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 Доброхвалов М. О.

Преподаватель Жангиров Т. Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами предобработки данных из библиотеки Scikit Learn

Ход работы

Загрузка данных

1. Скачан и загружен датасет в датафрейм. Исключены бинарные признаки и признаки времени (рис. 1).

```
df =
pd.read_csv('heart_failure_clinical_records_dataset.csv').drop(c
olumns =
['anaemia','diabetes','high_blood_pressure','sex','smoking','tim
e','DEATH_EVENT'])
print(df)
```

	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
...	...	...	...	...	...	...
294	62.0	61	38	155000.00	1.1	143
295	55.0	1820	38	270000.00	1.2	139
296	45.0	2060	60	742000.00	0.8	138
297	45.0	2413	38	140000.00	1.4	140
298	50.0	196	45	395000.00	1.6	136

299 rows × 6 columns

Рис. 1 — Загруженный датасет

2. Выполнено построение гистограммы признаков (рис. 2). Значения свойства *platelets* разделены на 1000 для более удобного представления.

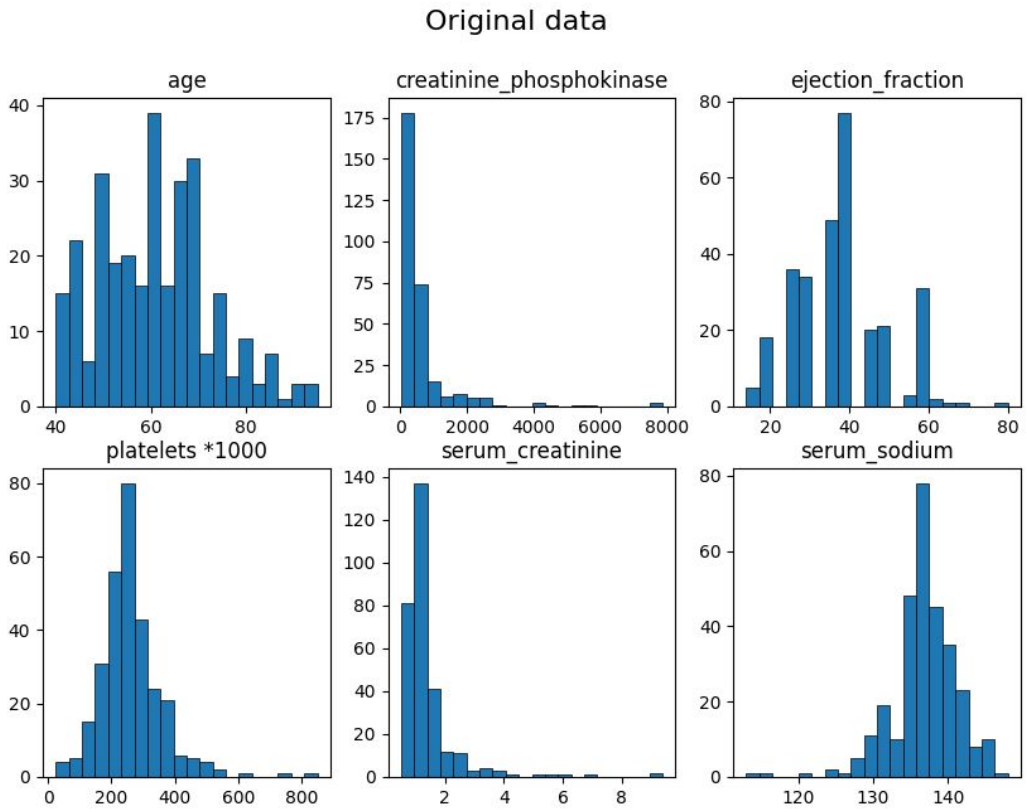

Рисунок 2 — Гистограмма признаков

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

Признак	Диапазон	Значение с наибольшим количеством наблюдений
age	(40, 100)	60
creatinine_phosphokinase	(0, 8000)	200
ejection_fraction	(10, 80)	38
platelets	$(0,\ 875) \cdot 10^3$	$250 \cdot 10^3$
serum_creatinine	(0.1, 9.75)	1.2
serum_sodium	(110, 150)	137

4. Выполнено преобразование датафрейма к формату numpy, т.к. библиотека Sklearn работает с этим форматом

Стандартизация данных

1. Выполнена стандартизация всех наблюдений на основе первых 150. После чего были построены гистограммы признаков (рис. 3)

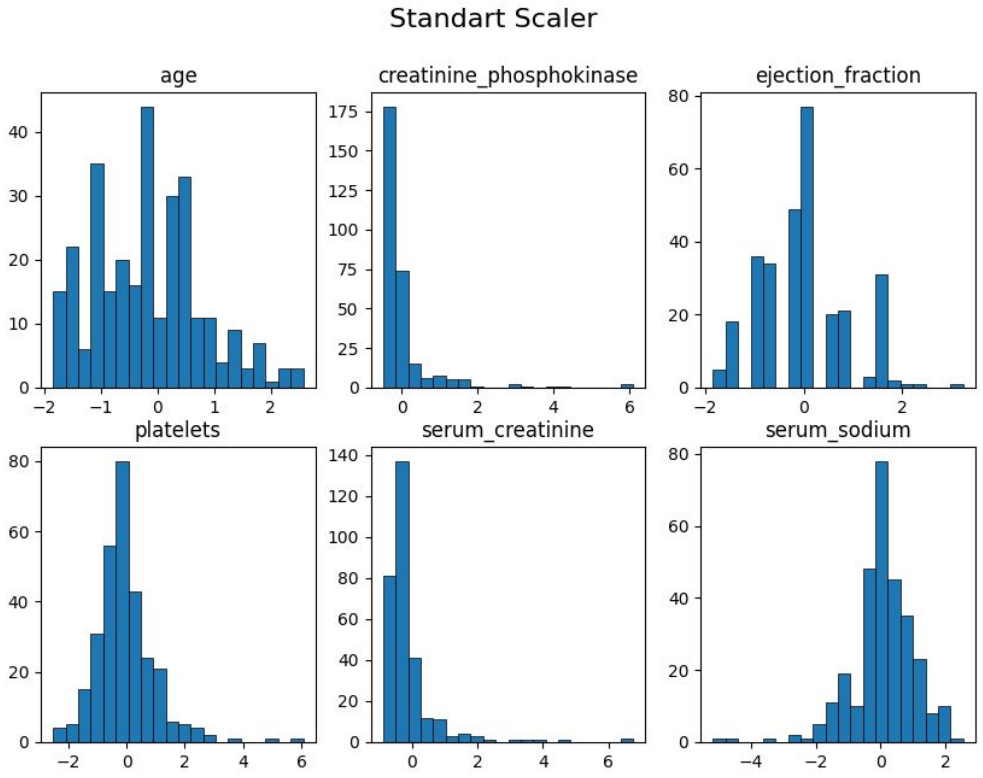

Рисунок 3 — Гистограмма стандартизированных признаков

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

Признак	Диапазон	Значение с наибольшим количеством наблюдений
age	(-2, 2.5)	-0.1
creatinine_phosphokinase	(-0.5, 6.2)	-0.2

ejection_fraction	(-2, 3.5)	0
platelets	(-3, 6.25)	0
serum_creatinine	(-1.5, 7)	-0.2
serum_sodium	(-5.5, 3)	0

Диапазон и значение с наибольшим количеством наблюдений изменились. Причиной является примененное преобразование. Вероятная формула будет приведена в пункте 4.

3. Была проведена стандартизация на полном наборе наблюдений

```
full_scaler = preprocessing.StandardScaler()
full_data_scaled = full_scaler.fit_transform(data)
```

4. Вычислено мат. ожидание и СКО каждой из 3 выборок.

Выборка	Статистика	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium
Оригинальная	мат. ожид.	60.834	581.839	38.084	263e3	1.394	136.625
	СКО	11.895	970.288	11.835	97e3	1.035	4.412
Стандартизированная на 150	мат. ожид.	-0.170	-0.021	0.011	-0.035	-0.109	0.038
	СКО	0.955	0.816	0.908	1.017	0.887	0.972
Стандартизированная	мат. ожид.	5.703e-16	0.0e+00	-3.268e-17	7.723e-17	1.426e-16	-8.674e-16
	СКО	1.002e+00	1.002e+00	1.002e+00	1.002e+00	1.002e+00	1.002e+00

На основании результатов можно сделать вывод о том, что преобразование имеет следующую форму:

$Y = \frac{X - \mu(X)}{std(X)}$, где $\mu(X)$ - мат. ожидание, а $std(X)$ - СКО.

5. В поля *mean_* и *var_* объекта *StandartScaler* записывается мат. ожидание и дисперсия величин, на основе которых будет производиться стандартизация.

Приведение к диапазону

1. С помощью *MinMaxScaler* выполнено приведение данных к диапазону (рис. 4)

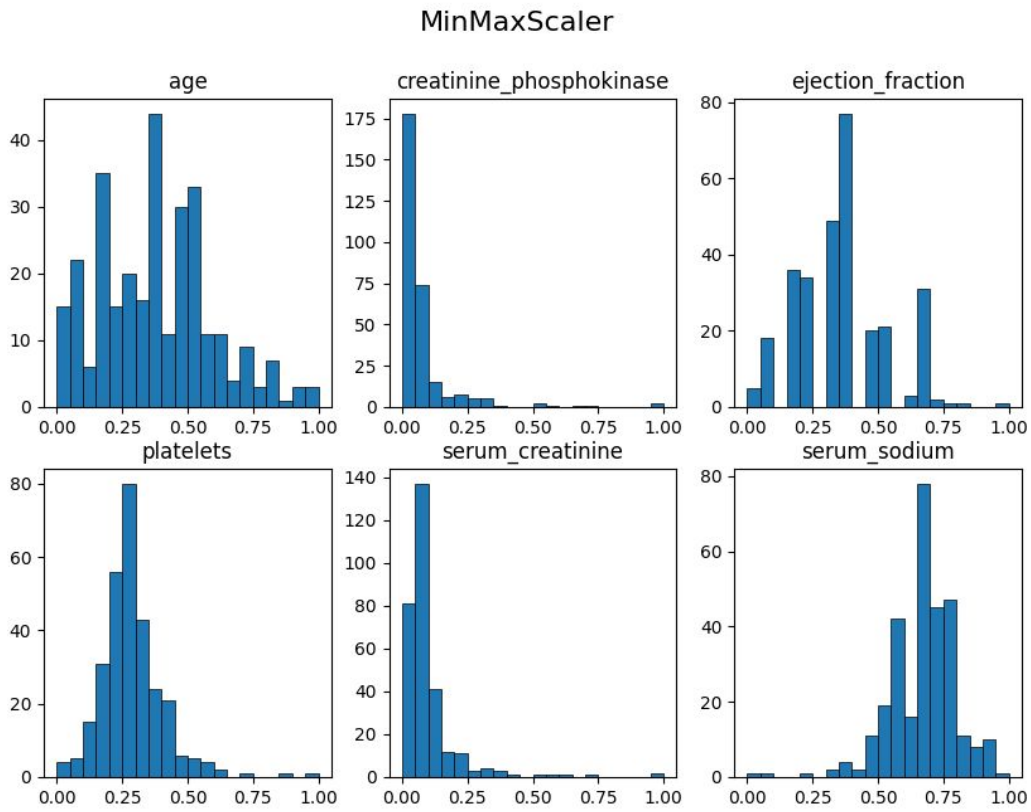

Рисунок 4 — Гистограмма после MinMaxScaler

Судя по гистограммам данные приводятся к диапазону [0,1]. Вероятным способом является следующая формула.

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

2. С помощью объекта *MinMaxScaler* были определены минимальное и максимальное значения каждого признака

	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium
мин.	4.00e+01	2.30e+01	1.40e+01	2.51e+04	5.00e-01	1.13e+02
макс.	9.500e+01	7.861e+03	8.000e+01	8.500e+05	9.400e+00	1.480e+02

3. С помощью *MaxAbsScaler* и *RobustScaler* выполнено приведение данных к диапазону (рис. 5 - 6)

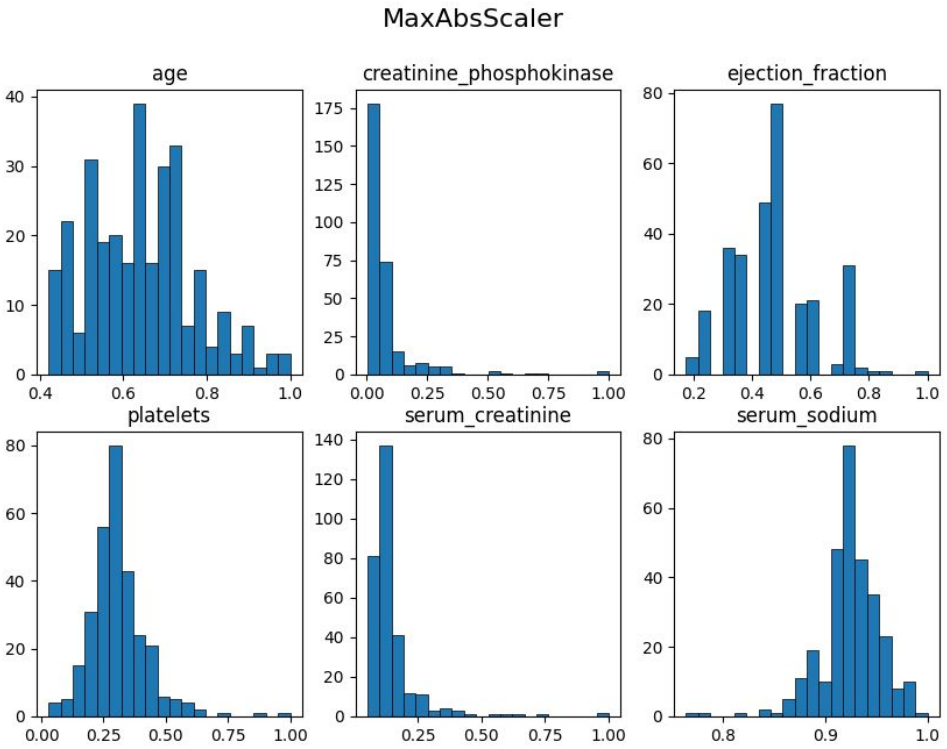

Рисунок 5 — Гистограмма после MaxAbsScaler

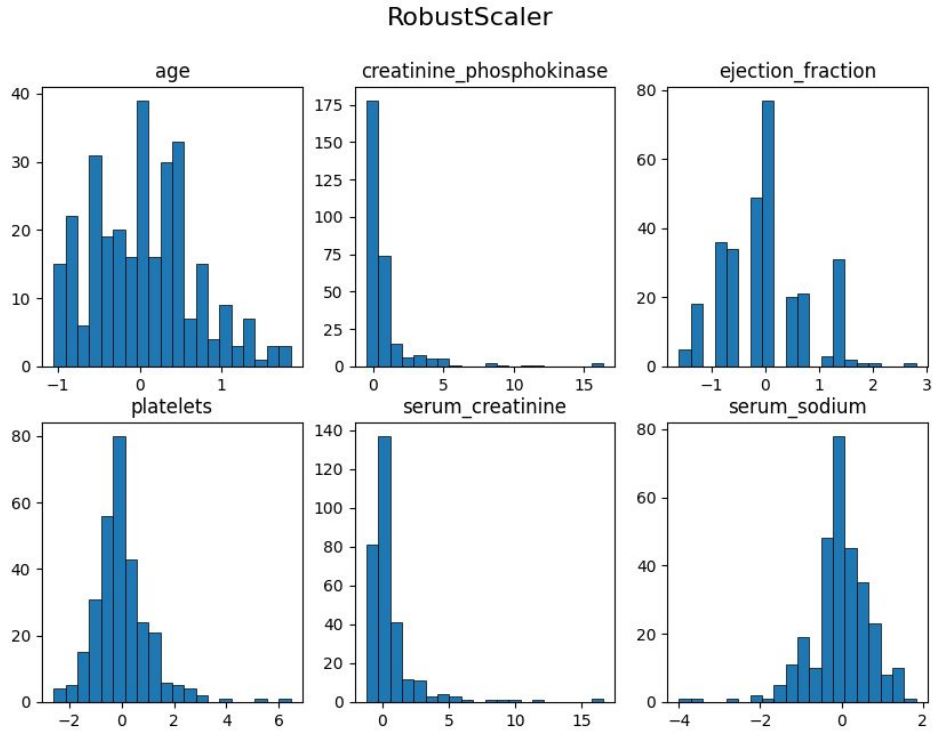

Рисунок 6 — Гистограмма после RobustScaler

MaxAbsScaler приводит данные таким образом, что максимальное по модулю значение равно 1. *RobustScaler* вычитает медиану и масшабирует по в соответствии с межквартильным размахом.

4. Также была написана функция, которая приводит данные к диапазону [-5,10].

```
def range_5_10(data):
    custom_scaler = preprocessing.MinMaxScaler().fit(data)
    return custom_scaler.transform(data)*15-5
```

Результат на рис. 7.

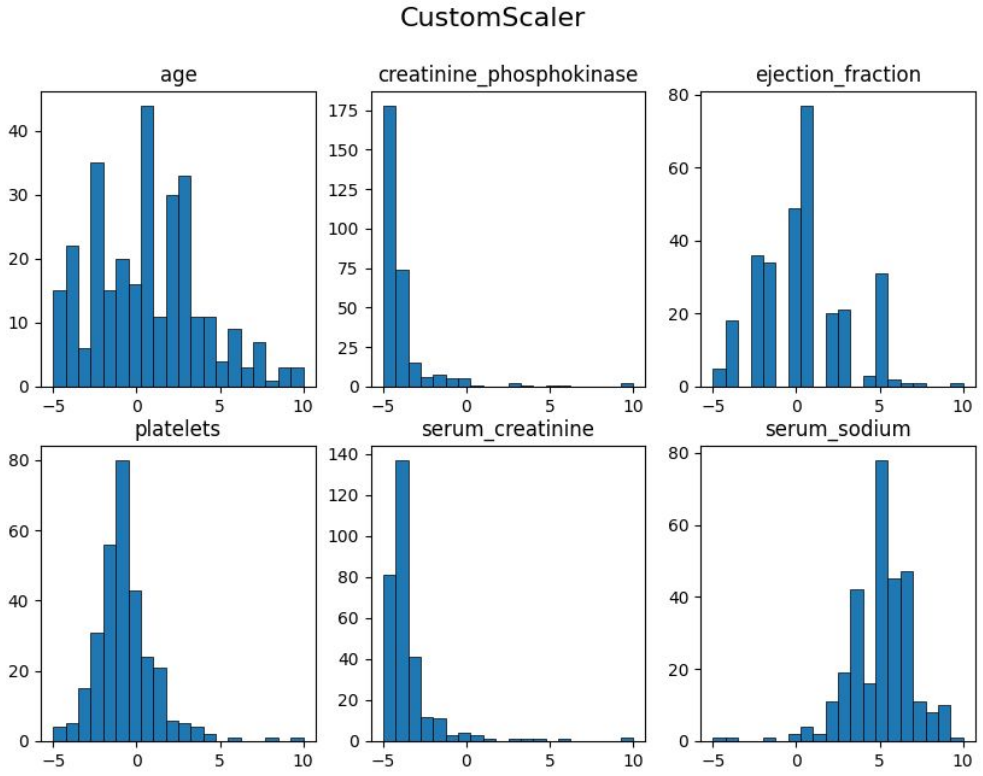

Рисунок 7 — Гистограмма после range_5_10

Нелинейные преобразования

1. С помощью QuantileTransformer данные были приведены к равномерному и нормальному распределениям (рис. 8-9).

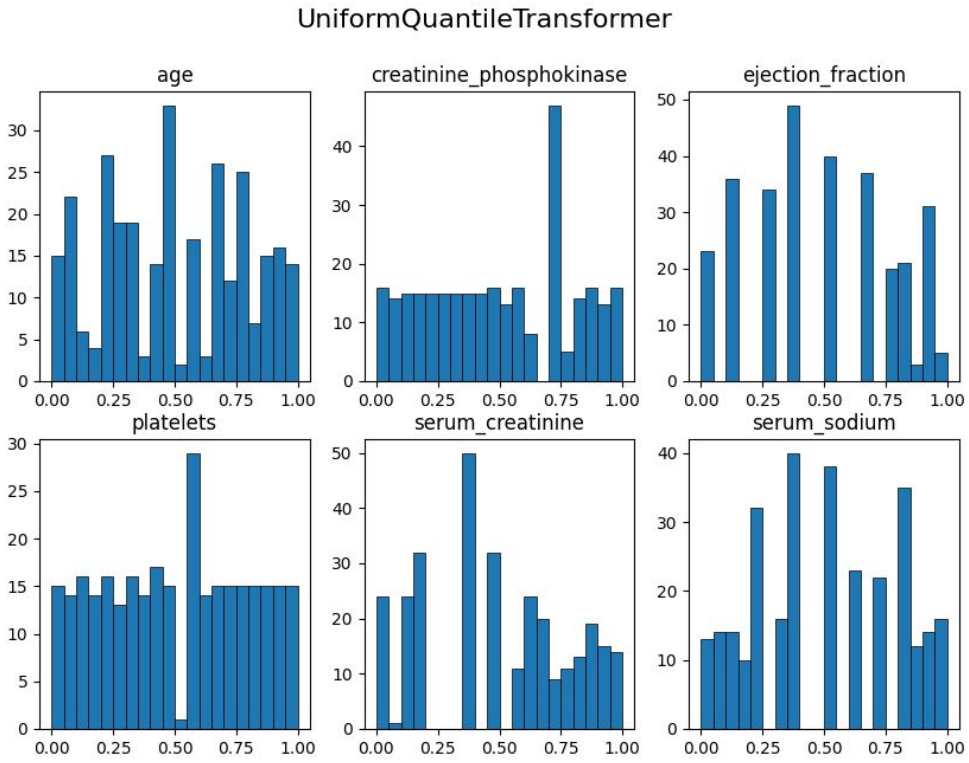

Рисунок 8 — Гистограмма после QuantileTransformer, равномерное распределение

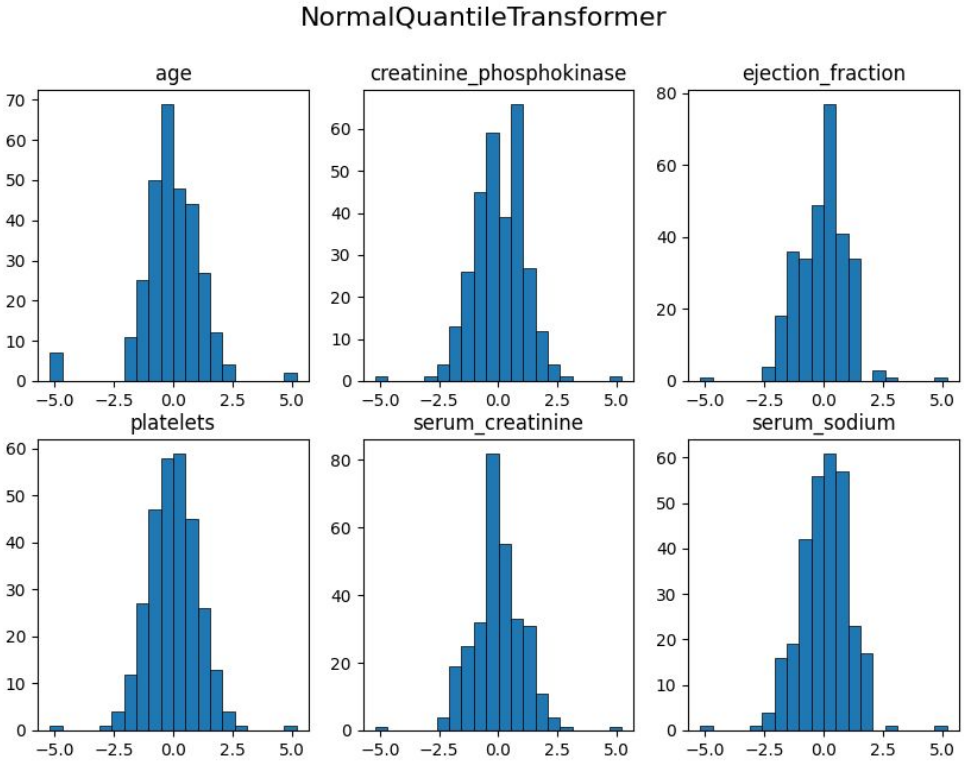

Рисунок 9 — Гистограмма после QuantileTransformer, нормальное распределение

Количество квантилей, используемых для дискретизации функции распределения. Чем больше количество квантилей (но не больше, чем количество наблюдений), тем ближе к требуемому распределению бу

2. С помощью *PowerTransformer* данные были приведены к нормальному распределениям (рис. 10).

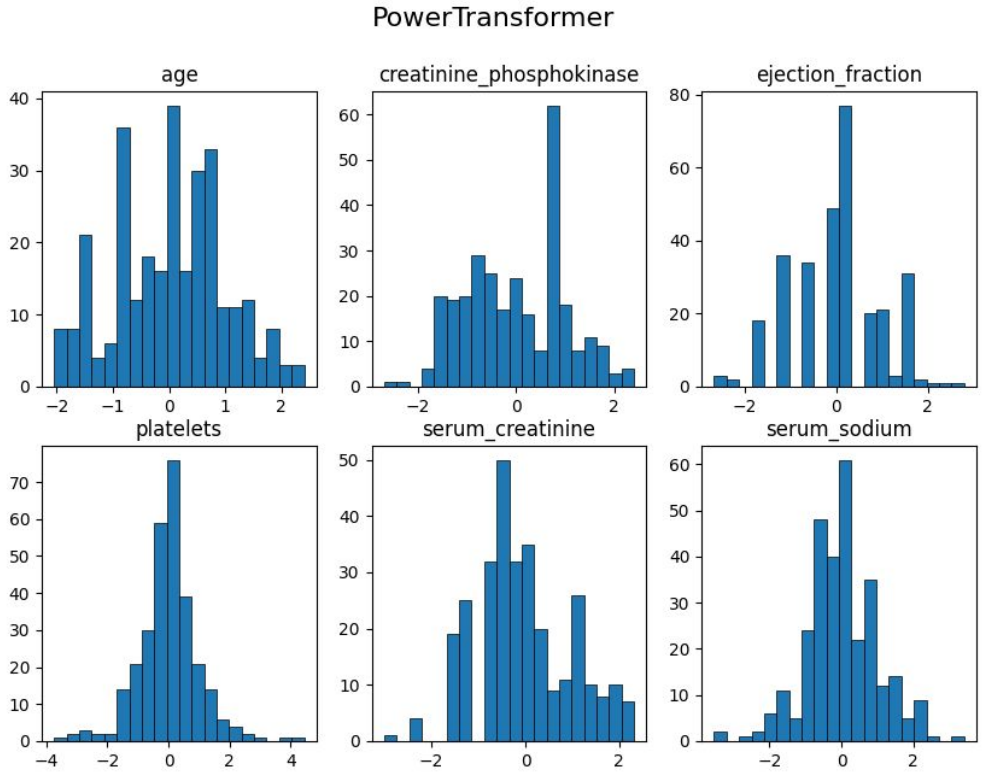

Рисунок 10 — Гистограмма после PowerTransformer

Дискретизация признаков

1. Была выполнена дискретизация признаков(рис. 11)

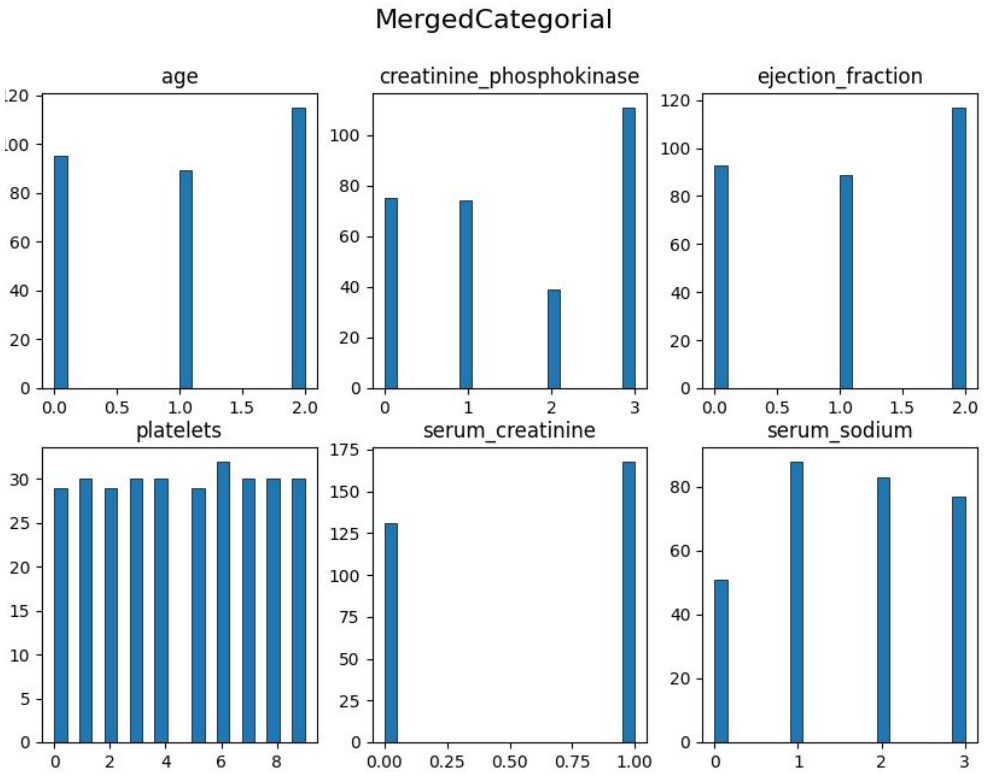

Рисунок 10 — Гистограмма после PowerTransformer

Диапазоны 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.]

Выводы

Было проведено ознакомление с методами предобработки данных с помощью методов библиотеки Scikit Learn.

После изучения стандартизации данных был сделан вывод, что при настройке на неполных данных происходит снижение качества результирующего набора данных.

После приведения к диапазону форма распределения сохраняется.

Нелинейные преобразования позволяют преобразовать форму распределения, например к равномерному или нормальному.

Также была проведена дискретизация данных.