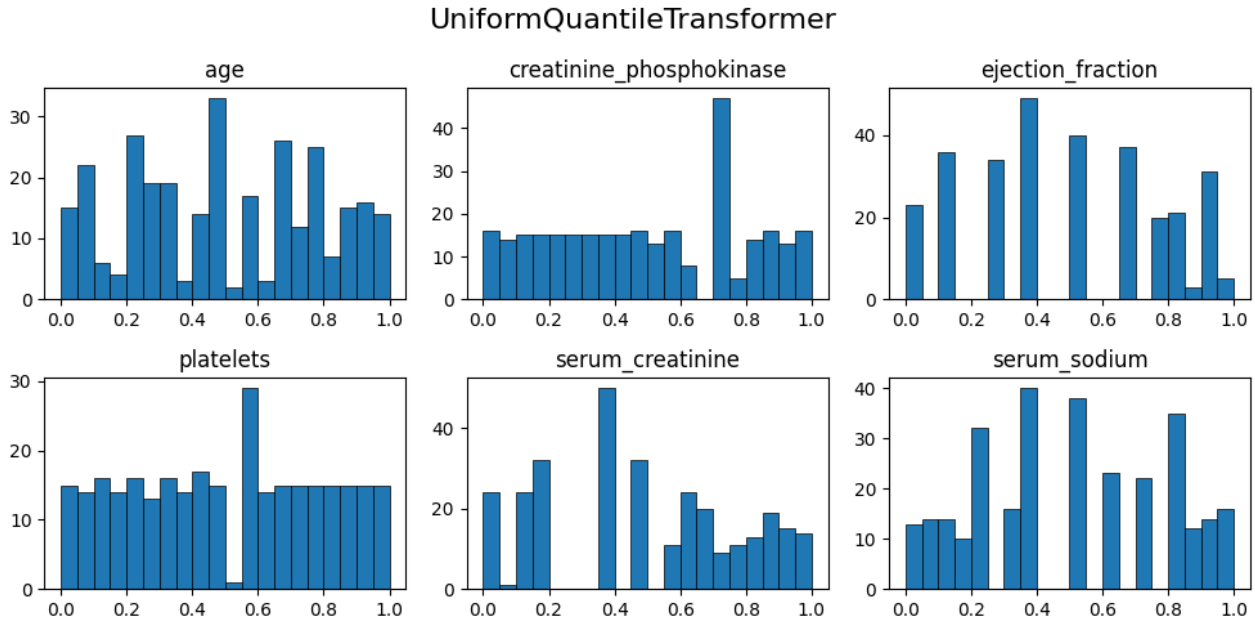

Рисунок 8 QuantileTransformer, равномерное распределение

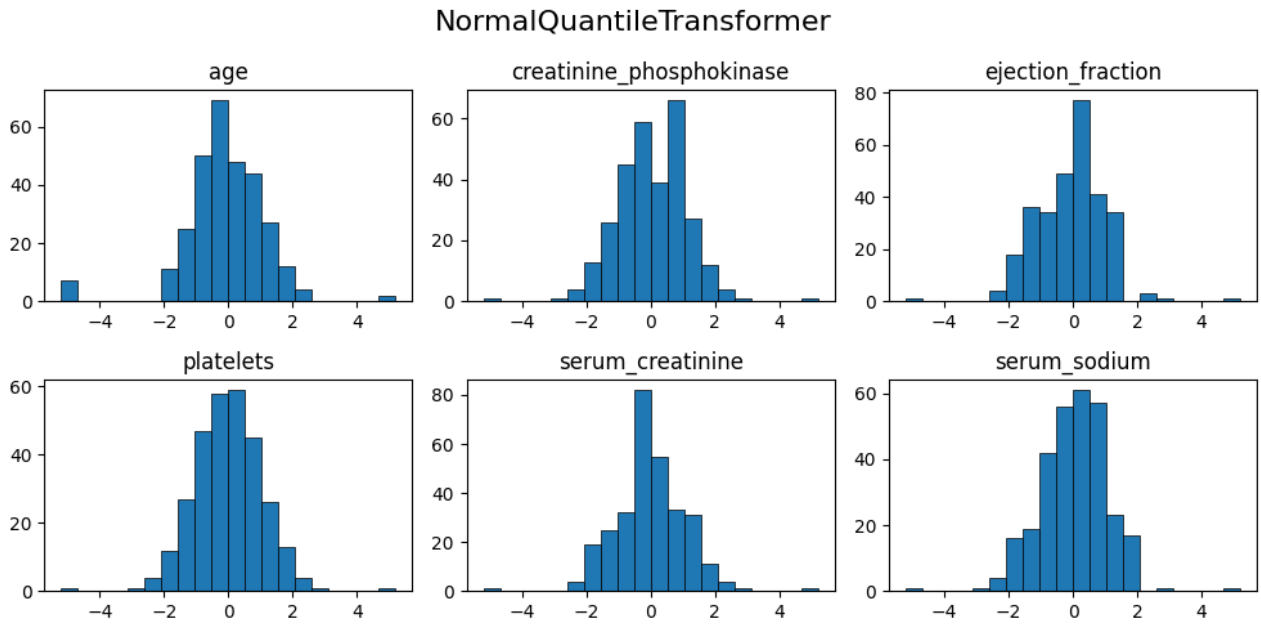

Рисунок 9 QuantileTransformer, нормальное распределение

Параметр n_quantiles определяет количество квантилей, используемых для дискретизации функции распределения. Чем больше квантилей (не больше количества наблюдений), тем ближе к требуемому распределению.

2. Данные приведены к нормальному распределению через PowerTransformer

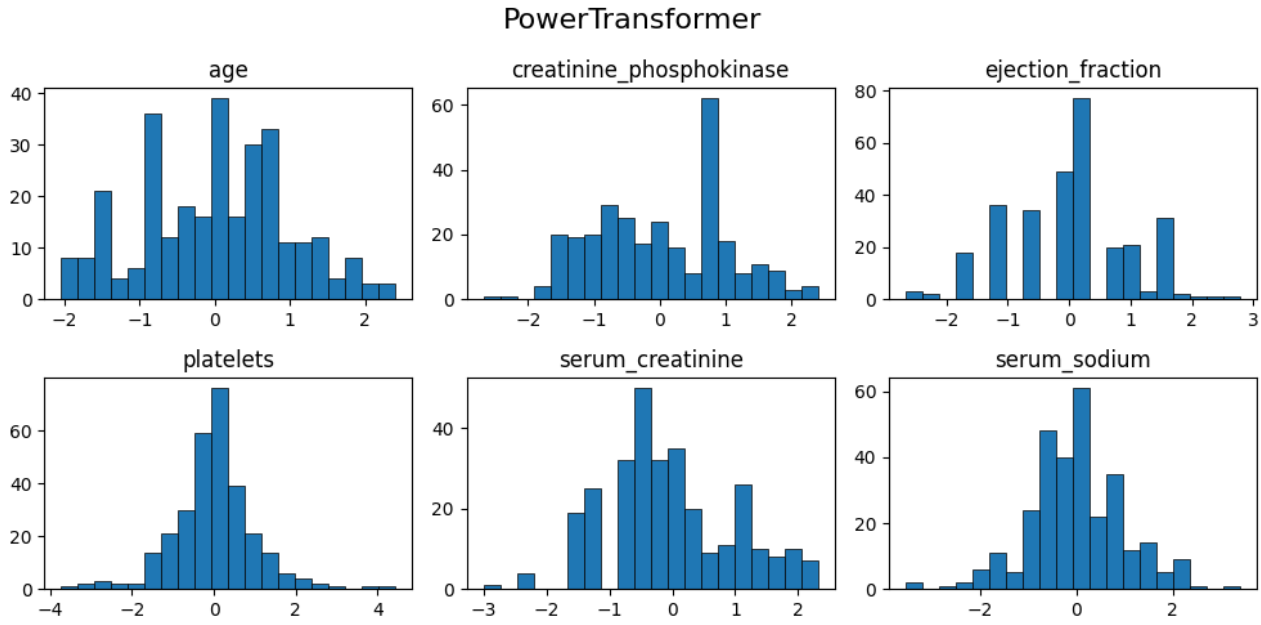

Рисунок 10 Гистограмма после PowerTransformer

Дискретизация признаков

1. Для дискретизации признаков был использован KBinsDescretizer.

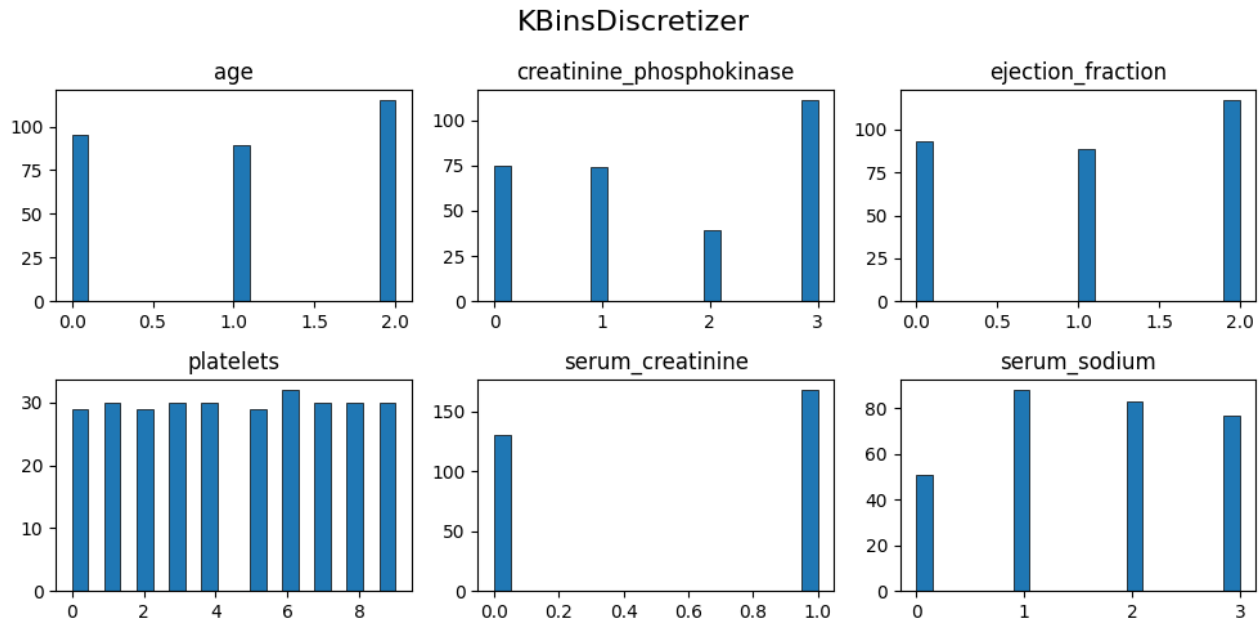

Рисунок 11 Гистограмма признаков после дискретизации

Диапазоны интервалов:

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.])

Выводы:

В результате выполнения лабораторной работы было проведено ознакомление с методами предобработки данных с помощью библиотеки Sciti Learn.

Применяя стандартизацию данных, было установлено, что стандартизация на неполной выборке снижает качество стандартизации.

При приведении данных к диапазону изменяются лишь границы диапазона наблюдений без изменения формы распределения.

Нелинейные преобразования, наоборот, изменяют саму форму распределения. В лабораторной работе было проведено преобразование к нормальной и равномерной форме.

Также была проведена дискретизация данных, которая позволяет разбивать данные на классы.