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. 6307 _____ Михайлов И. Т.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами предобработки данных из библиотеки Scikit Learn.

Ход выполнения работы

Загрузка данных

Данные заданного датасета загружены в датафрейм. Исключены бинарные признаки и признак времени.

	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 x 6 columns]

Рис. 1 - Исходные данные

Гистограммы признаков:

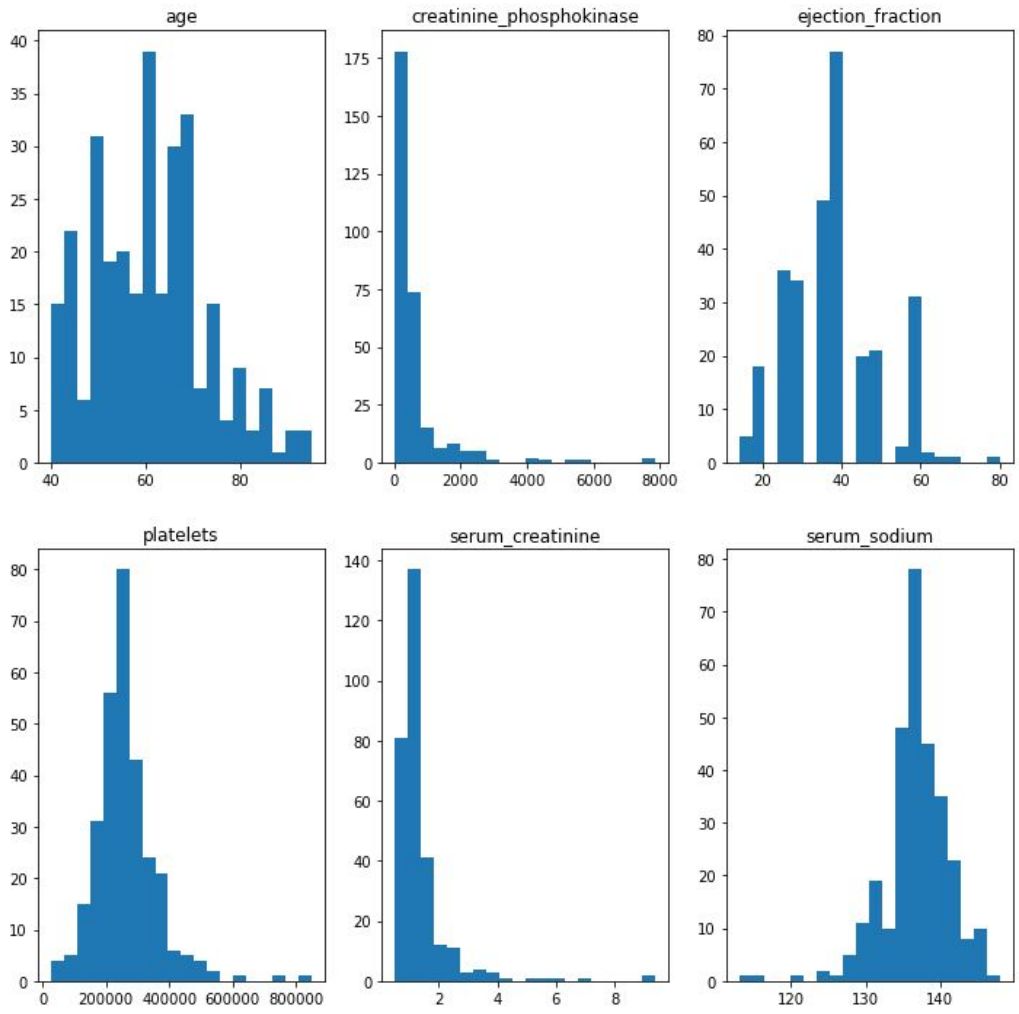


Рис. 2 - Гистограммы признаков

Определим по гистограммам диапазоны значений признаков и значения, у которых лежит наибольшее количество наблюдений.

Название признака	Диапазон значений	Значение, у которого лежит наибольшее число наблюдений
age	(40; 95)	60
creatinine_phosphokinase	(0; 7800)	175
ejection_fraction	(10; 80)	37
platelets	(33000; 840000)	270000
serum_creatinine	(0,4; 9,4)	1
serum_sodium	(112; 149)	137

Таблица 1 - диапазоны значений и наиболее частые значения признаков

Стандартизация данных

2. Данные стандартизированы на основе первых 150 наблюдений.

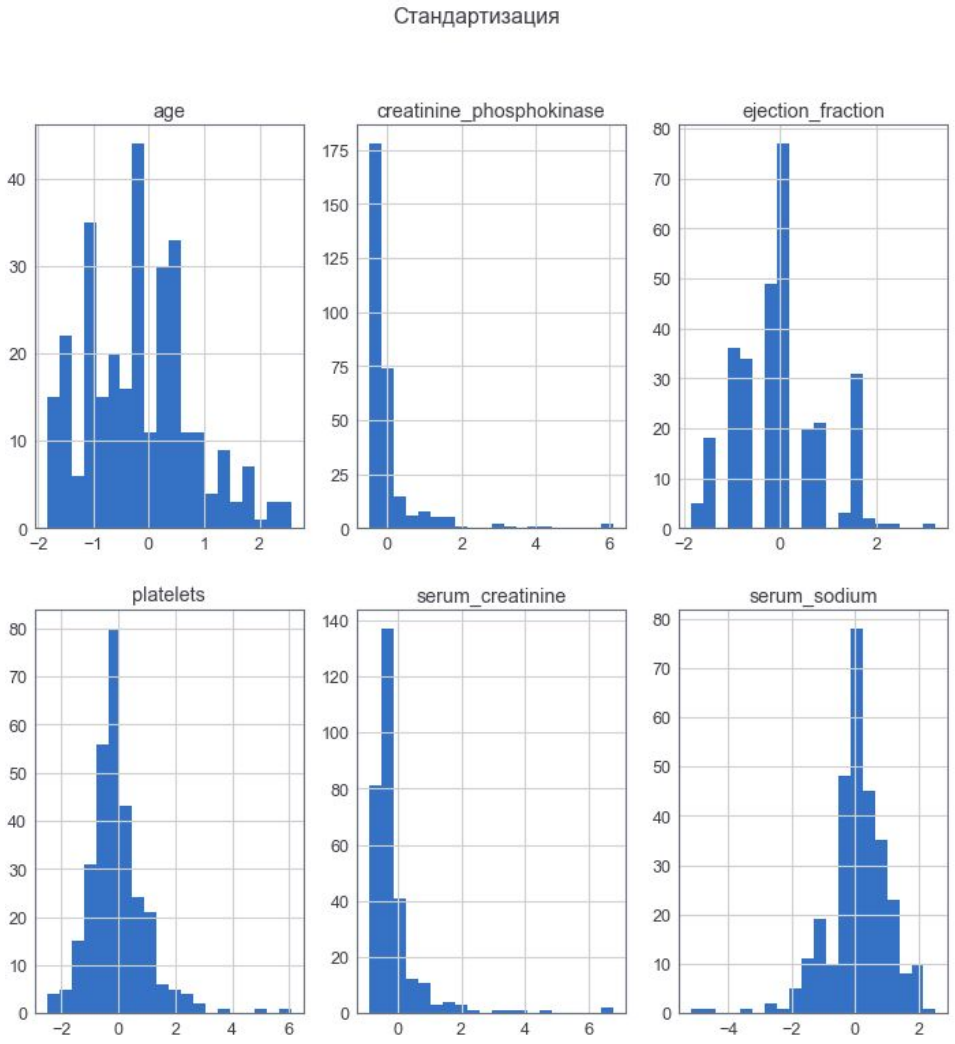


Рис. 2 - Гистограммы стандартизированных признаков

Название признака	Диапазон	Значение, у которого лежит наибольшее число наблюдений
age	(-1,8; 2,6)	-0,2
creatinine_phosphokinase	(-0,8; 6,1)	-0,4
ejection_fraction	(-2; 3)	0
platelets	(-2,5; 6,1)	0

serum_creatinine	(-1; 6,6)	-0,2
serum_sodium	(-5; 2,4)	0

Таблица 2 - диапазоны и наиболее частые значения признаков после стандартизации

4. Диапазоны значений и значения, у которых лежит наибольшее число наблюдений, изменились. Это связано с применением функции StandardScaler - мат ожидание столбцов приблизилось к нулю, а стандартное отклонение к 1.

	До стандарт.	Стандрт. 150	Стандарт. все	До стандарт.	Стандрт. 150	Стандарт. все
	Мат. ожидание	Мат. ожидание	Мат. ожидание	СКО	СКО	СКО
age	60,833	-0.169	5.703e-16	11,875	0.953	1
creatinine_phosphokinase	581,839	-0.021	0.000e+00	968,664	0.814	1
ejection_fraction	38,084	0.011	-3.267e-17	11,815	0.906	1
platelets	263358,029	-0.035	7.723e-17	97640,551	1.015	1
serum_creatinine	1,394	-0.108	1.425e-16	1,033	0.885	1
serum_sodium	136,625	0.038	-8.673e-16	4,405	0.970	1

Таблица 3 - Мат. ожидание и СКО признаков до и после дискретизации

5. Формула, по которой производилась стандартизация имеет следующий вид:

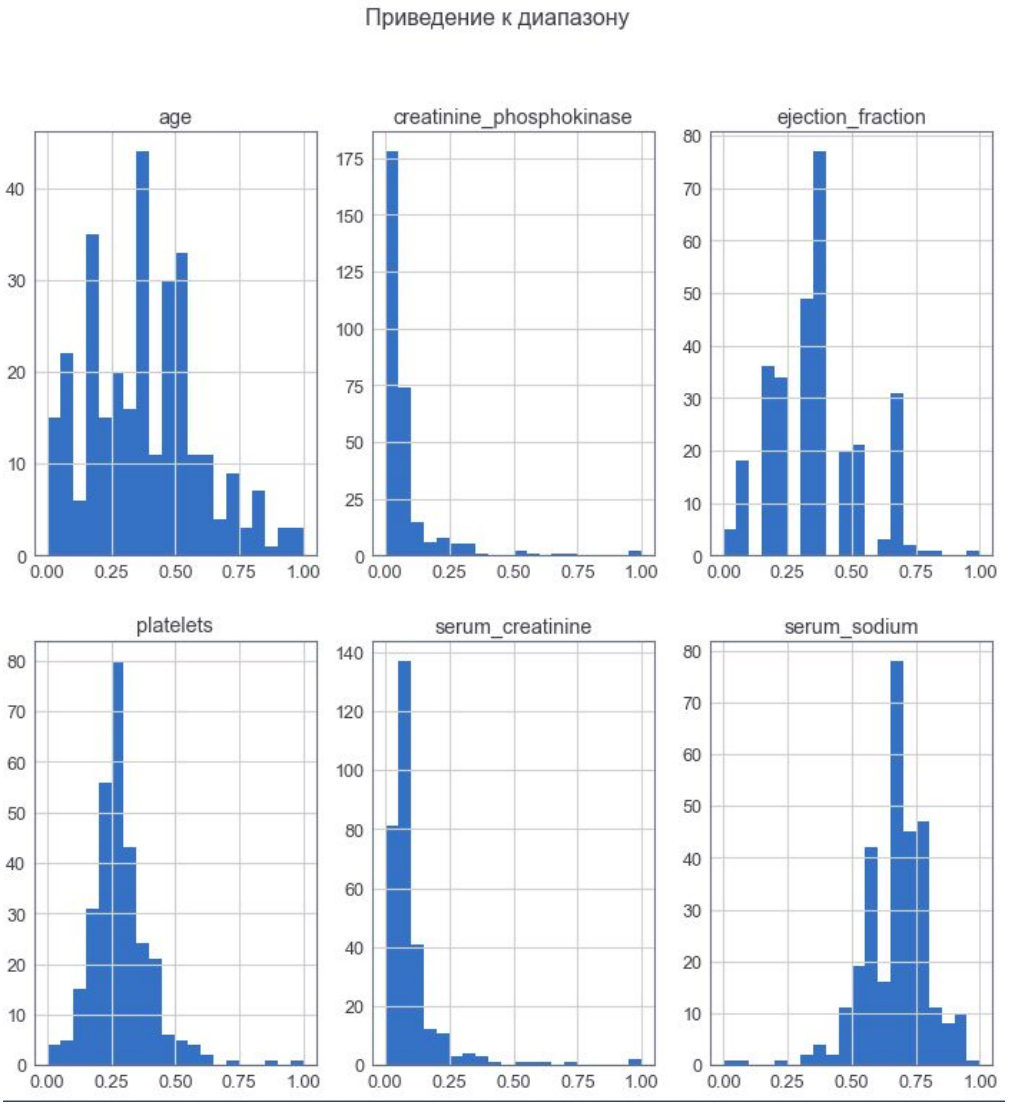
$$z_i = \frac{x_i - \mu}{\sigma}$$, где μ - мат. ожидание, а σ - СКО.

6. Поля mean_ и var_ объекта scaler содержат мат. ожидание и дисперсию соответственно.

7. Стандартизация по всем записям дает мат. ожидание незначительно отклоняющееся от 0 и СКО равное 1. Значения мат. ожидания и СКО после стандартизации по 150 признакам отличаются от 0 и 1 соответственно, так как стандартизация проведена не по всем признакам

Приведение к диапазону

1. Гистограмма для признаков после использования MinMaxScaler.



2. Значения приведены к диапазону [0, 1]. Это диапазон по умолчанию для данного скейлера.

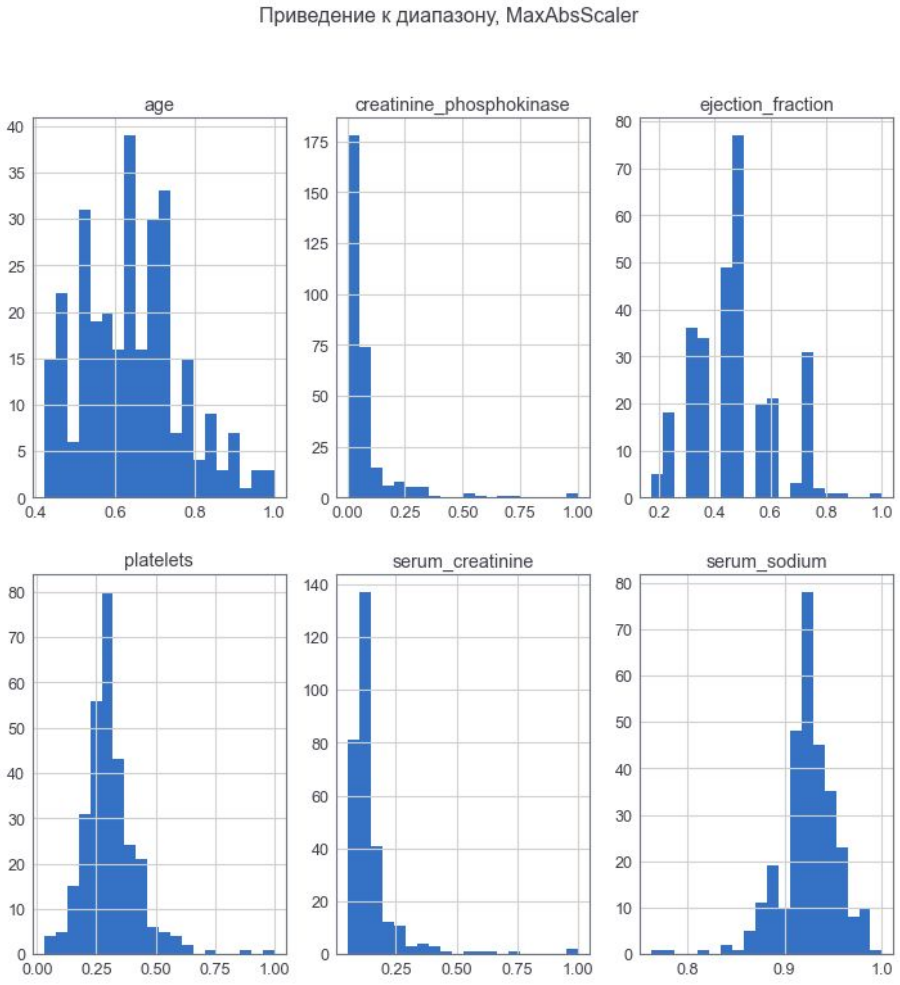
3. Минимальное и максимальное значение в данных для каждого признака:

	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

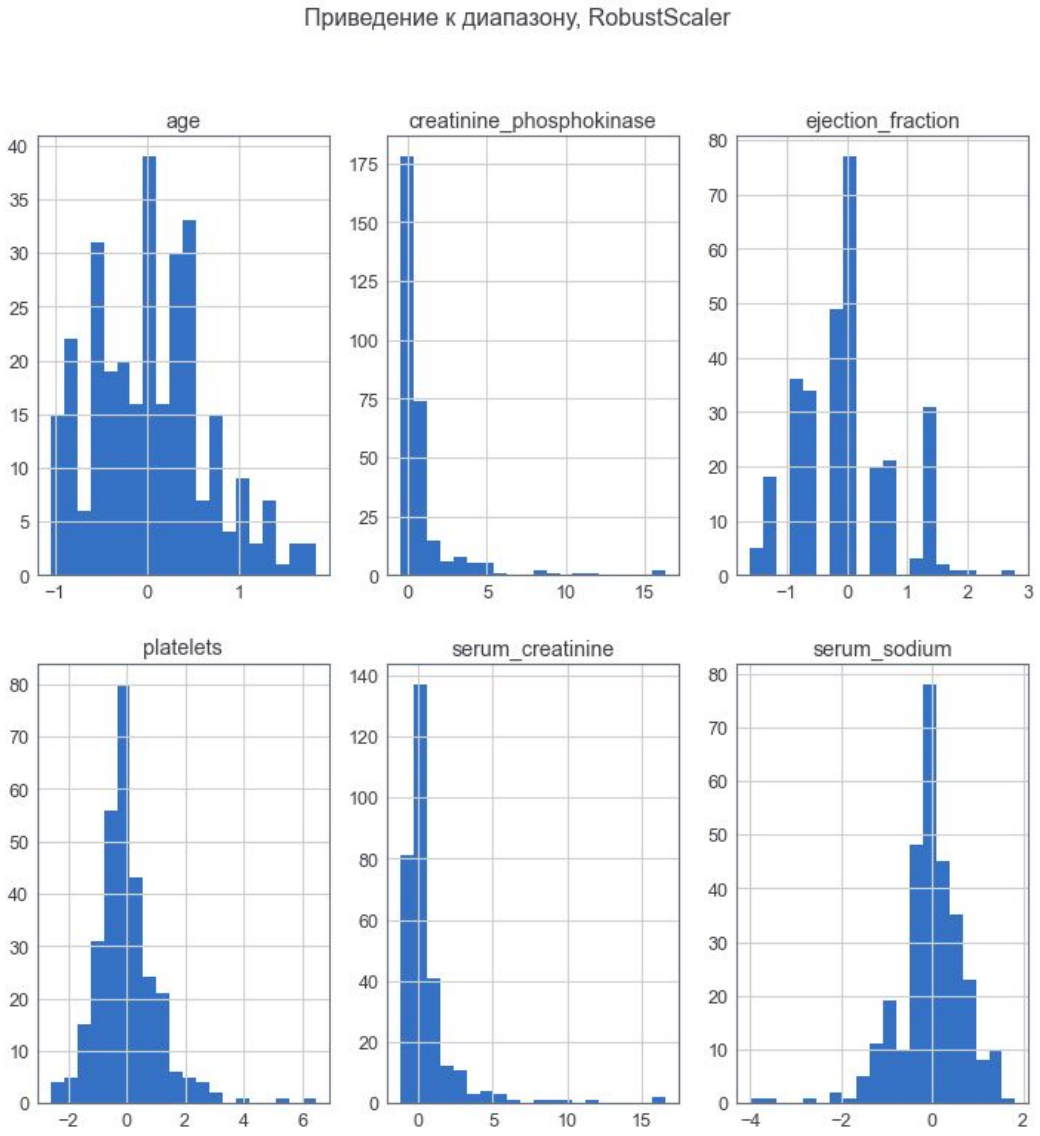
Таблица 4 - Минимальное и максимальное значение в данных для каждого признака

4. MaxAbsScaler



MaxAbsScaler приводит значения к диапазону [-1, 1]. Так как все значения положительны, они лежат в диапазоне [0, 1].

RobustScaler

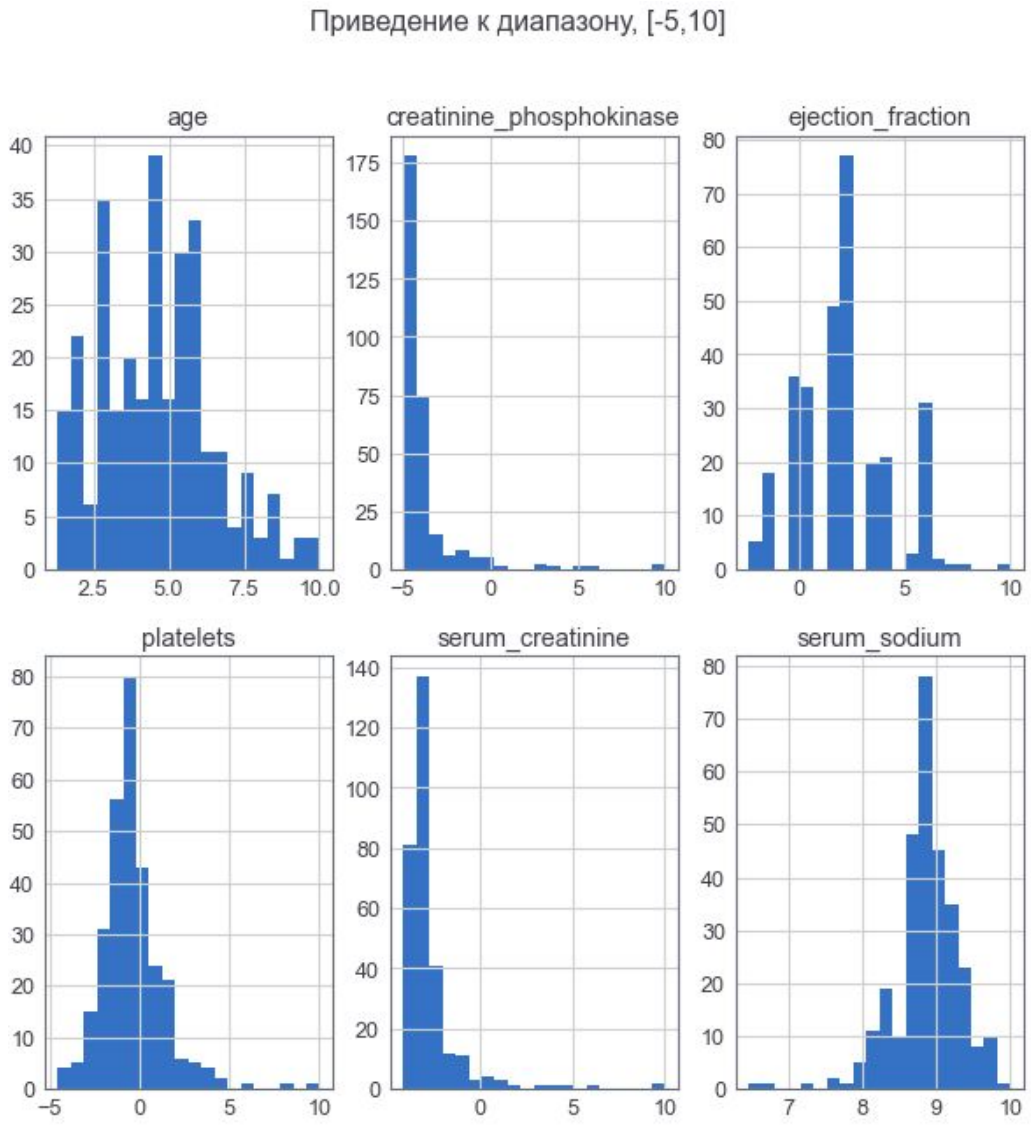


В отличие от предыдущих скейлеров RobustScaler не приводит данные к заранее заданному диапазону. RobustScaler преобразует вектор признаков путем вычитания медианы, а затем деления на диапазон между четвертями (значение 75% - значение 25%).

5. Данные приведены к диапазону [-5, 10]

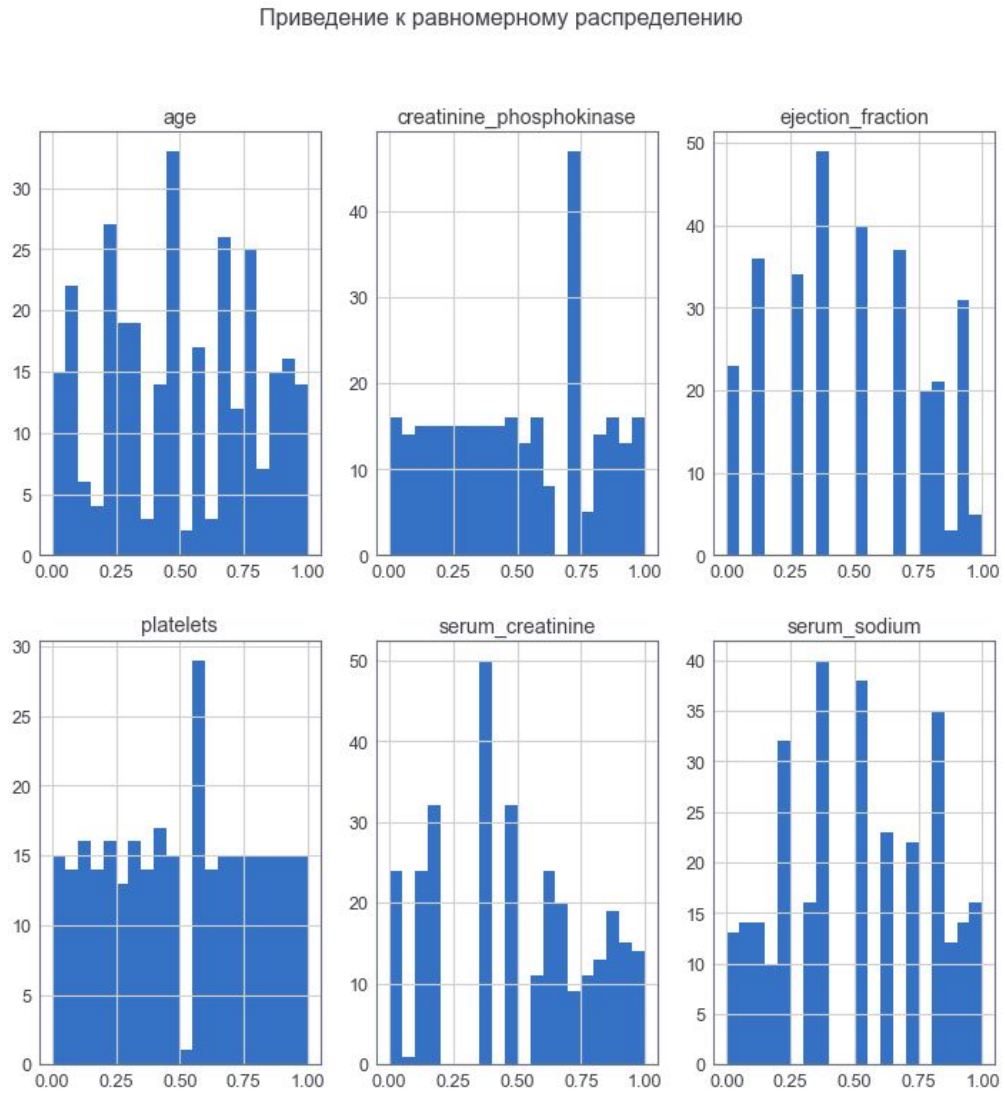
```
def to_range(data, mul, sub):
        return(data*mul - sub)

m5_10_scaler = preprocessing.MaxAbsScaler().fit(data)
data_m5_10_scaled = m5_10_scaler.transform(data)
data_m5_10_scaled = to_range(data_m5_10_scaled, 15, 5)
```



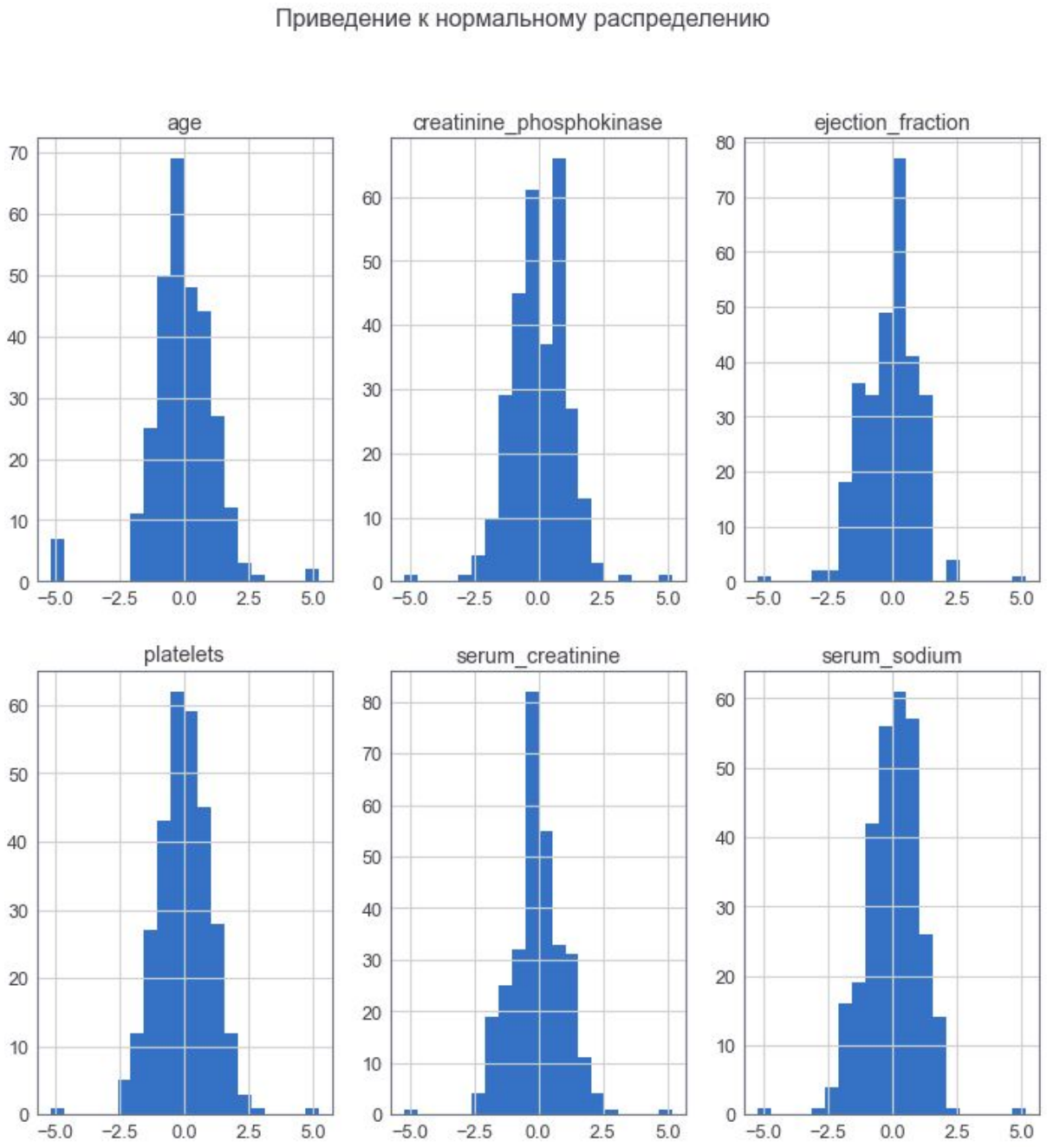
Нелинейные преобразования

2. Данные приведены к равномерному распределению:

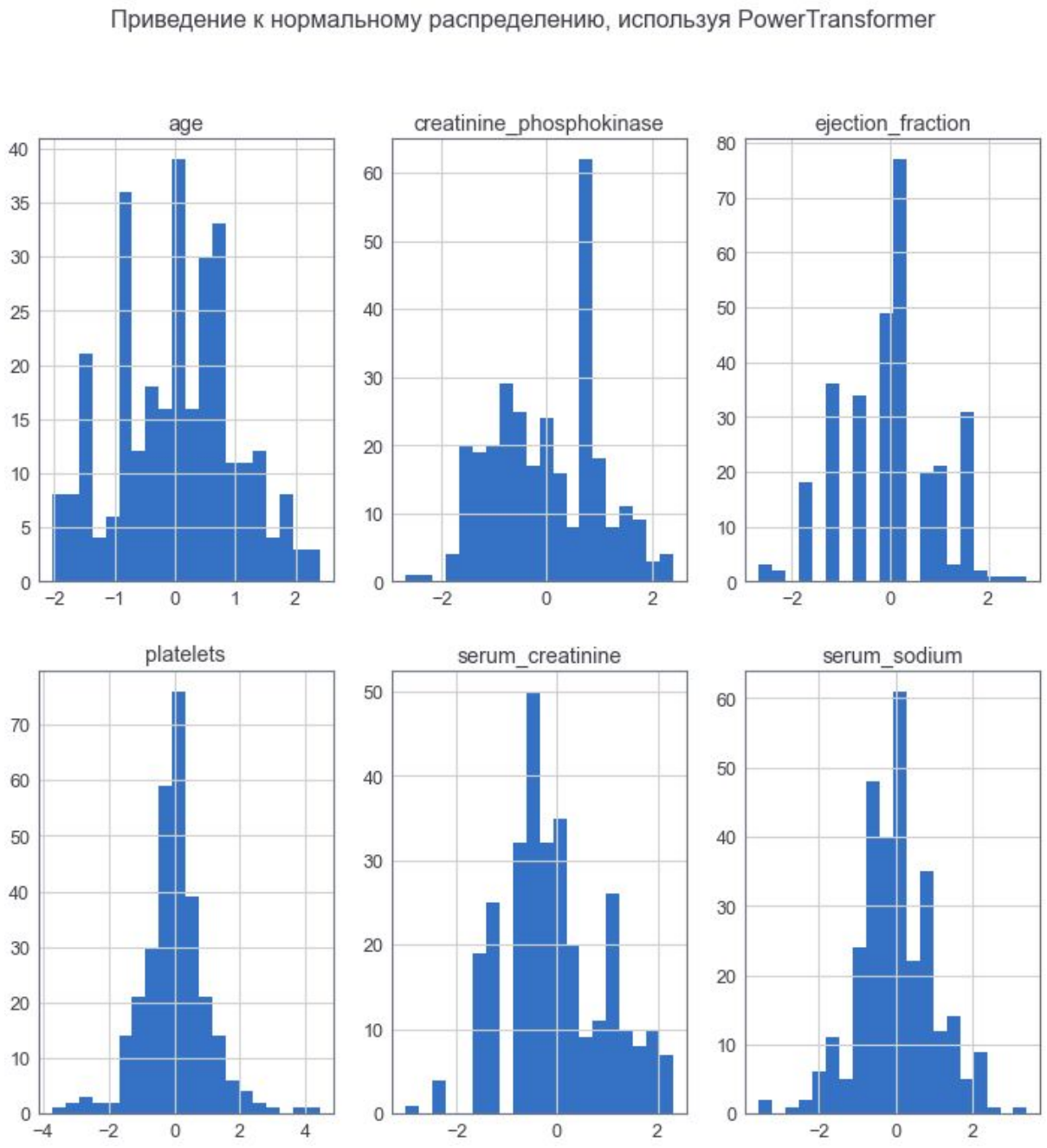


3. n_quantilies - количество квантилей в распределении, чем больше квантилей, тем точнее распределение. При этом это число должно быть меньше числа измерений.

4. Приведение к нормальному распределению, помощью параметра output_distribution=‘normal’.



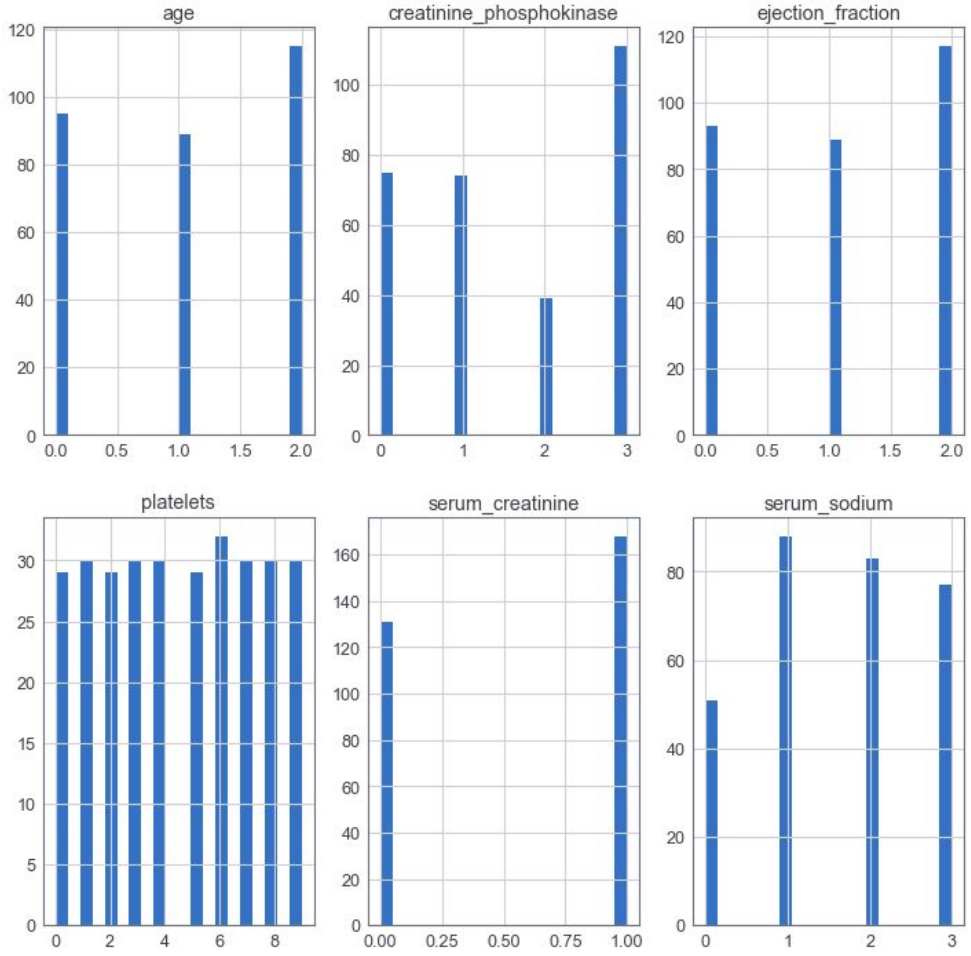
6. Приведение данные к нормальному распределению используя PowerTransformer:



Дискретизация признаков

2. Дискретизация признаков с использованием KBinsDiscretizer.

Приведение к нормальному распределению, используя PowerTransformer



3. Диапазоны каждого интервала для каждого признака:

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 данной работе были рассмотрены методы предобработки данных из библиотеки Scikit Learn. Были построены гистограммы признаков, для наглядной демонстрации работы методов. Были изучены сходства и различия в работе методов.

Рассмотрены алгоритмы масштабирования и стандартизации, которые могут помочь функциям получить более удобную форму для алгоритмов машинного обучения, методы нелинейного преобразование, которые помогут избавиться от выбросов данных. Дискретизация признаков может быть использована для разделения признаков на группы.