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. 6304 _____ Ковынев М.В.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Цель

Ознакомиться с методами предобработки данных из библиотеки Scikit Learn

Ход работы

1. Загружен датасет по ссылке: https://www.kaggle.com/andrewmvd/heart-failure-clinical-dat. Данные представлены в виде csv таблицы.
2. Создан Python скрипт. Загружен датасет в датафрейм, и исключены бинарные признаки и признак времени.

```
"C:\Program Files (x86)\Microsoft Visual Studio\Shared\Python37_64\python.exe"
      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

3. Построены гистограммы признаков

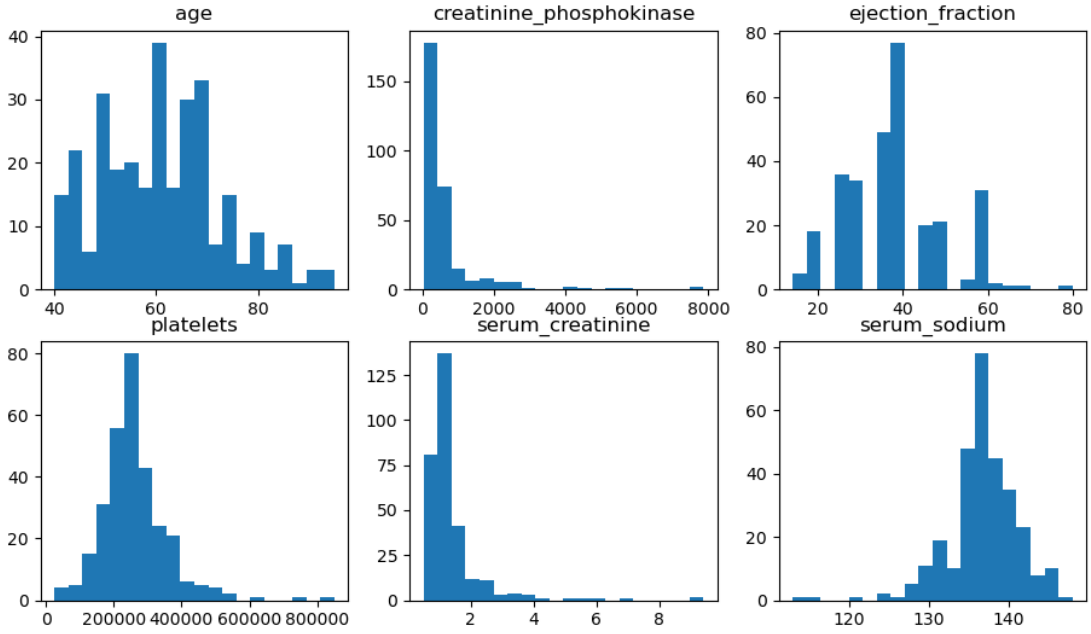

Рисунок 2 — Гистограммы признаков

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

   Таблица 1

Название признака	**Наибольшее кол-во наблюдений**	**Минимум**	**Максимум**
age	60	40	95
creatinine_phosphokinase	582	23	7861
ejection_fraction	35	14	80
platelets	263358.03	25100	850000
serum_creatinine	1	0.5	9.4
serum_sodium	136	113	148

5. Подключен модуль Sklearn. Настроена стандартизацию на основе первых 150 наблюдений используя StandardScaler. Стандартизованы все данные (data_scaled_150). Построены гистограммы стандартизированных данных.

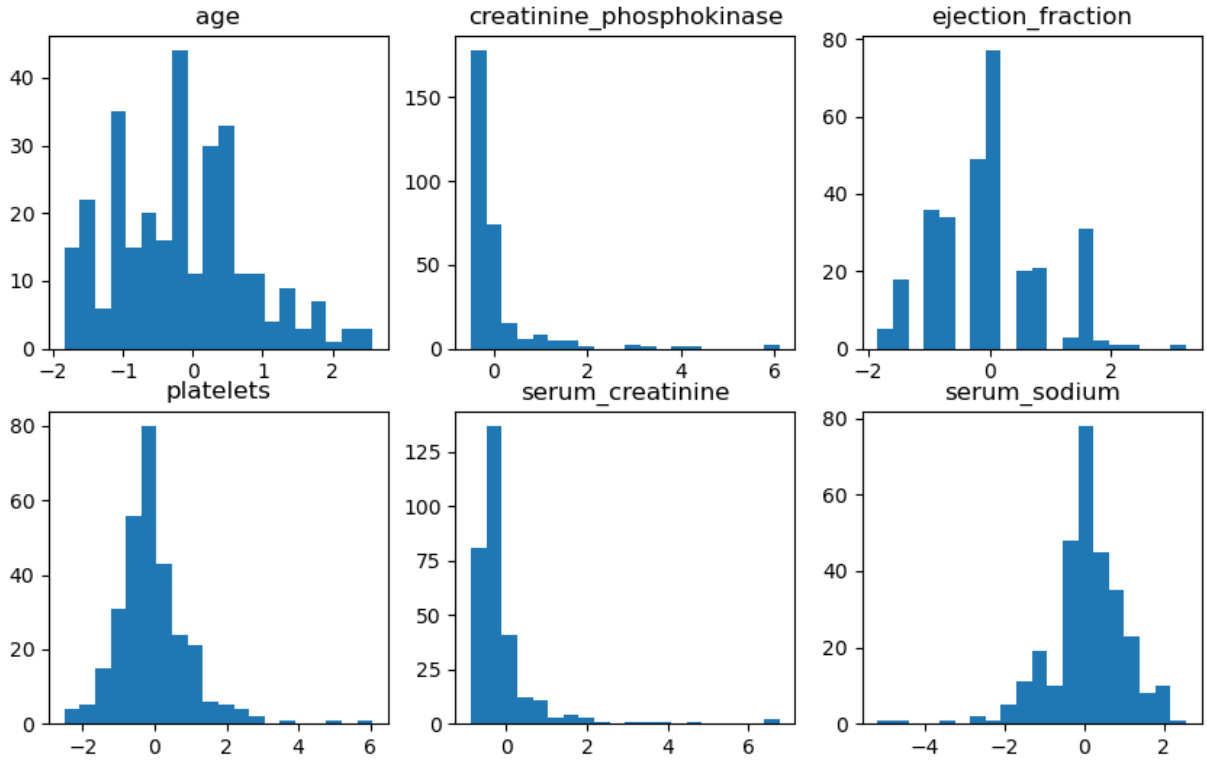

Рисунок 3 — Гистограммы признаков StandardScaler 150

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

   Таблица 1

Название признака	**Наибольшее кол-во наблюдений**	**Минимум**	**Максимум**
age	-0.23	-1.84	2.57
creatinine_phosphokinase	-0.02	-0.49	6.09
ejection_fraction	-0.22	-1.83	3.22
platelets	-0.03	-2.51	6.06
serum_creatinine	-0.44	-0.87	6.75
serum_sodium	-0.09	-5.16	2.54

7. В виду применения преобразования стандартизации диапазон значений и наибольшее кол-во наблюдений изменились.

8. Проведена настройка стандартизации на всех данных (data_scalled_full).
9. Рассчитаны мат. ожидание и СКО до и после стандартизации. На основании этих значений выведены для каждого признака формулы, по которым они стандартизировались.

   Таблица 2

Выборка	**data**		**data_scaled_150**		**data_scalled_full**	
Метрика	mean	std	mean	std	mean	std
age	60.83	11.87	-0.16	0.95	5.7e-16	0.99
creatinine_phosphokinase	581.83	968.66	-0.02	0.81	0.0	1
ejection_fraction	38.08	11.81	0.01	0.90	-3e-17	1
platelets	263358.02	97640.54	-0.03	1.01	7.7e-17	1
serum_creatinine	1.39	1.03	-0.10	0.88	1.4e-1	1
serum_sodium	136.62	4.40	0.03	0.97	-8e-16	0.99

10. На основании таблицы выяснено, что стандартизация имеет следующий вид:

$$Y = \frac{X - \mu(X)}{std(X)}$$

$\mu(X)$ – mean (мат ожидание), $std(X)$ – std (СКО).

11. Сравнены значения из формул с полями mean_ и var_ объекта scaler. mean_ - мат. ожидание, var_ - дисперсия величин, на основании которых производится стандартизация данных.

Выборка	**scaller_150**		**scaller_full**	
Метрика	mean_	var_	mean_	var_
age	62.94	154.99	60.83	141.01
creatinine_phosphokinase	607.15	1415488.82	581.83	938309.88
ejection_fraction	37.94	170.02	38.08	139.59
platelets	266746.74	9252860499.07	263358.02	9533676546.27
serum_creatinine	1.52	1.36	1.39	1.06
serum_sodium	136.45	20.60	136.62	19.40

12. Приведены данные к диапазону используя MinMaxScaler

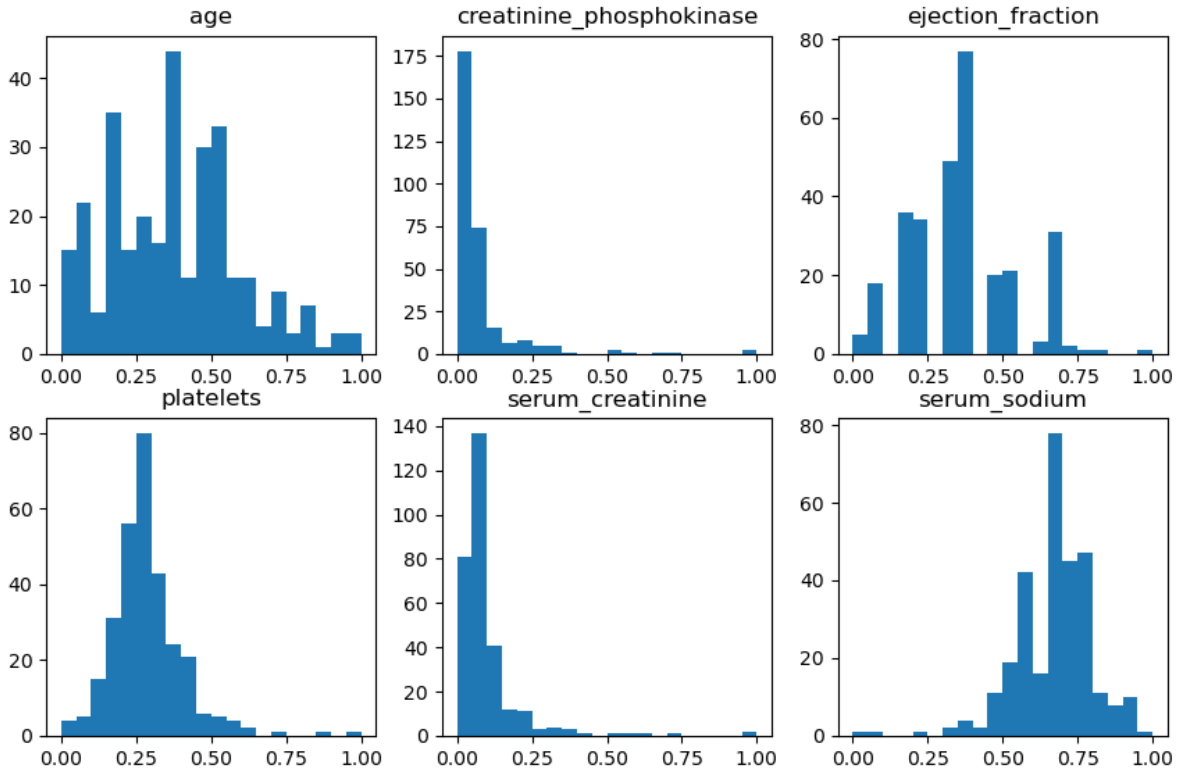

Рисунок 4 — Гистограммы признаков MinMaxScaler

Данные приводятся к диапазону [0, 1]. Преобразование можно получить следующим образом.

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

13. Через параметры MinMaxScaler определены минимальное и максимальное значение в данных для каждого признака. Данные совпадают с таблицей 1.

14. Аналогично трансформированы данные используя MaxAbsScaler и RobustScaler. Построены гистограммы. Определены к какому диапазону приводятся данные

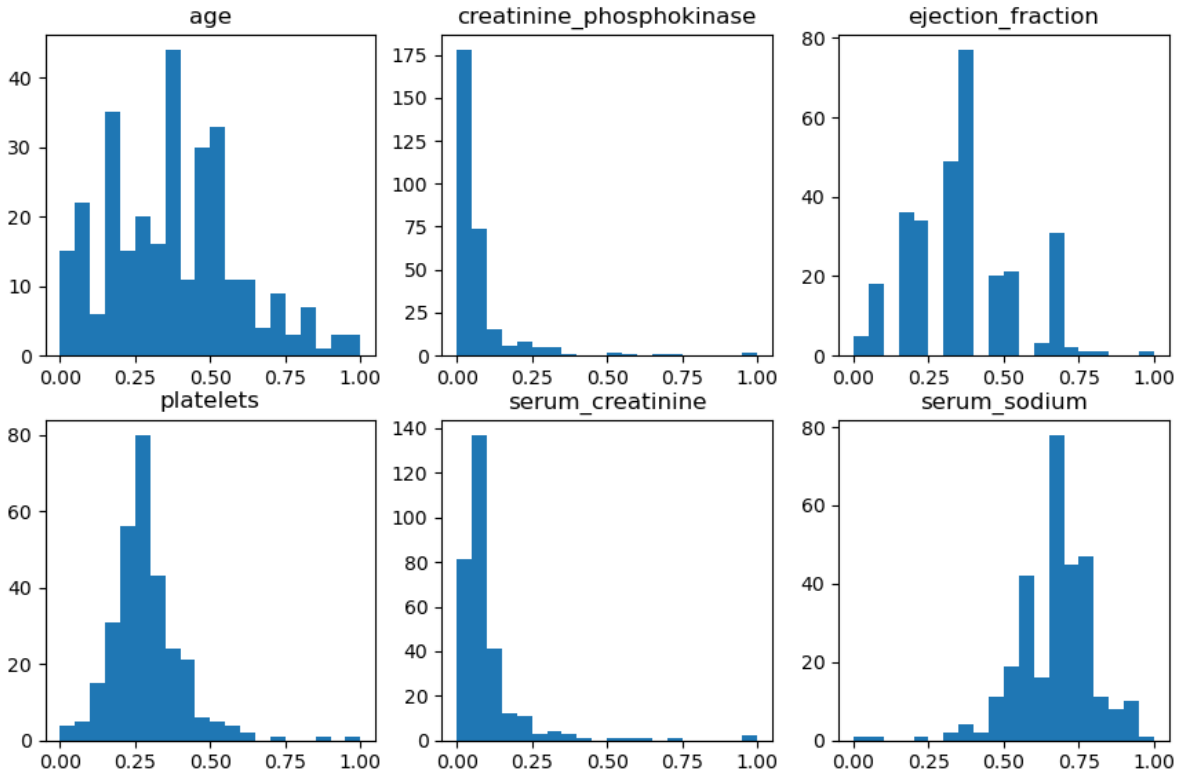

Рисунок 5 — Гистограммы признаков MaxAbsScaler

Данные приводятся к интервалу [0, 1] как и MinMaxScaler.

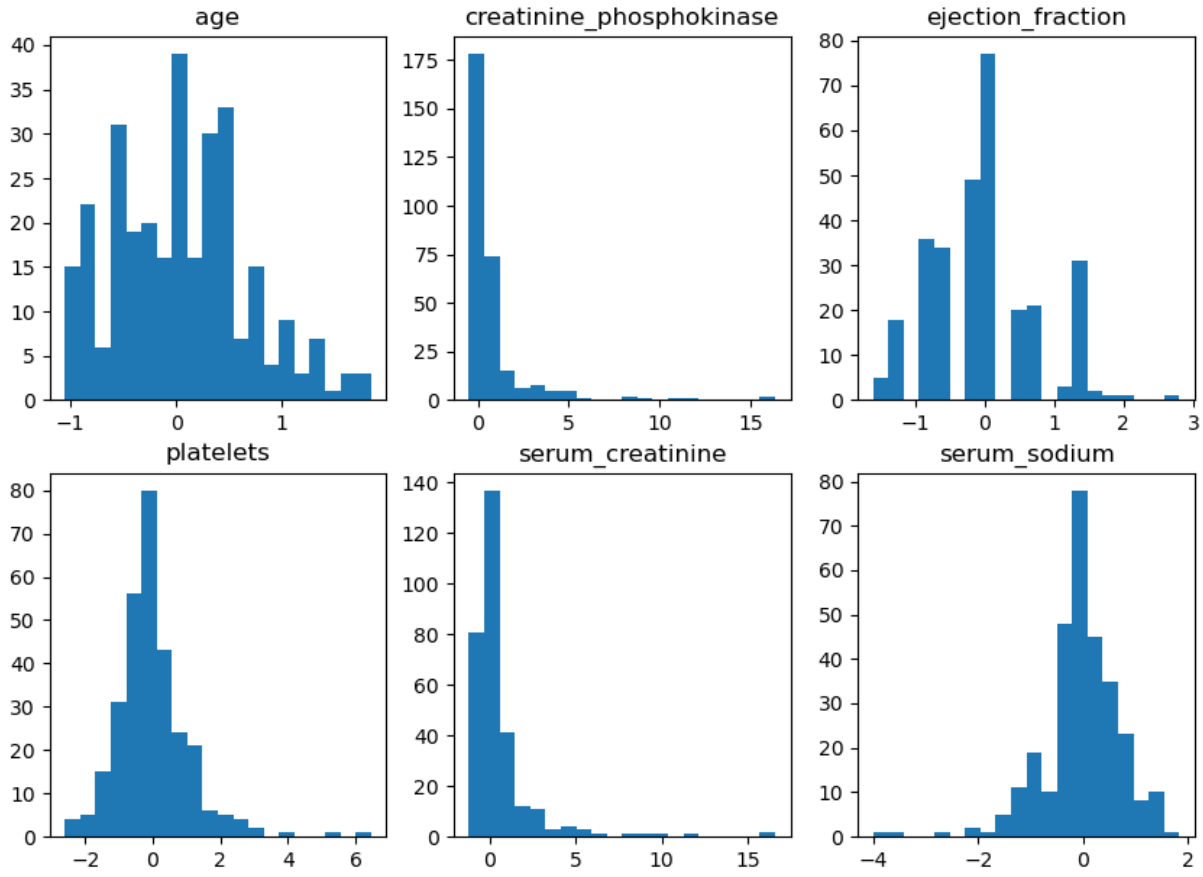

Рисунок 6 — Гистограммы признаков RobustScaler

RobustScaler отнимает медиану и масштабирует данные в соответствии с межквартильным размахом (диапазон между 1-м квартилем (25-й квантиль) и 3-м квартилем (75-й квантиль)).

15. Написана функция, которая приводит все данные к диапазону [-5 10]

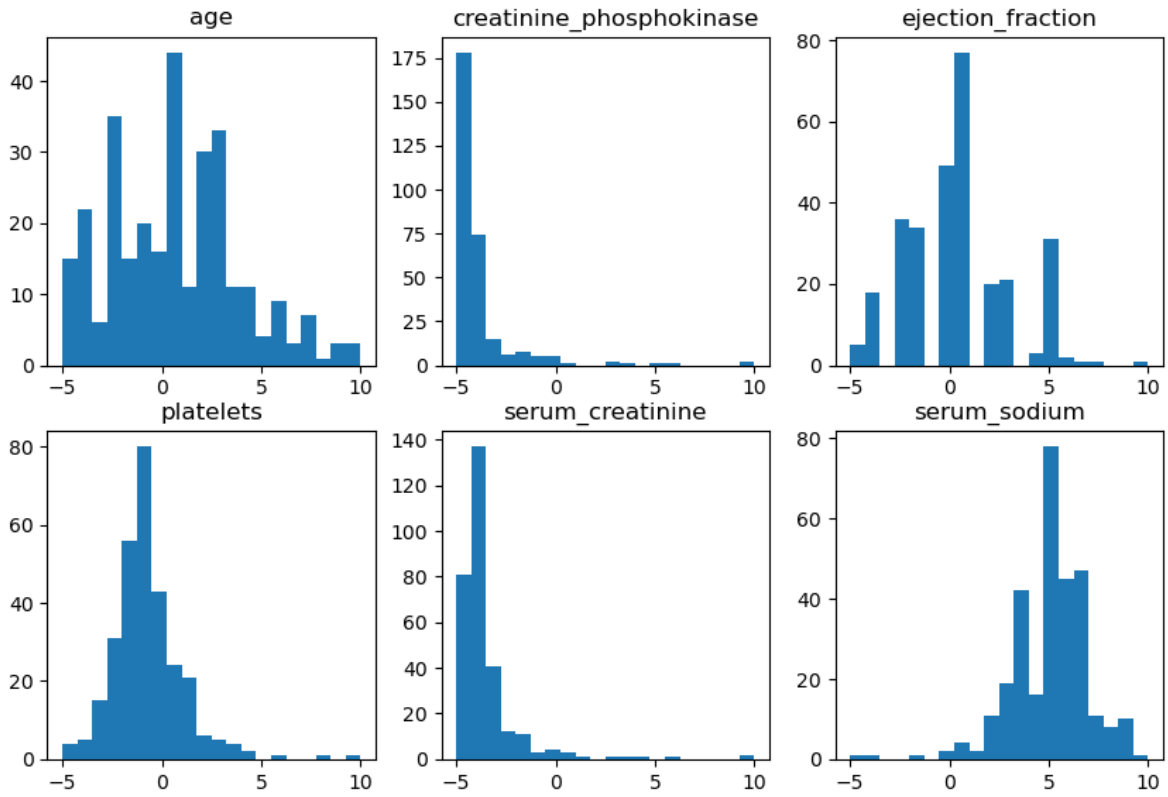

Рисунок 6 — Гистограммы признаков [-5 10]

16. Приведите данные к равномерному и нормальному распределению используя QuantileTransformer

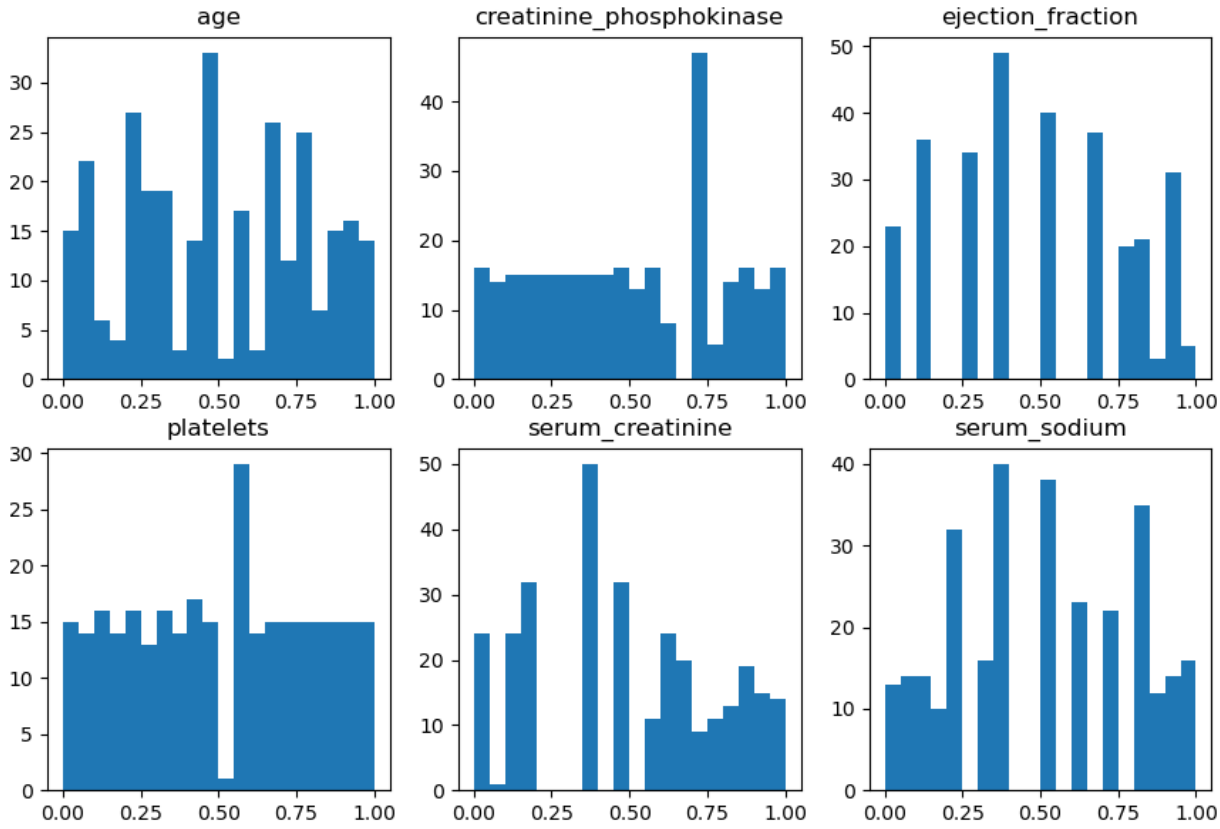

Рисунок 7 — Гистограммы признаков QuantileTransformer (равномерное)

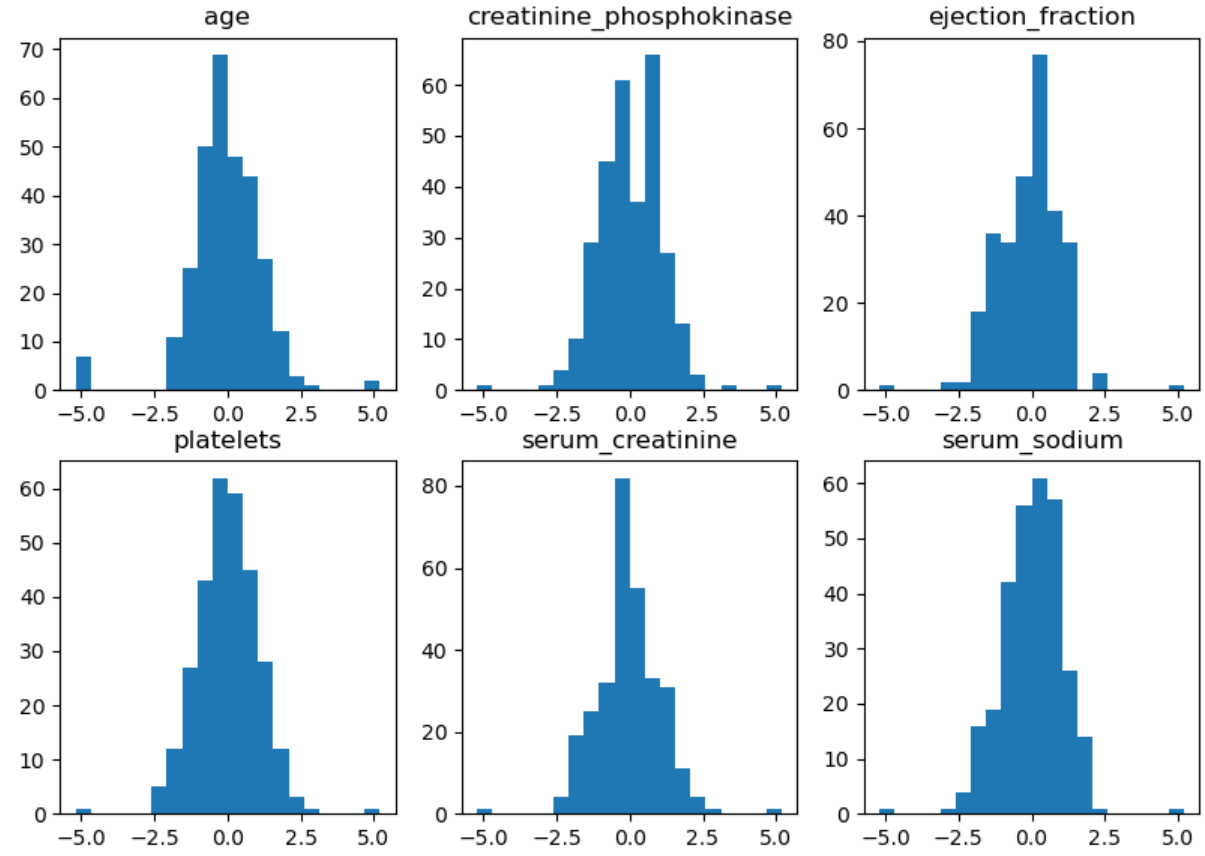

Рисунок 8 — Гистограммы признаков QuantileTransformer (нормальное)

17. n_quantiles – параметров, задающий количество квантилей, которые используются для дискретизации функции распределению. Чем больше квантилей, тем ближе функция к заданному распределению.
18. Самостоятельно приведите данные к нормальному распределению используя PowerTransformer

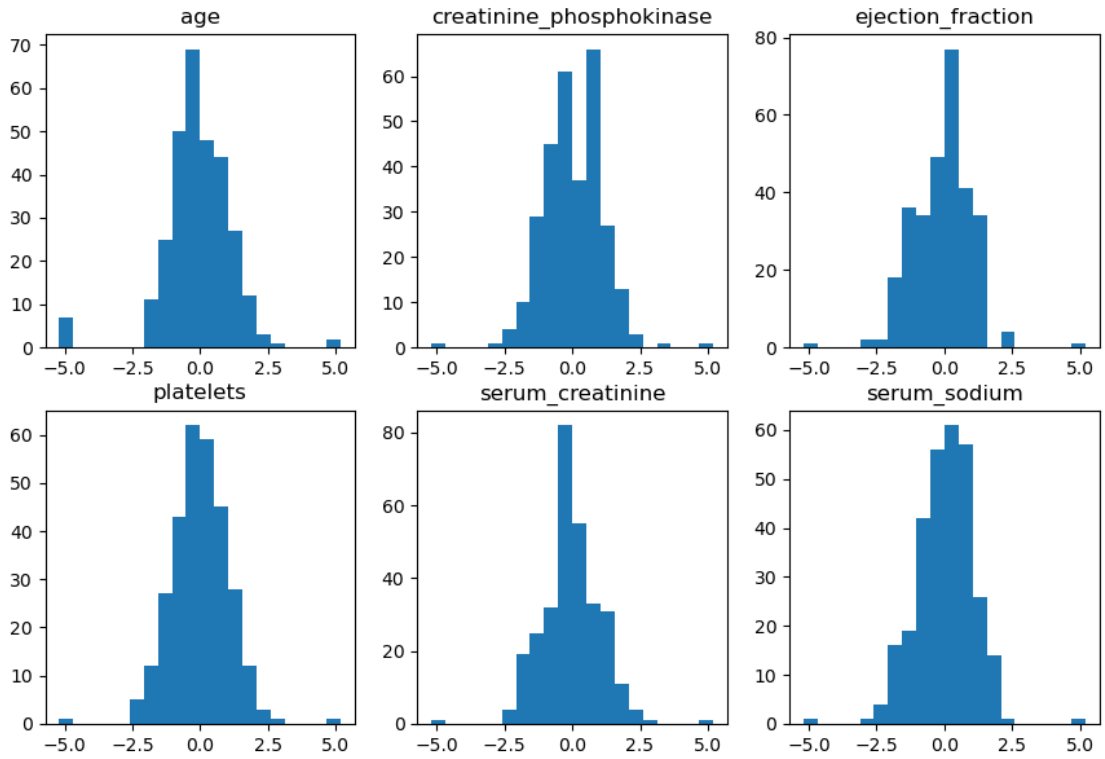

Рисунок 9 — Гистограммы признаков PowerTransformer

19. Проведена дискретизация признаков, используя KBinsDiscretizer, на следующее количество диапазонов:
    - age - 3
    - creatinine_phosphokinase - 4
    - ejection_fraction – 3
    - platelets - 10
    - serum_creatinine - 2
    - serum_sodium – 4

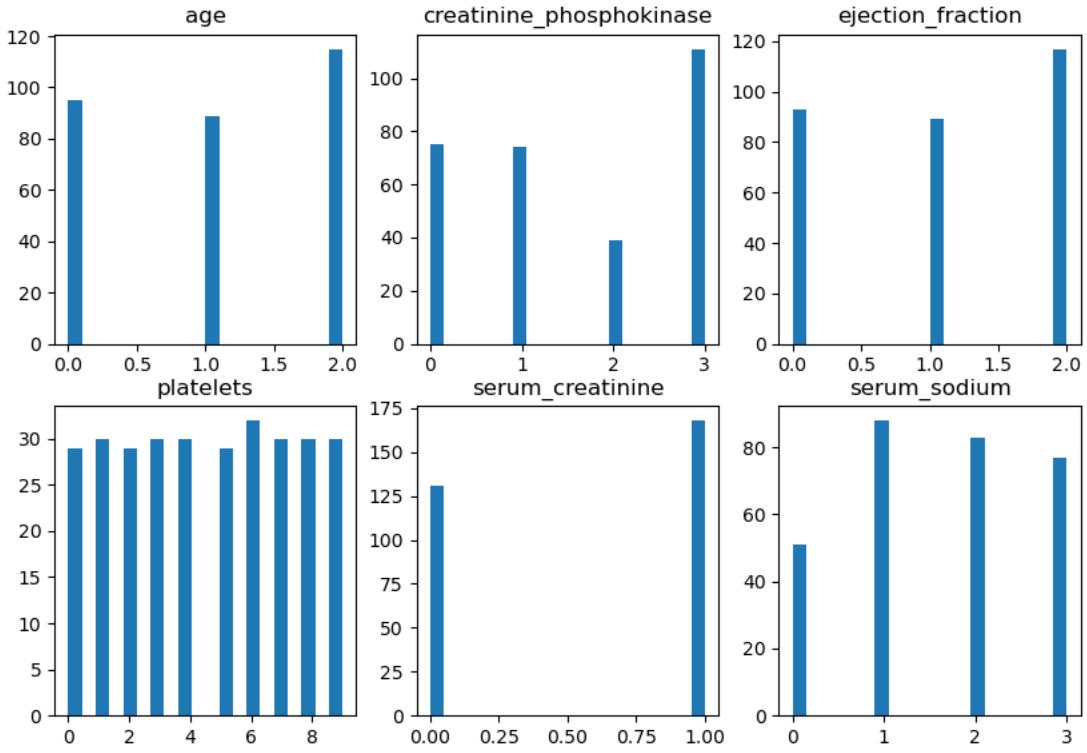

Рисунок 9 — Гистограммы признаков KBinsDiscretizer

20.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.]

Вывод

В ходе выполнения данной лабораторной работы было выполнено ознакомление с методами предобработки данных из библиотеки Scikit Learn. Выяснено, что приведение к диапазону не меняет форму распределению, на частичных данных происходит снижение качества результирующего набора.