

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. 8304

Сергеев А.Д.

Преподаватель

Санкт-Петербург

2021

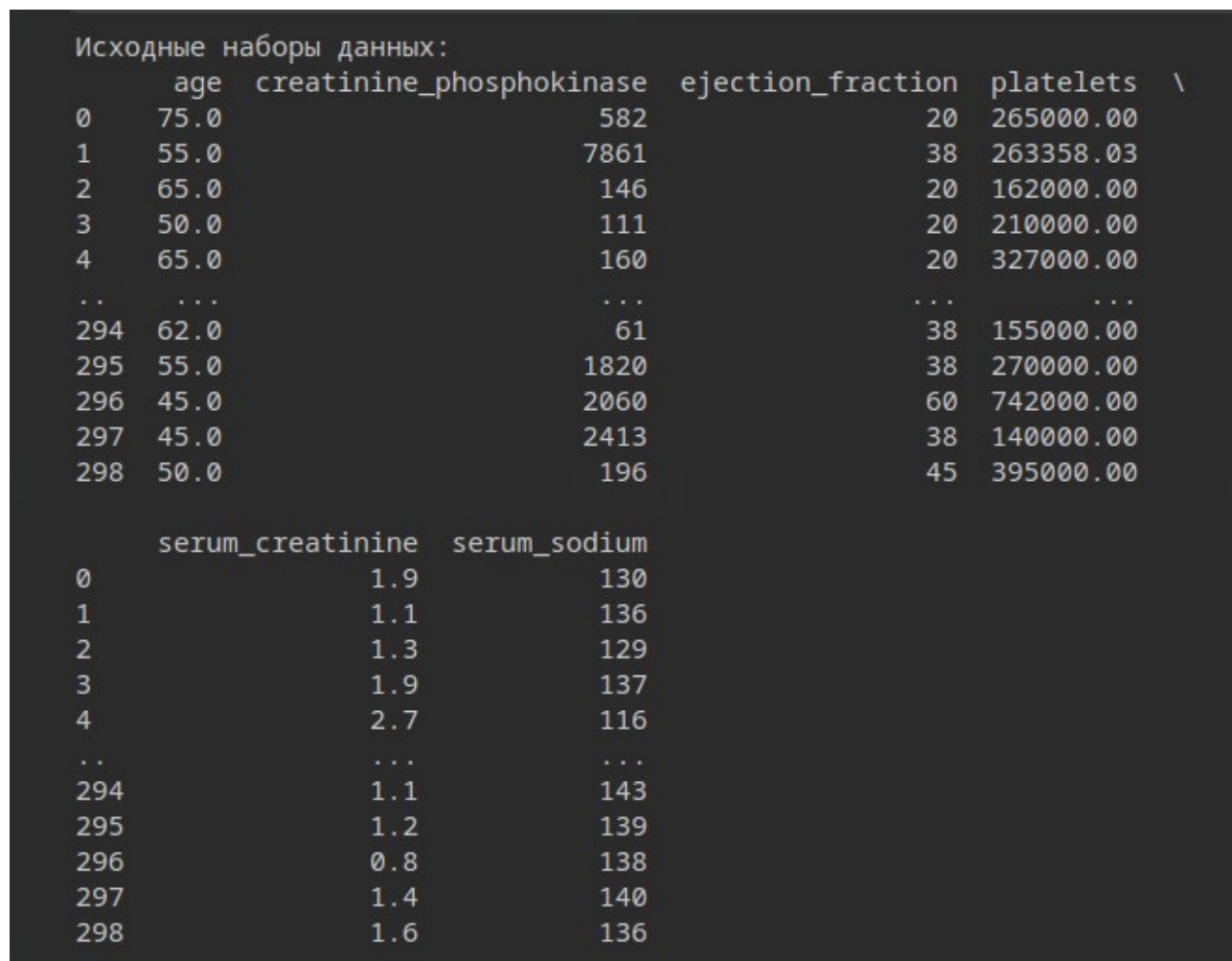
Цель работы.

Ознакомиться с методами предобработки данных из библиотеки Scikit Learn.

Ход работы.

Данные были загружены в датафрейм, ненужные атрибуты исключены.

Полученные данные:



```
Исходные наборы данных:
```

	age	creatinine_phosphokinase	ejection_fraction	platelets	\
0	75.0	582	20	265000.00	
1	55.0	7861	38	263358.03	
2	65.0	146	20	162000.00	
3	50.0	111	20	210000.00	
4	65.0	160	20	327000.00	
..
294	62.0	61	38	155000.00	
295	55.0	1820	38	270000.00	
296	45.0	2060	60	742000.00	
297	45.0	2413	38	140000.00	
298	50.0	196	45	395000.00	

	serum_creatinine	serum_sodium
0	1.9	130
1	1.1	136
2	1.3	129
3	1.9	137
4	2.7	116
..
294	1.1	143
295	1.2	139
296	0.8	138
297	1.4	140
298	1.6	136

Рисунок 1 - исходные данные в датафрейме

Были построены гистограммы признаков:

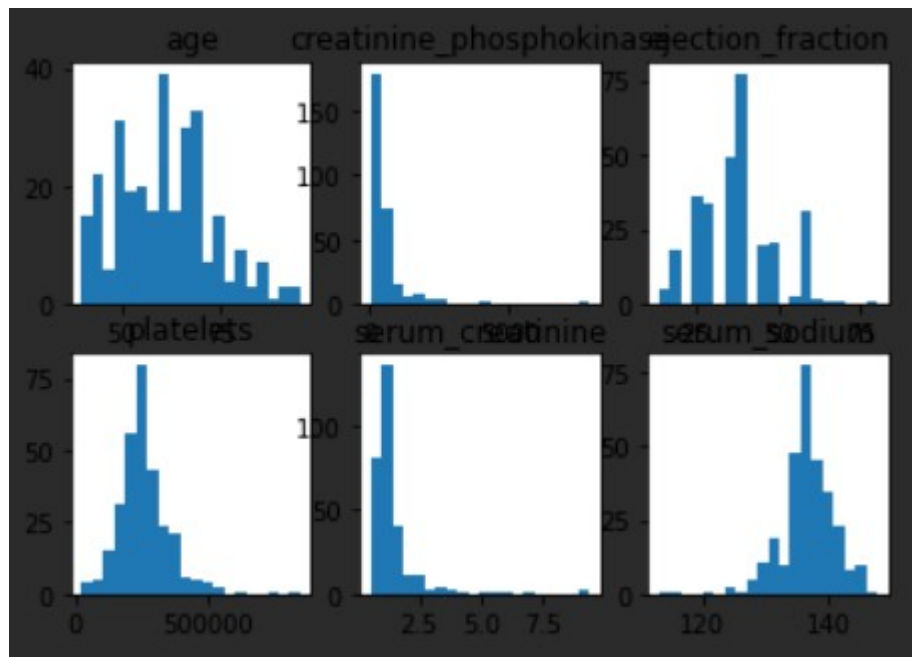


Рисунок 2 — гистограммы исходных данных

На основании гистограмм были определены диапазоны и медианы для каждого из признаков:

Название признака	Минимум	Максимум	Медиана
age	40	95	60
creatinine_phosphokinase	23	7861	250
ejection_fraction	14	80	38
platelets	25100	850000	262000
serum_creatinine	0.5	9.4	1.1
serum_sodium	113	148	137

Датафрейм преобразован к двумерному массиву.

Была настроена стандартизация на основе первых 150 наблюдений, после чего все данные были стандартизованы при помощи *StandartScaler*. Были построены гистограммы стандартизованных данных:

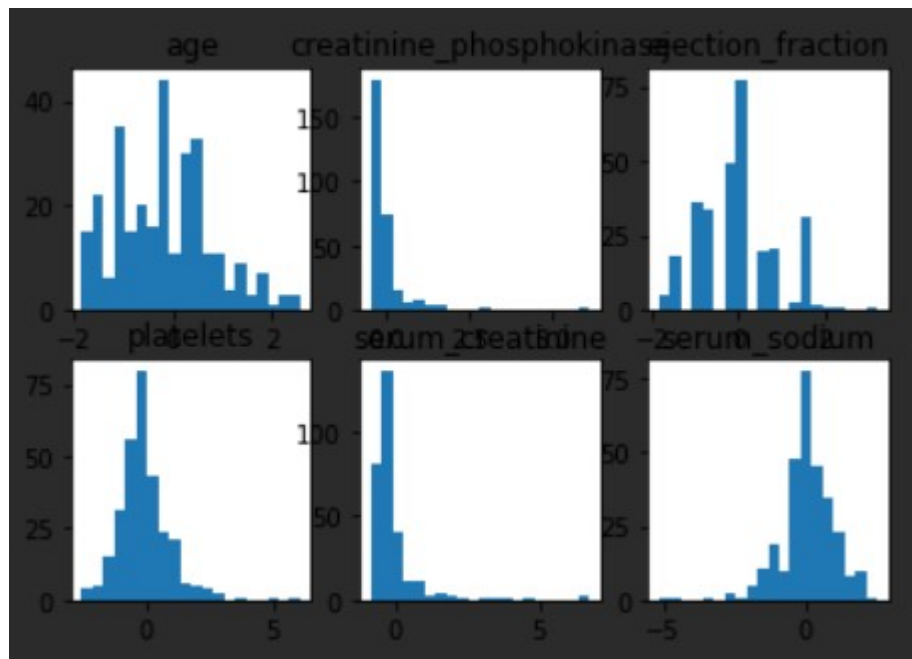


Рисунок 3 — гистограммы стандартизованных данных

По внешнему виду гистограммы стандартизованных данных похожи на гистограммы исходных данных. Масштаб оси ординат остался прежним, а масштаб оси абсцисс сильно изменился. В общем можно сказать, что значения на оси абсцисс на всех графиках принадлежат промежутку $[-10 .. 10]$.

Для каждого параметра были высчитаны значения математического ожидания и среднеквадратического отклонения до и после стандартизации:

Название признака	МО до	МО после	СКО до	СКО после
age	60.834	-0.170	11.875	0.954
creatinine_phosphokinase	581.839	-0.021	968.664	0.814
ejection_fraction	30.084	0.011	11.815	0.906
platelets	263358.029	-0.035	97640.548	1.015
serum_creatinine	1.394	-0.109	1.033	0.885
serum_sodium	136.625	0.038	4.405	0.970

Из полученных данных можно сделать вывод о том, что для стандартизации данных была использована формула: $z \approx (x - u) / s$, где x — исходное значение, u — математическое ожидание параметра, s — среднеквадратическое отклонение параметра.

Значения полей `mean_` и `var_` объекта `scaler` подтверждают предположение:

Название признака	scaler.mean_	scaler.var_
age	62.947	154.997
creatinine_phosphokinase	607.153	1415488.823
ejection_fraction	37.947	170.024
platelets	266746.749	9252860499.079
serum_creatinine	1.521	1.361
serum_sodium	136.453	26.608

Была проведена настройка стандартизации на всех данных. Результаты стали точнее, так как значения математического ожидания приблизились к 0, а среднеквадратического отклонения — к 1.

При помощи MinMaxScaler данные были приведены к диапазону [0, 1]:

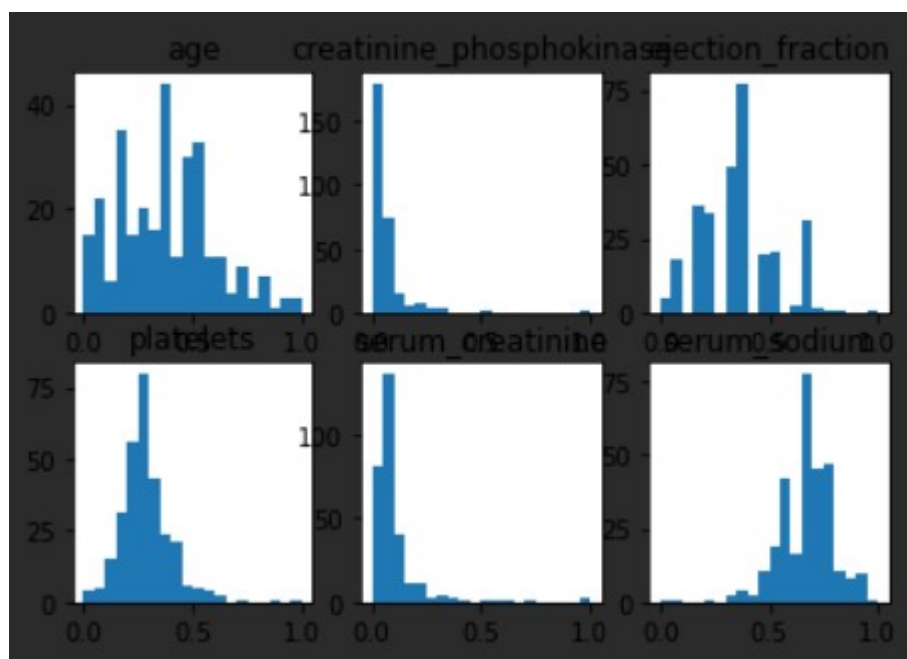


Рисунок 4 — гистограмма приведенных к диапазону данных

По внешнему виду гистограммы приведенных к диапазону данных похожи на гистограммы исходных данных. Масштаб оси ординат остался прежним, а масштаб оси абсцисс изменился, теперь все данные принадлежат отрезку [0, 1].

Таблица, в которой указаны минимальные и максимальные значения для каждого признака, была приведена выше. Значения не изменились.

Данные были трансформированы с помощью *MaxAbsScaler* и *RobustScaler*.
 Были построены гистограммы:

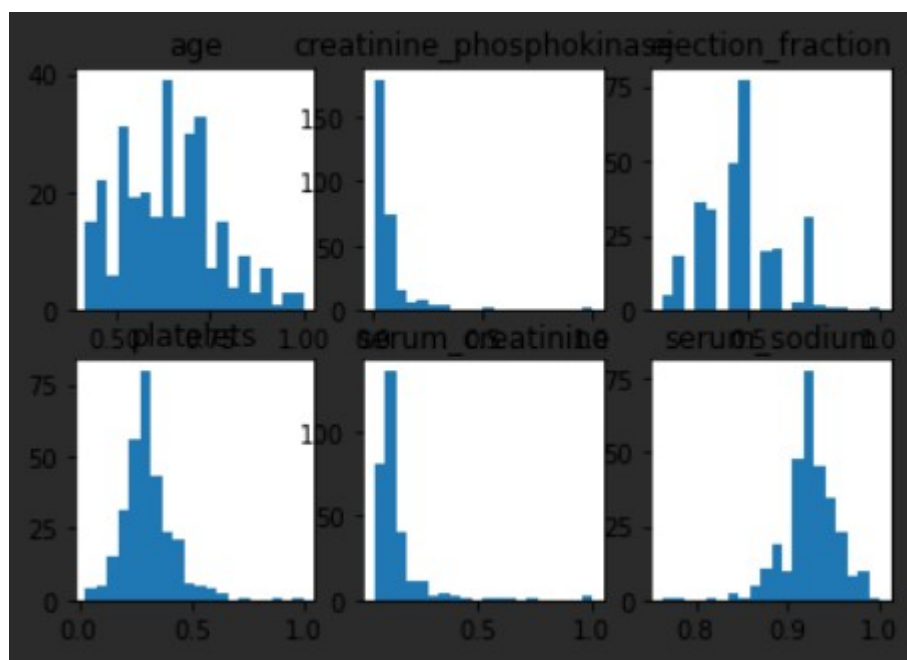


Рисунок 5 — гистограмма данных, обработанных *MaxAbsScaler*

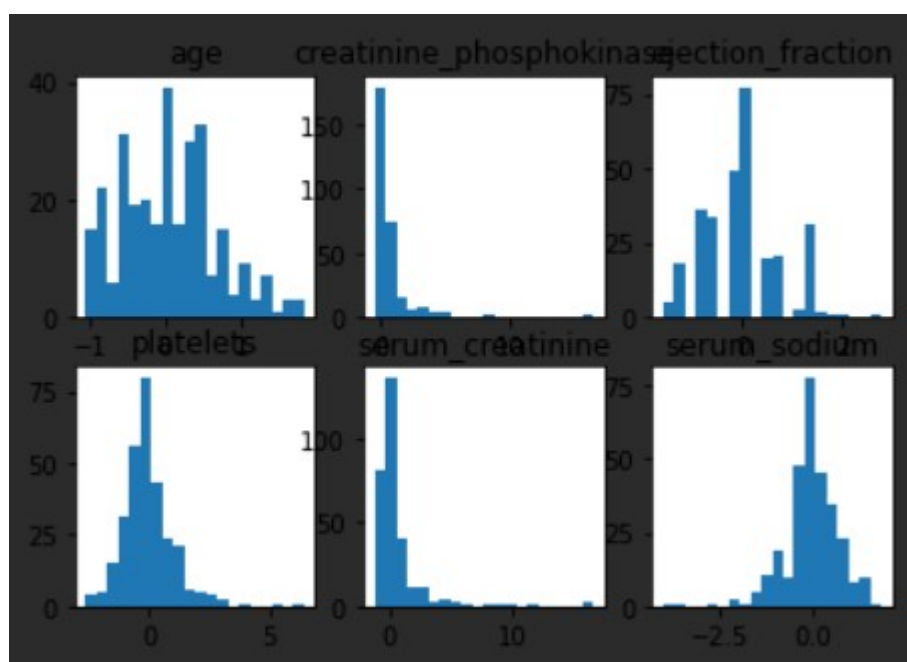


Рисунок 6 - гистограмма данных, обработанных *RobustScaler*

Основываясь на документации *MaxAbsScaler* и *RobustScaler*, можно сказать, что первый также приводит данные к диапазону $[0, 1]$, а второй — к интерквартильному диапазону (между 25 и 75 квантилью).

Для того, чтобы привести данные к диапазону $[-5, 10]$ достаточно передать в конструктор *MinMaxScaler* параметр *feature_range*, равный кортежу $(-5, 10)$.

Данные были обработаны *QuantileTransformer* для получения равномерного распределения:

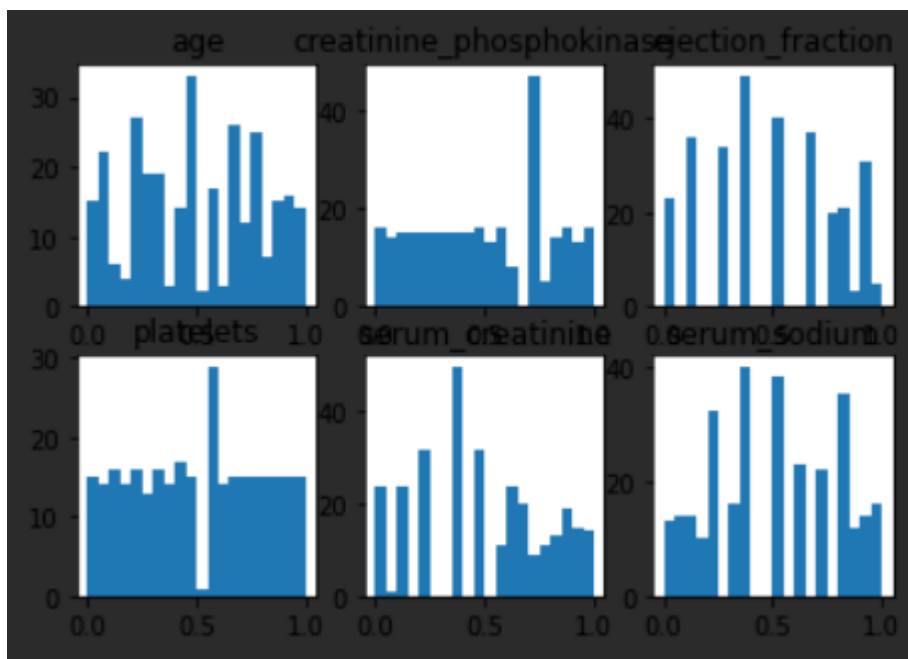


Рисунок 7 — гистограмма равномерно распределенных данных

По внешнему виду гистограммы равномерно распределенных данных не похожи на гистограммы исходных данных.

Параметр *n_quantiles* представляет из себя количество фрагментов, на которое будет разбита квантильная функция при вычислении. Согласно документации, максимально точное значение дает количество квантилей, равное количеству измерений.

При передаче в конструктор *QuantileTransformer* параметра *output_distribution*, равного *"normal"*, данные будут распределены нормально:

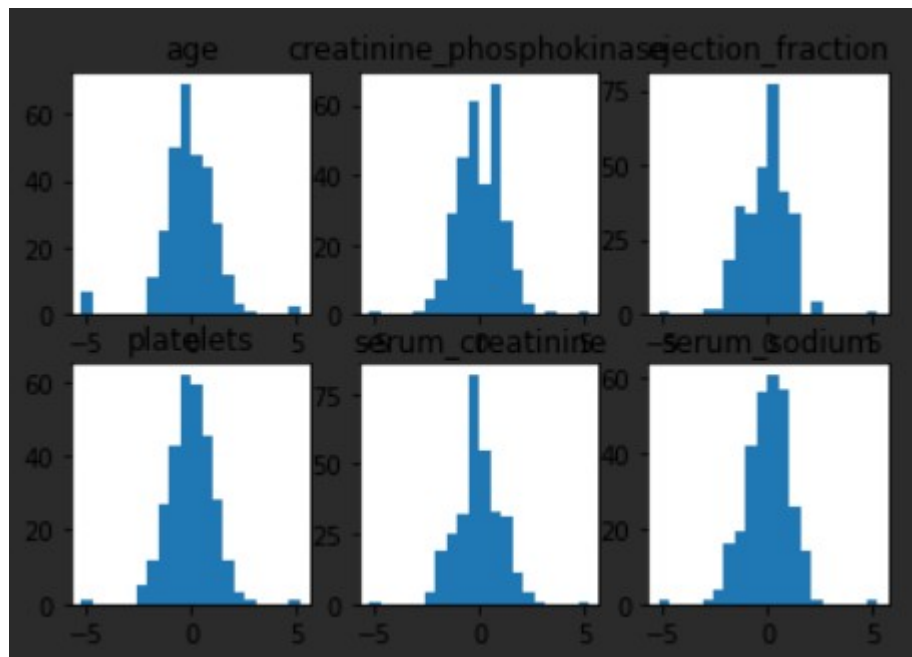


Рисунок 8 — нормально распределенные данные при помощи *QuantileTransformer*

Также данные могут быть обработаны и распределены нормально при использовании *PowerTransformer*.

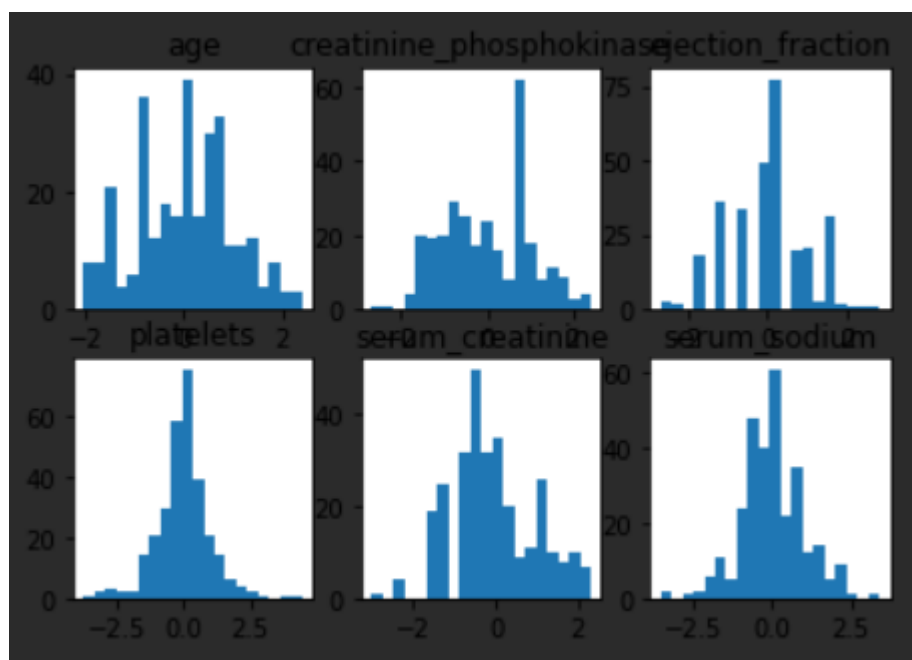


Рисунок 9 — нормально распределенные данные при помощи *PowerTransformer*

Была проведена дискретизация данных с использованием *KBinsDiscretizer*.

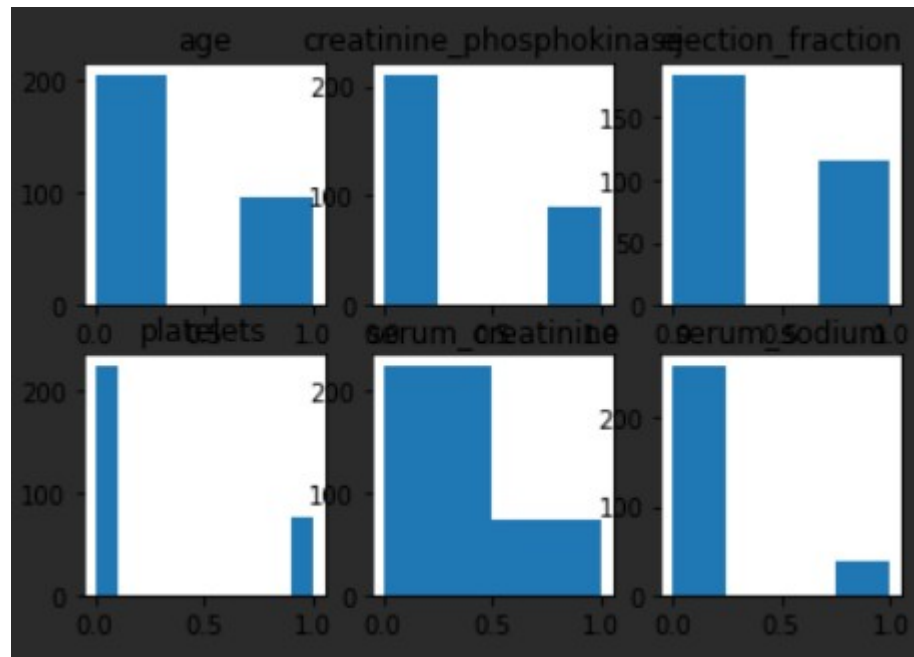


Рисунок 10 — дискретизированные данные при помощи *KBinsDiscretizer*

Количество столбцов в гистограмме определенного признака равно значению параметра *n_bins* для него. Внешний вид гистограммы похож на внешний вид гистограммы исходных данных с уменьшенным количеством столбцов.

Через параметр *bin_edges_* были получены границы диапазонов для каждого признака:

Название признака	scaler.bin_edges_
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.

В ходе лабораторной работы было успешно произведено ознакомление с методами предобработки данных из библиотеки Scikit Learn.