МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №1

по дисциплине «Предобработка данных»

Тема: Машинное обучение

Студент гр. 6304 _____ Виноградов К.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Загрузка данных.

Загрузим датасет в датафрейм, и исключить бинарные признаки и признак времени. Результат представлен на рис.1.

```
      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]

Process finished with exit code 0
```

Рисунок 1 – Начальные признаки

Построим диаграммы признаков. Результат представлен на рис.2.

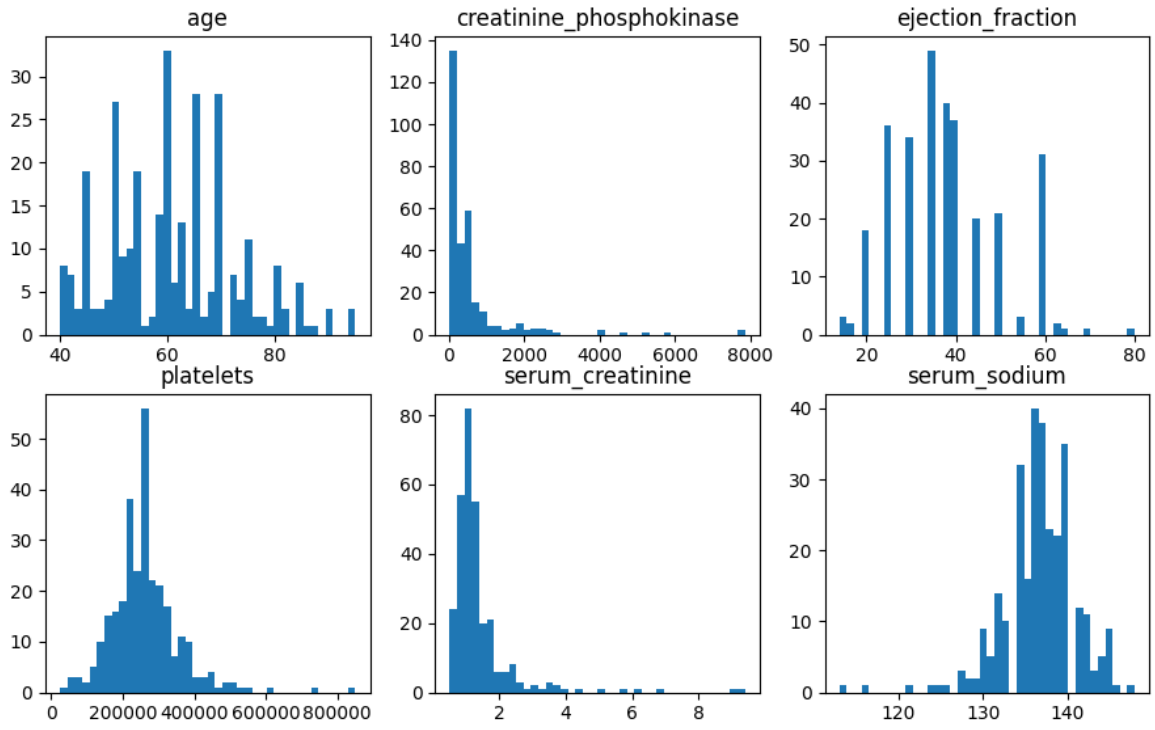

Рисунок 2 – Диаграммы признаков

На основании гистограмм определите диапазоны значений для каждого из признаков, а также возле какого значения лежит наибольшее количество наблюдений. Результат на рис. 3.

```
Traits info
{'name': 'age', 'left edge': 40.0, 'right edge': 95.0, 'max': {'value': 33.0, 'left edge': 59.25, 'right edge': 60.625}}
{'name': 'creatinine_phosphokinase', 'left edge': 23.0, 'right edge': 7861.0, 'max': {'value': 135.0, 'left edge': 23.0, 'right edge': 218.95}}
{'name': 'ejection_fraction', 'left edge': 14.0, 'right edge': 80.0, 'max': {'value': 49.0, 'left edge': 33.8, 'right edge': 35.45}}
{'name': 'platelets', 'left edge': 25100.0, 'right edge': 850000.0, 'max': {'value': 56.0, 'left edge': 251947.5, 'right edge': 272570.0}}
{'name': 'serum_creatinine', 'left edge': 0.5, 'right edge': 9.4, 'max': {'value': 82.0, 'left edge': 0.9450000000000001, 'right edge': 1.1675}}
{'name': 'serum_sodium', 'left edge': 113.0, 'right edge': 148.0, 'max': {'value': 40.0, 'left edge': 135.75, 'right edge': 136.625}}
```

Рисунок 3 – Информация о признаках

Стандартизация данных.

Стандартизируем первые 150 наблюдений с помощью StandartScaler и построим гистограммы. Результат на рис. 4.

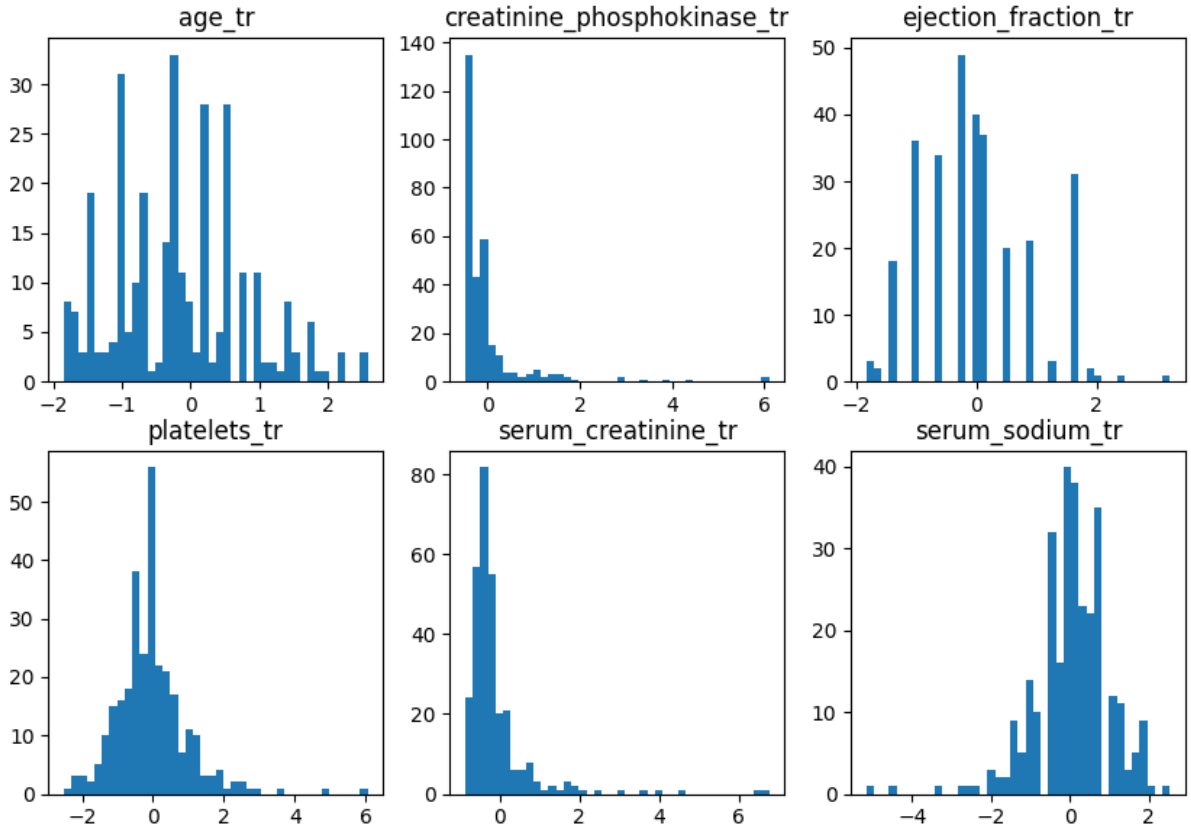

Рисунок 4 – Гистограммы после стандартизации первых 150 измерений

Далее стандартизируем все признаки и построим гистограммы. Результат на рис. 5.

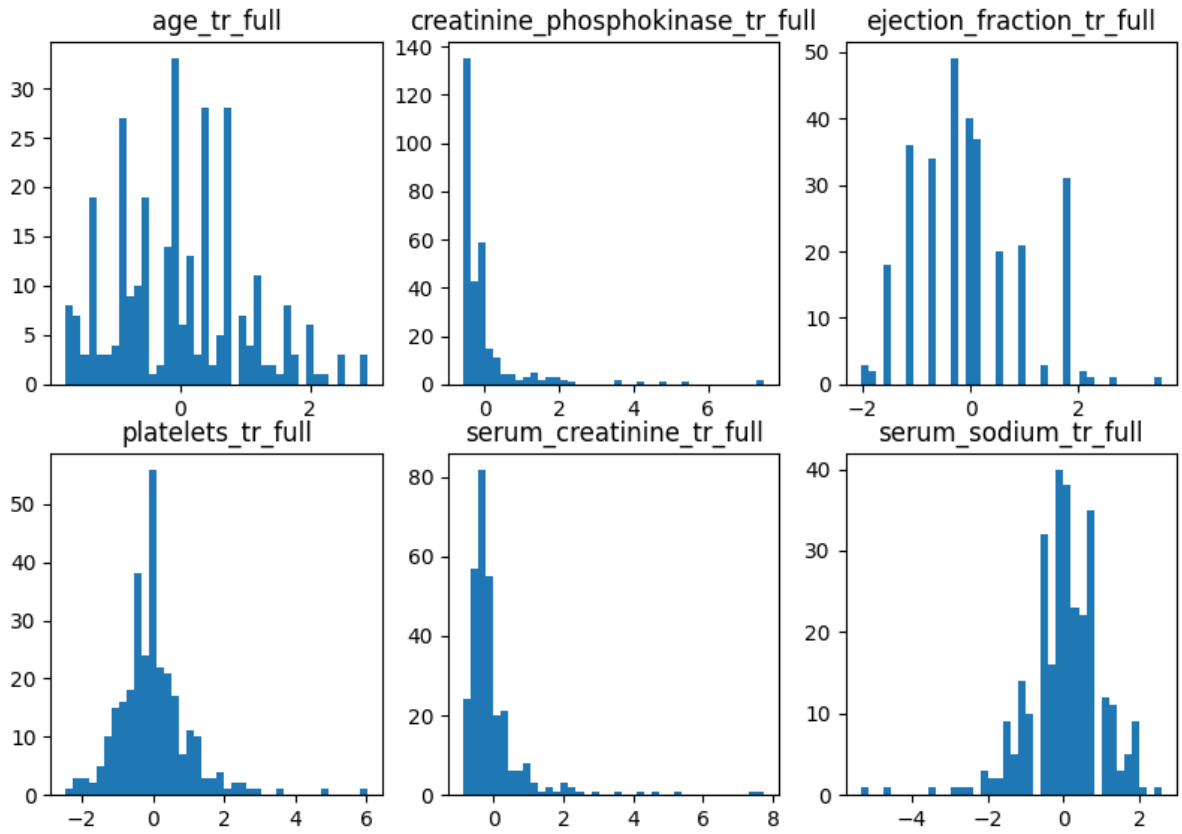

Рисунок 5 – Гистограммы после стандартизации

Были рассчитаны мат. ожидания и дисперсии данных до стандартизации, после стандартизации первых 150 элементов и после стандартизации всех элементов. Также были изучены данные о мат. ожидании и дисперсии внутри объектов StandartScaler. Результаты на рис. 6

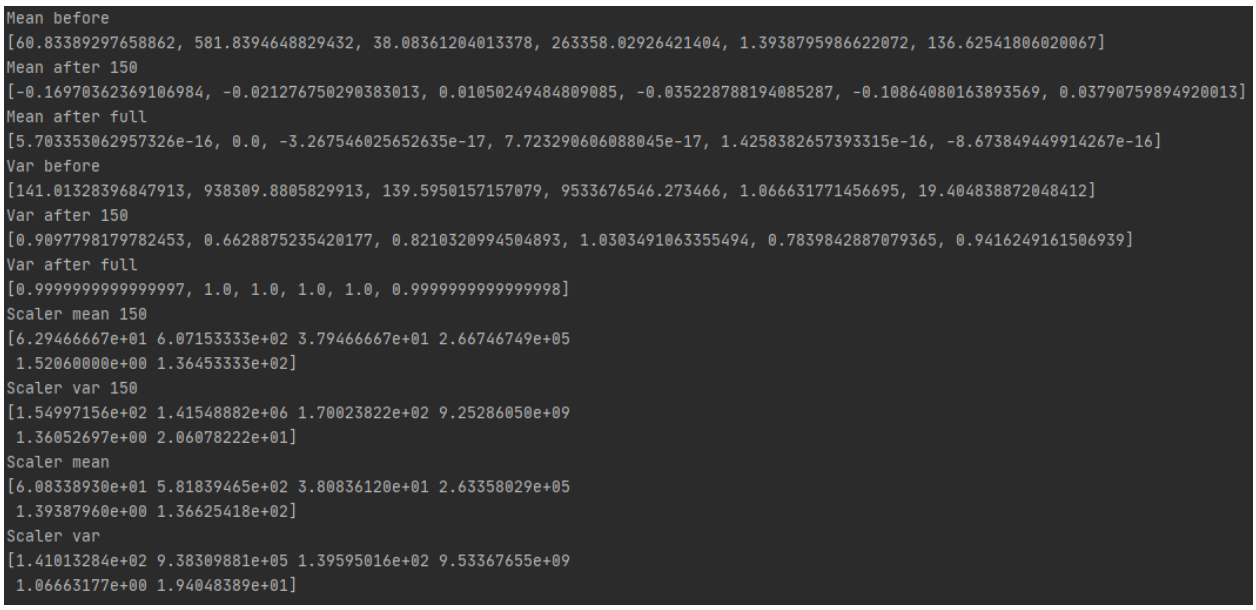

Рисунок 6 – Информация о мат. ожидании и дисперсии

Судя по полученным данным StandartScaler центрирует данные относительно мат. ожидания и нормирует относительно дисперсии, причем данные о сдвиге хранить в переменной mean_ а о скалировании в переменной var_. Можно сделать вывод что формулой нормировки является

$$Stand = \frac{Init - Mean}{Var}$$

Стандартизация по 150 элементам оказывается неполной так как в ней участвуют не все данные.

Приведение к диапазону.

Данные приведены к диапазону с помощью MinMaxScaler. Также извлечены параметры метода. Результаты на рис. 7 и 8.

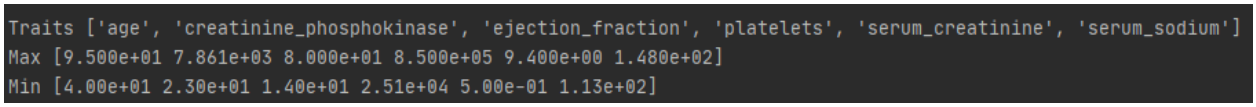

Рисунок 7 – Параметры MinMaxScaler

На гистограммах видно что данное преобразование масштабирует данные к промежутку [0, 1]. Преобразование производится по формуле

$$Scaled = \frac{Init - \min(Init)}{\max(Init) - \min(Init)} * \left(MaxRange(def = 1) - MinRange(def = 0)\right) + MinRange(\text{def} = 0)$$

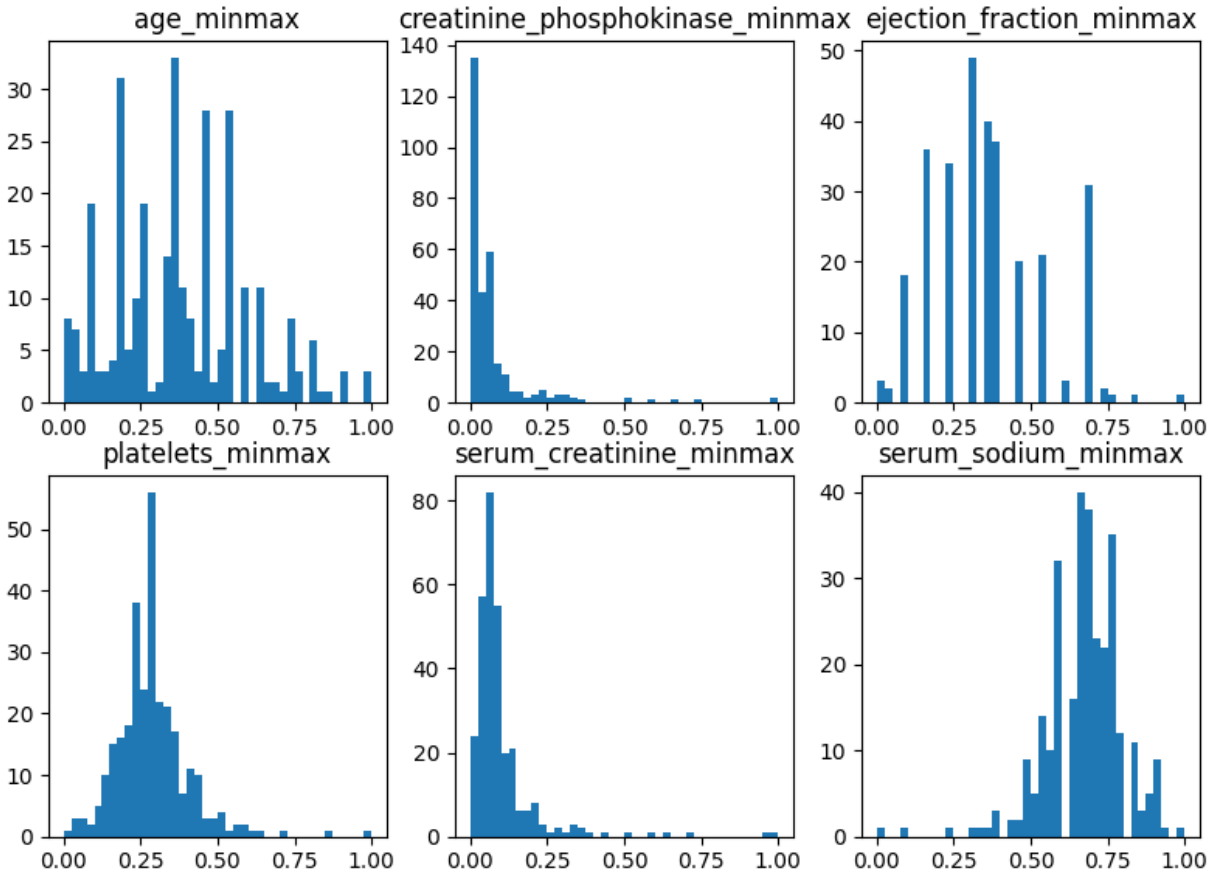

Рисунок 8 – Преобразование MinMaxScaler

Похожим образом было проведено преобразование с помощью MaxAbsScaler и RobustScaler. Результаты на рис. 9 и 10.

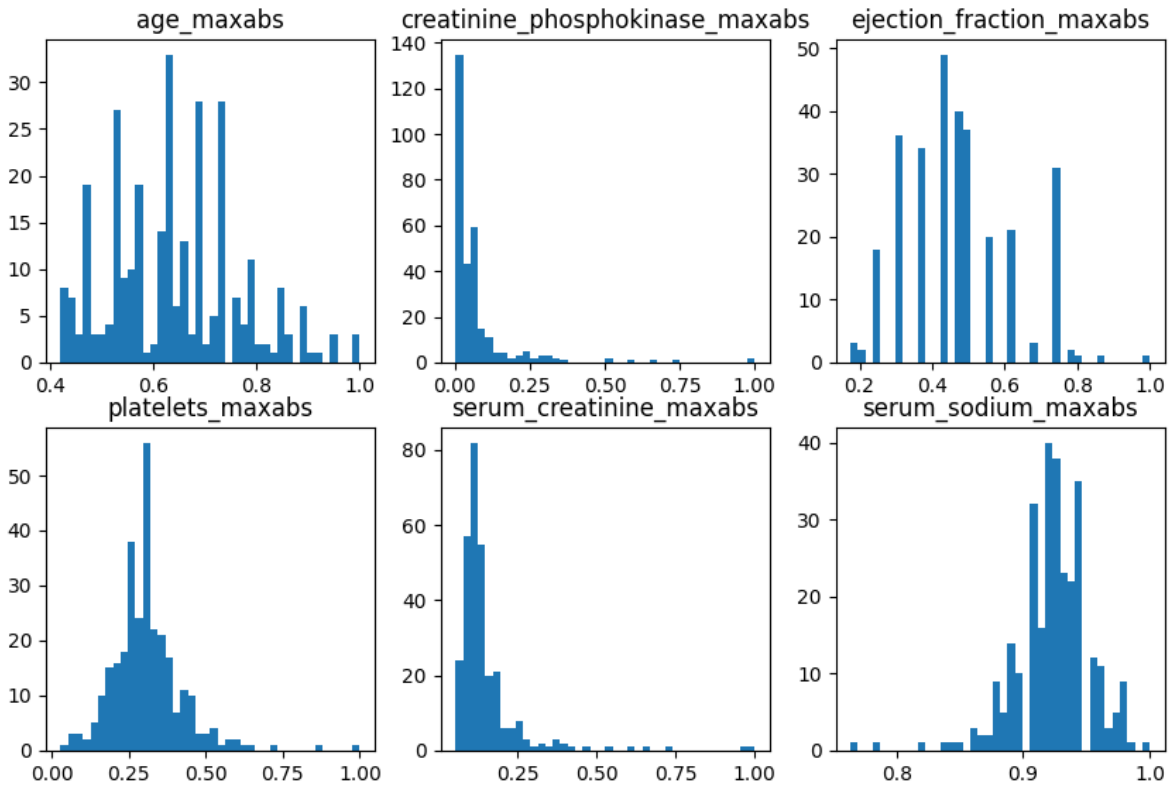

Рисунок 9 – Преобразование MaxAbsScaler

Преобразование MaxAbsScaler скалирует данные относительно модуля наибольшего значения по формуле

$$Scaled = \frac{Init}{\max\left(abs(Init)\right)}$$

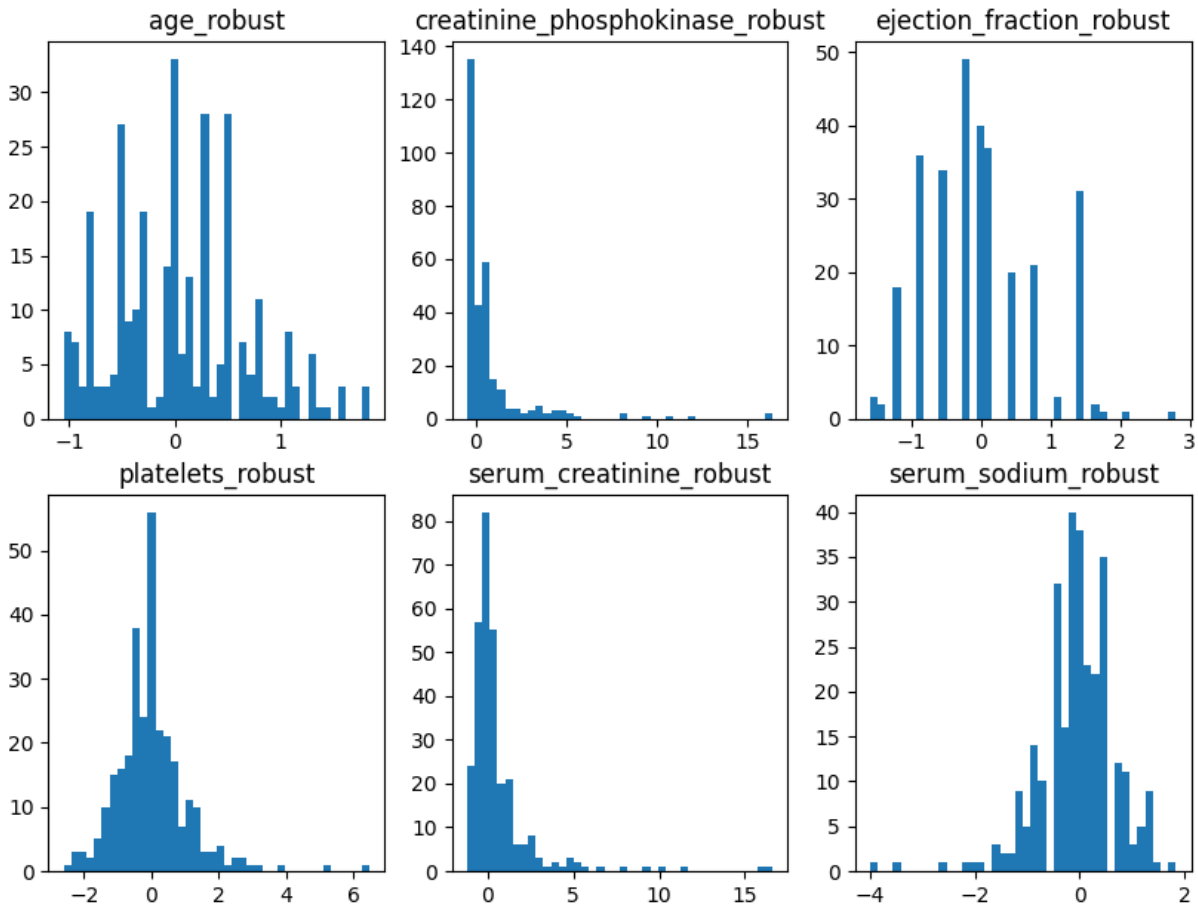

Рисунок 10 – Преобразование RobustScaler

Преобразование RobustScaler сдвигает данные по медиане и скалирует данные относительно IRQ по формуле

$$Scaled = \frac{Init - Median(Init)}{\text{IRQ(Init)}}$$

Написана функция приведения данных к промежутку [-5, 10] с помощью формулы

$$Scaled = \frac{Init - \min(Init)}{\max(Init) - \min(Init)} * \big(10 - (-5)\big) + (-5)$$

Результат выполнения функции на рис. 11.

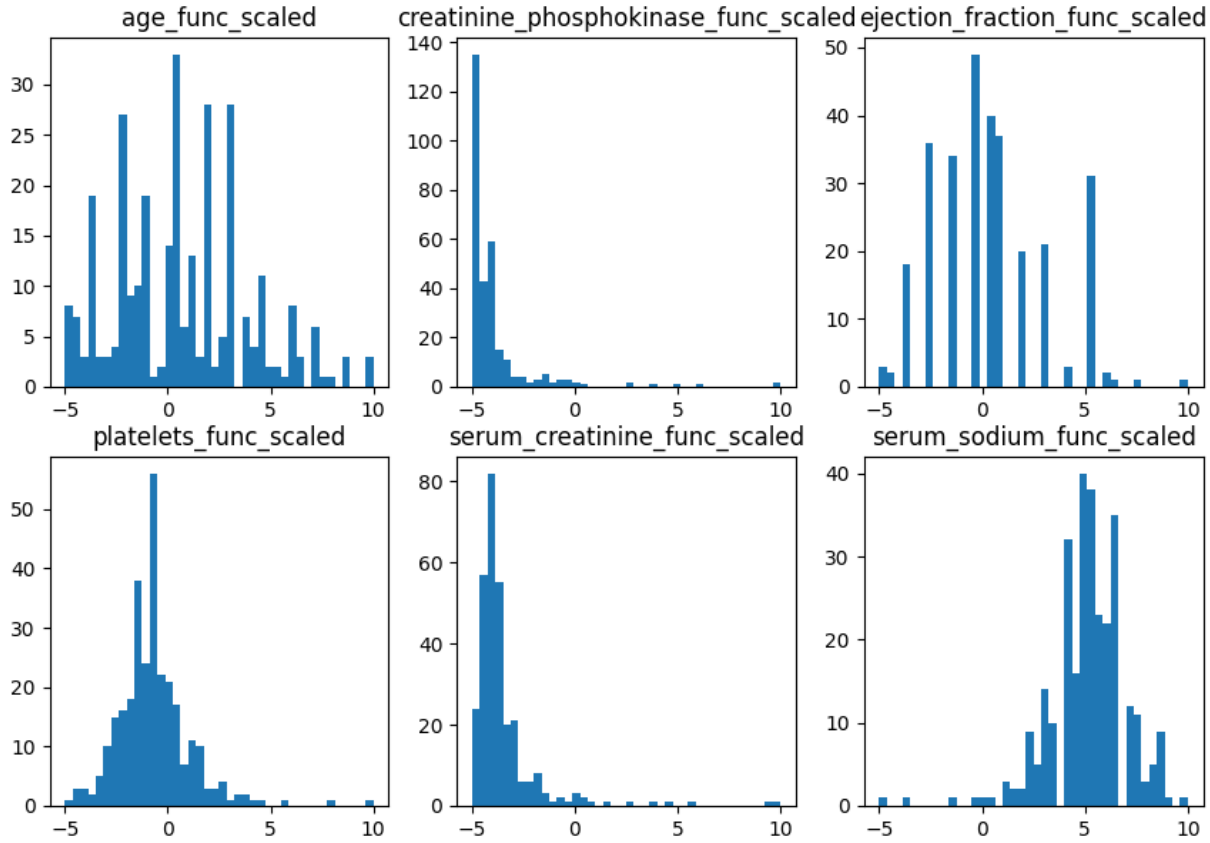

Рисунок 11 – Приведение к диапазону [-5, 10]

Нелинейные преобразования.

С помощью QuantileTransformer данные приведены к равномерному и нормальному распределениям. Результаты на рис. 12 и 13.

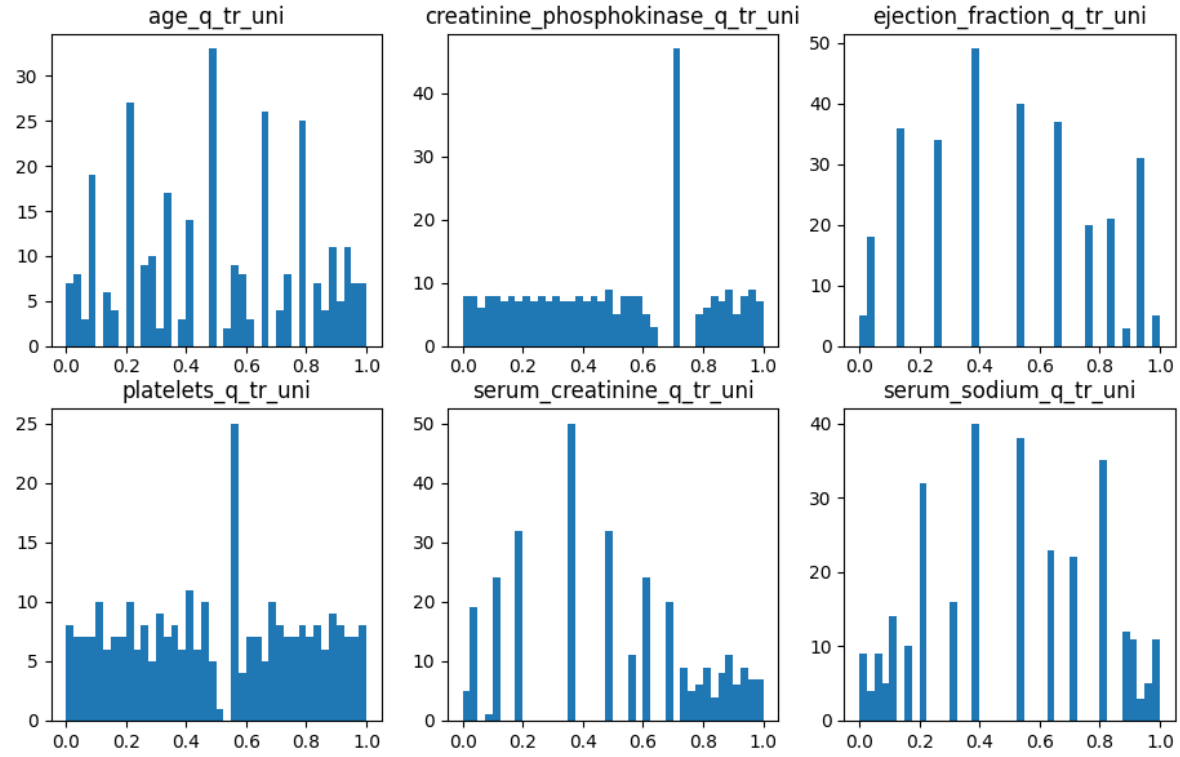

Рисунок 12 – Равномерное распределение с помощью QuantileTransformer

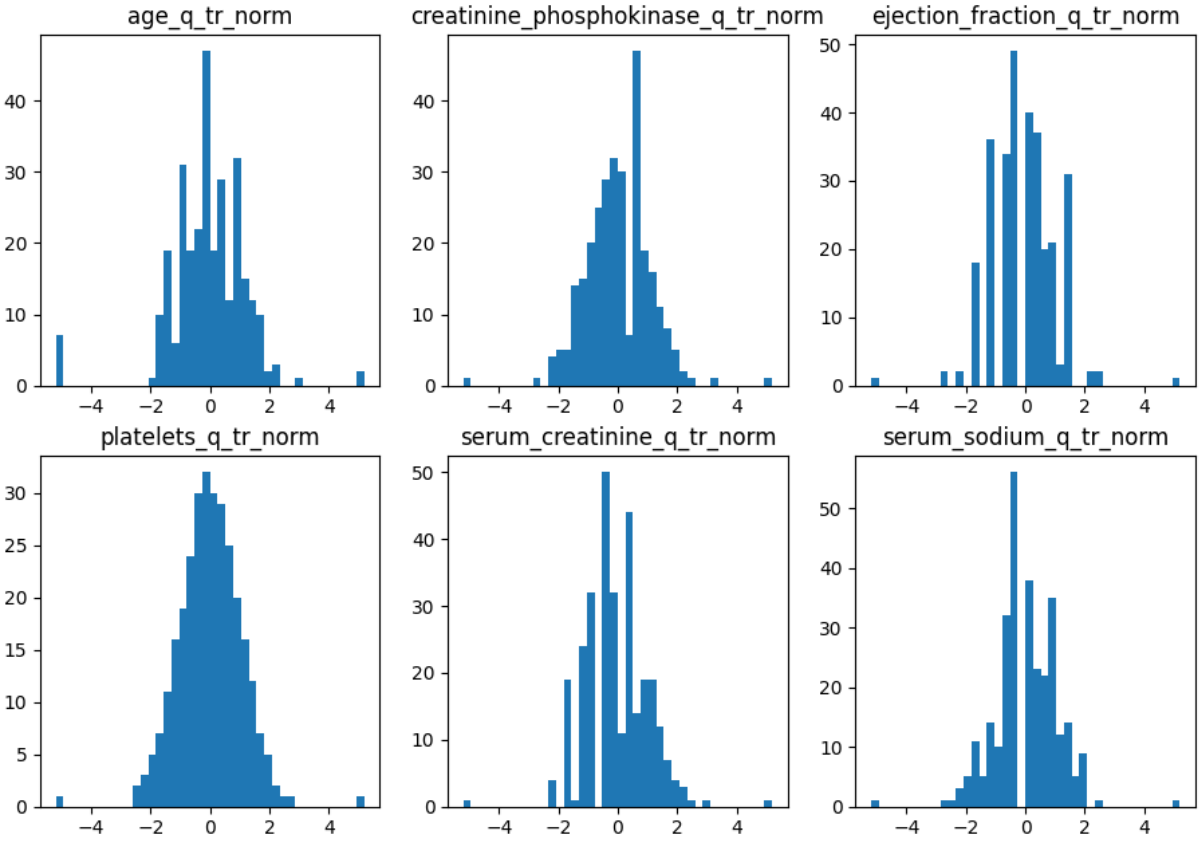

Рисунок 13 – Нормальное распределение с помощью QuantileTransformer

Параметр n_quantiles отвечает за частоту дискретизации функции распределения. Чем значение больше тем большее количество процентилей вычисляется.

Похожим образом данные приведены к нормальному распределению с помощью PowerTransformer. Результаты на рис. 14.

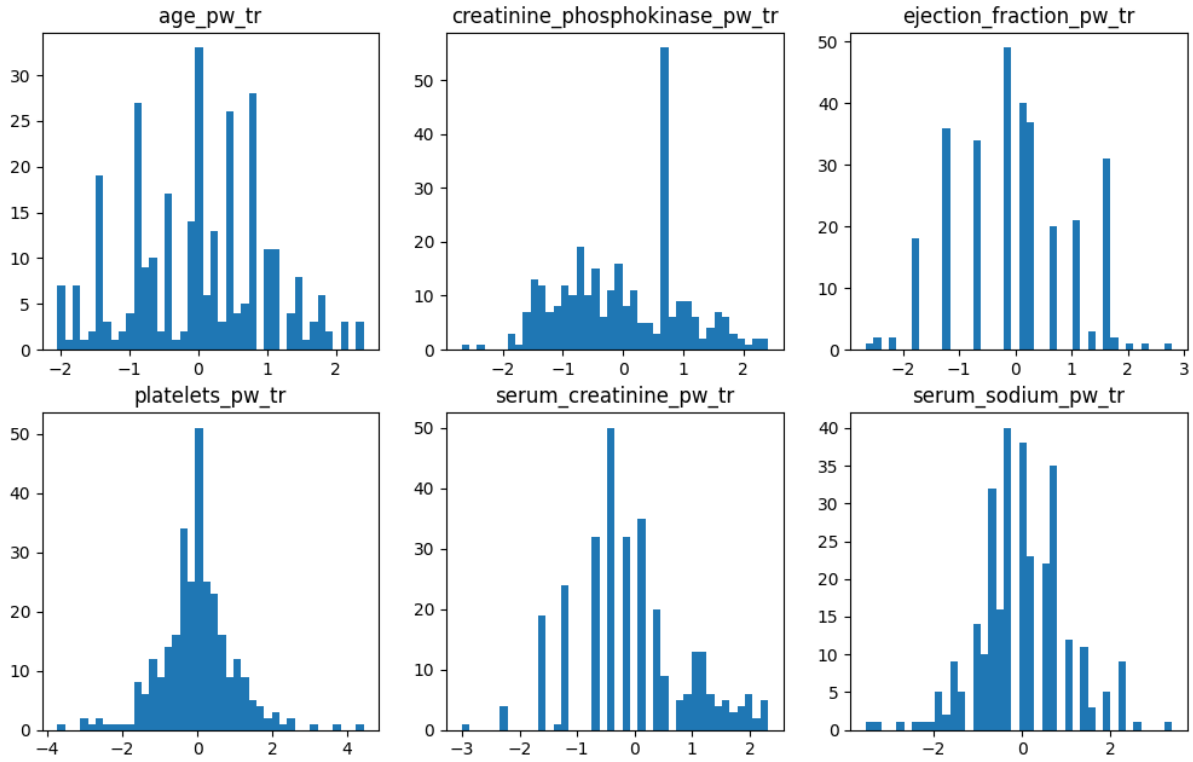

Рисунок 14 – Нормальное распределение с помощью PowerTransformer

Дискретизация признаков

Была проведена дискретизация признаков с помощью KBinsDiscretizer, преобразование похоже на стандартное построение гистограммы. Также были выведены диапазоны каждого интервала. Результаты на рис. 15 и 16.

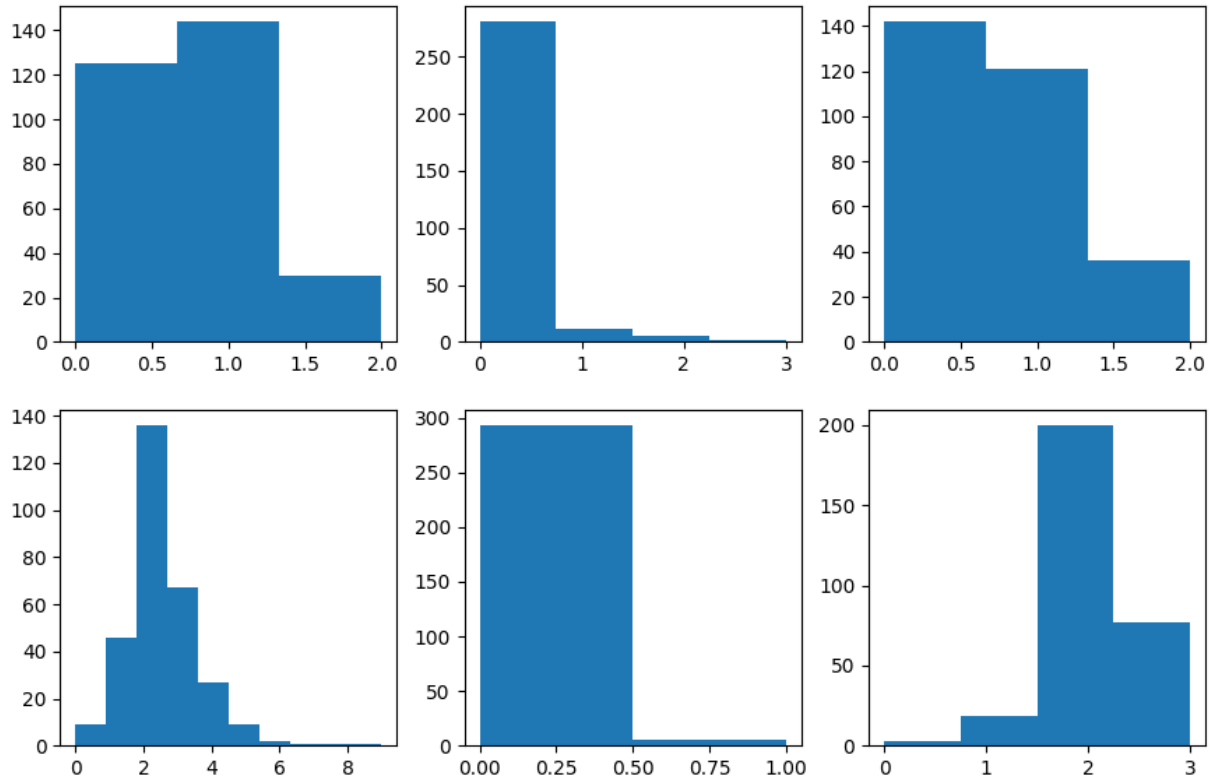

Рисунок 15 – Дискретизация с помощью KBinsDiscretizer

```
age [array([40.        , 58.33333333, 76.66666667, 95.        ])]
creatinine_phosphokinase [array([  23. , 1982.5, 3942. , 5901.5, 7861. ])]
ejection_fraction [array([14., 36., 58., 80.])]
platelets [array([ 25100., 107590., 190080., 272570., 355060., 437550., 520040.,
       602530., 685020., 767510., 850000.])]
serum_creatinine [array([0.5 , 4.95, 9.4 ])]
serum_sodium [array([113.  , 121.75, 130.5 , 139.25, 148.  ])]
```

Рисунок 16 – Диапазоны интервалов

Выводы

В ходе выполнения данной лабораторной работы были изучены подходы к предобработке данных с помощью алгоритмов библиотеки sklearn.

Была рассмотрена и изучена стандартизация.

Были рассмотрены методы скалирования и приведения к диапзону.

Были рассмотрены методы нелинейных преобразований данных.

Были рассмотрены методы дискретизации данных.