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. 8303 _____ Гришин К. И.
Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург
2021

Цель работы

Ознакомиться с методами предобработки данных из библиотеки *Scikit Learn*.

Ход выполнения работы

Загрузка данных

1. Загрузить датасет (https://www.kaggle.com/andrewmvd/heart-failure-clinical-data).
2. Создать Python скрипт, загружающий датасет и исключающий нечисловые признаки (отсортированные данные приведены на рис. 1).

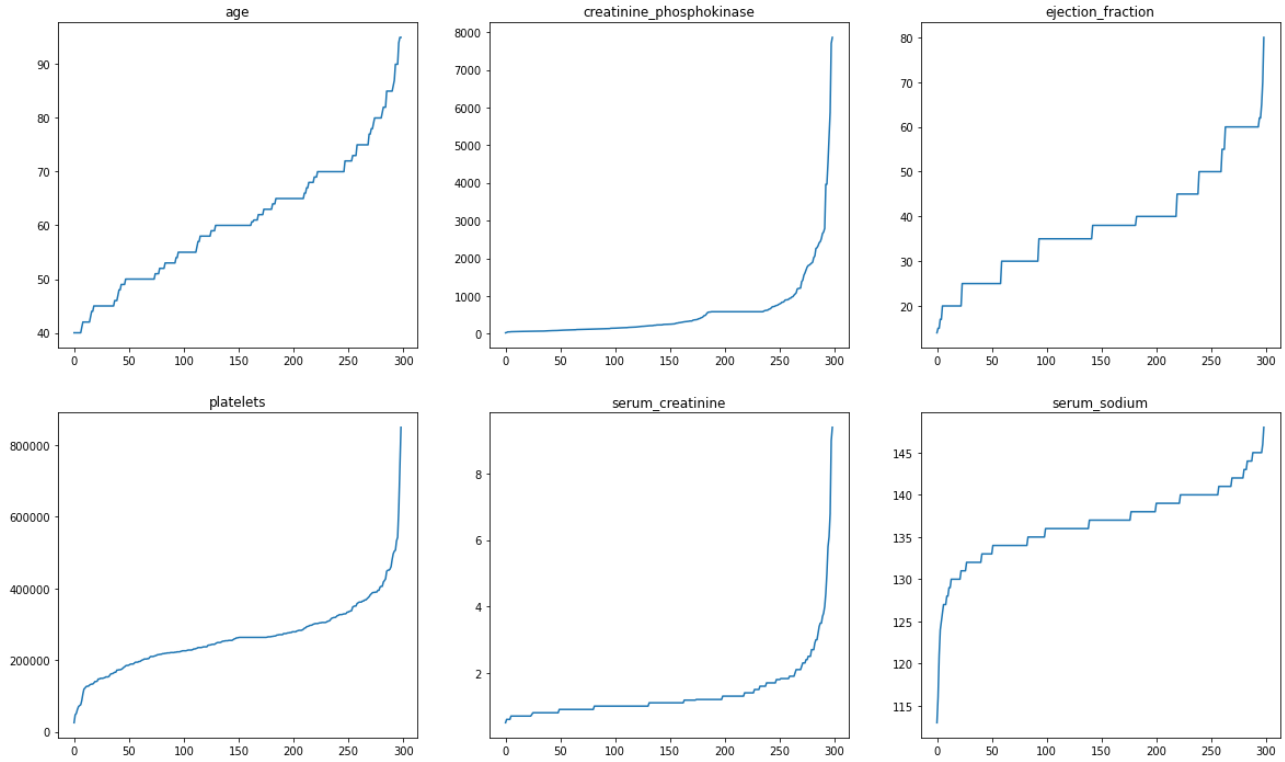

Рисунок 1. Загруженные данные.

3. Построить гистограммы признаков (рис. 2).

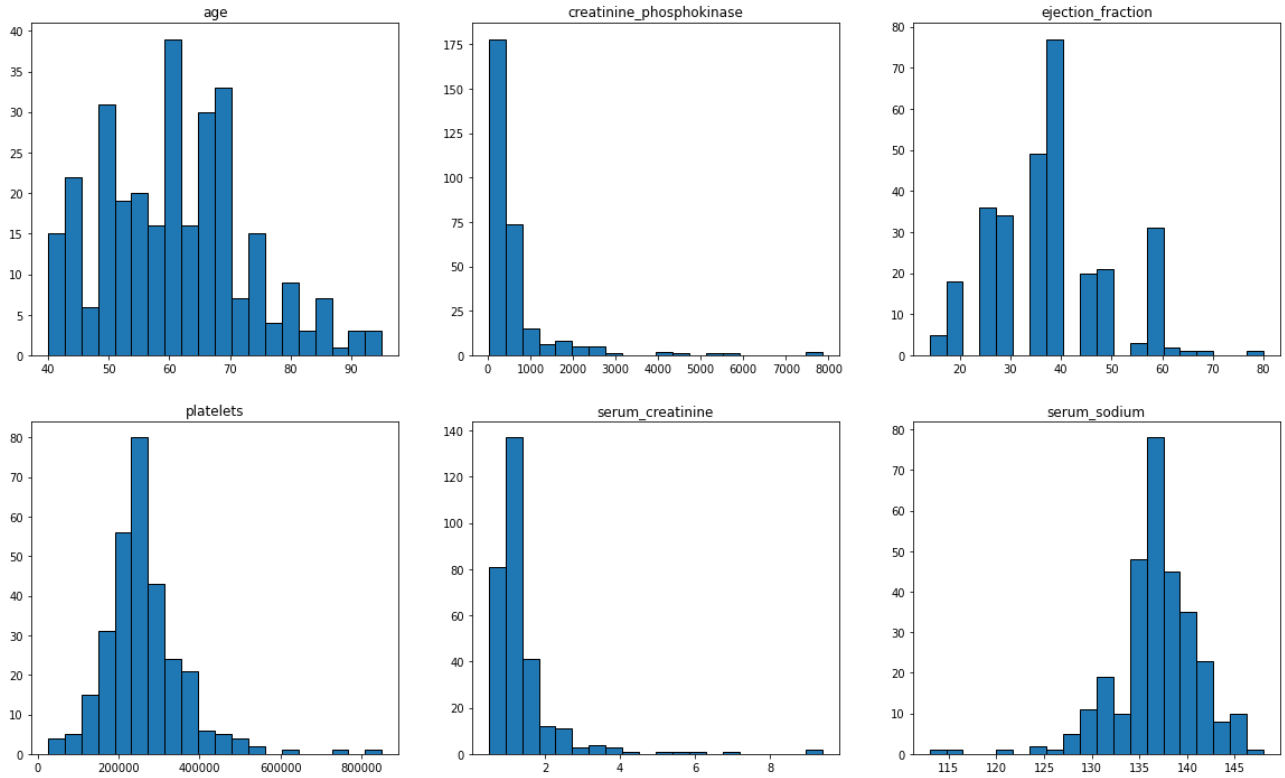

Рисунок 2. Гистограммы загруженных данных.

4. Определить диапазоны значений для каждого из признаков, а также возле какого значения лежит наибольшее количество наблюдений (табл. 1).

	Диапазон	Наибольшее количество наблюдений
age	(40, 95)	60
creatine_phosphokinase	(0, 7900)	0
ejection_fraction	(14, 80)	40
platelets	(25000, 850000)	225000
serum_creatine	(0.5, 9.2)	1
serum_sodium	(113, 148)	136

Таблица 1. Описание диапазонов значений.

5. Преобразовать *Pandas* датафрейм в массив *NumPy*.

Стандартизация данных

1. Подключить модуль *Sklearn* и настроить стандартизацию по первым 150 наблюдениям используя *StandartScaler*.

2. Стандартизировать данные.

3. Построить гистограммы признаков (рис. 3).

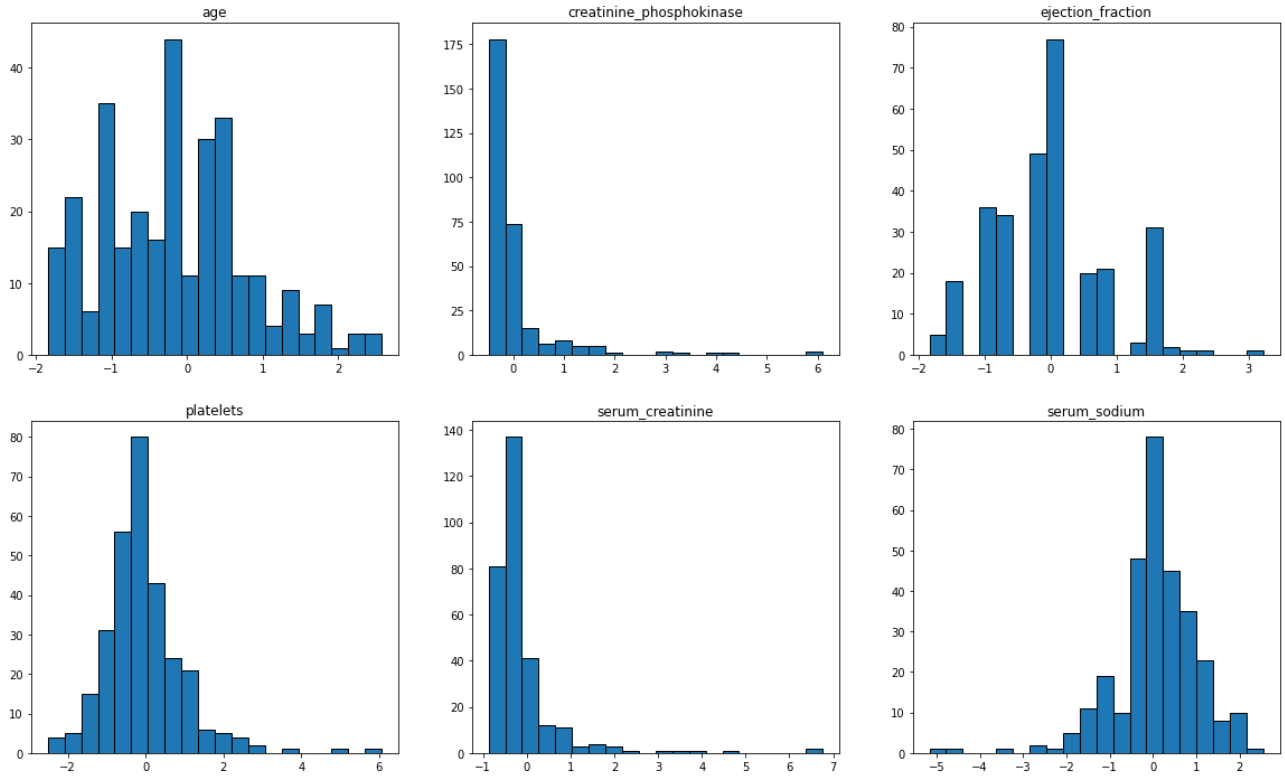

Рисунок 3. Гистограммы признаков стандартизированных данных.

4. Сравнить данные до и после стандартизации.

На основе полученных гистограмм составлена таблица с диапазонами наблюдений и значениями, около которых лежит наибольшее число наблюдений (табл. 2).

	Диапазон	Наибольшее количество наблюдений
age	(-1.75, 2.6)	-0.2
creatine_phosphokinase	(-0.5, 6.1)	-0.4
ejection_fraction	(-2.2, 3.2)	0.1
platelets	(-2.5, 6)	-0.2
serum_creatine	(-0.9, 6.6)	0.1
serum_sodium	(-5.1, 2.5)	0

Таблица 2. Описание диапазонов значений, стандартизированных по первым 150 наблюдениям.

Из табл. 2 видно, что наибольшее количество наблюдений находится возле 0, а сам разброс значений находится в пределах одного порядка.

5. Рассчитать мат. ожидание и СКО до и после стандартизации. Вывести формулы, по которым они стандартизировались (табл. 3).

	μ до	σ до	μ после	σ после
age	60.834	11.895	-0.170	0.955
creatine_phosphokinase	581.839	970.288	-0.021	0.816
ejection_fraction	38.084	11.835	0.011	0.908
platelets	263358.029	97804.237	-0.035	1.017
serum_creatine	1.394	1.035	-0.109	0.887
serum_sodium	136.625	4.412	0.038	0.972

Таблица 3. Мат. ожидание и СКО данных до и после стандартизации по первым 150 наблюдениям.

Исходя из таблицы 3, можно сделать вывод, что стандартизация данных приводит к тому, что математическое ожидание становится равным нулю, а среднеквадратичное отклонение равным 1. Из этого следует, что формула преобразования данных принимает вид:

$$y = \frac{x - \mu(X)}{\sigma(X)}$$

где X – исходный набор данных.

6. Сравненить со значениями объекта настроенной стандартизации (табл. 4).

	mean_	var_	σ
Age	62.95	155.00	12.45
creatine_phosphokinase	607.15	1415488.82	1189.74
ejection_fraction	37.95	170.02	13.04
Platelets	266746.8	9.252860e+09	96191.79
serum_creatine	1.52	1.36	1.17
serum_sodium	136.45	20.61	4.54

Таблица 4. Таблица значений объект-трансформатора, настроенного по первым 150 наблюдениям.

Из таблицы видно, что значения математического ожидания и дисперсии близки к соответствующим значениям исходного набора данных, однако все же отличаются. Из-за этих отличий и наблюдается математическое ожидание и среднеквадратичное отклонение стандартизированного набора данных от 0 и 1 соответственно.

7. Настроить стандартизацю на всех данных. Сравненить со стандартизацией для первых 150 наблюдений.

   Сравнение исходного набора данных со стандартизированным приведено в таблице 5. Значения объекта настроенной стандартизации приведены в таблице 6.

	μ до	σ до	μ после	σ после
age	60.834	11.895	0	1.002
creatine_phosphokinase	581.839	970.288	0	1.002
ejection_fraction	38.084	11.835	0	1.002
platelets	263358.029	97804.237	0	1.002
serum_creatine	1.394	1.035	0	1.002
serum_sodium	136.625	4.412	0	1.002

Таблица 5. Мат. ожидание и СКО данных до и после стандартизации по всем наблюдениям.

	mean_	var_	σ
Age	60.83	141.01	11.87
creatine_phosphokinase	581.84	938309.88	968.66
ejection_fraction	38.08	139.60	11.82
Platelets	263358.03	9.533677e+9	97640.55
serum_creatine	1.39	1.07	1.03
serum_sodium	136.63	1.03	4.41

Таблица 6. Таблица значений объект-трансформатора, настроенного по всем наблюдениям

Настройка стандартизации по полной выборке привела к тому, что у всех наборов данных математическое ожидание равно 0, а среднеквадратичное отклонение 1 (несмотря на присутствующую погрешность вычислений, все данные имеют одинаковое значение, а значит – стандартизированы).

Приведение к диапазону

1. Привести данные к диапазону, используя *MinMaxScaler*.
2. Построить гистограммы признаков (рис. 4)

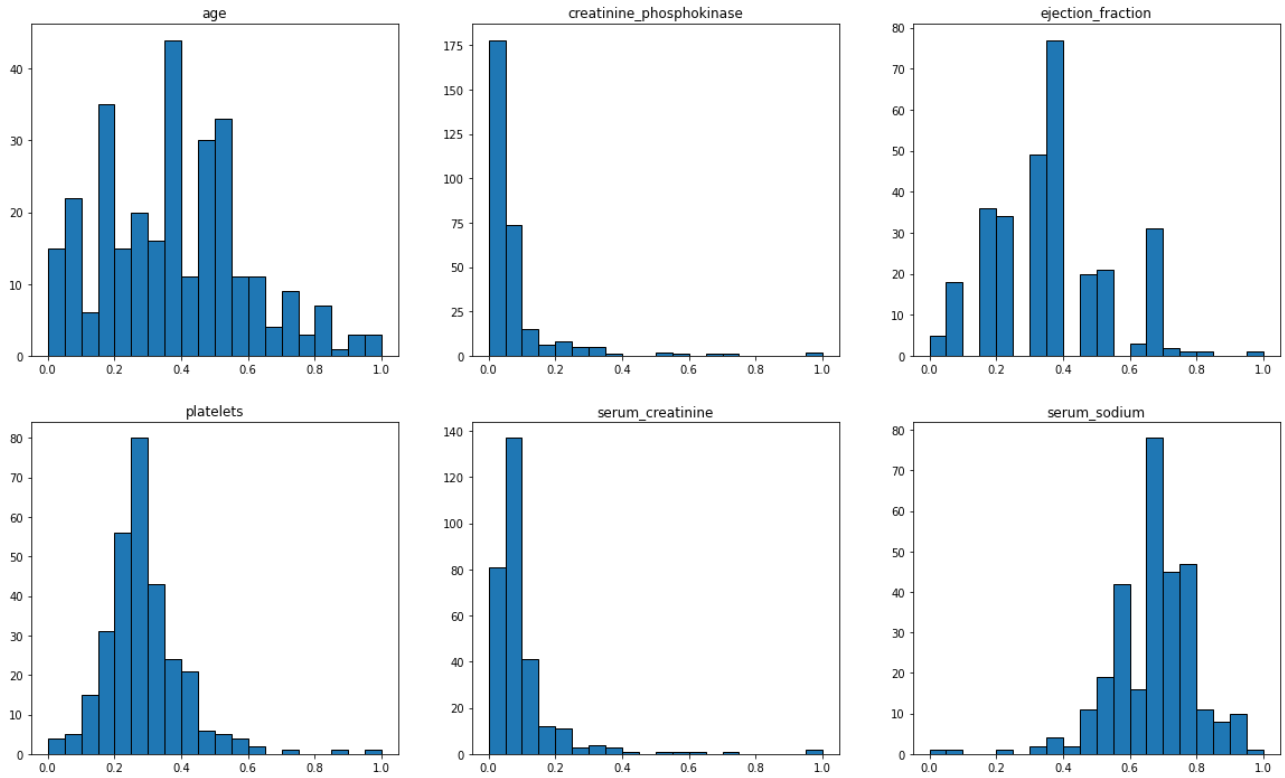

Рисунок 4. Гистограммы значений, отмасштабированных по диапазону с использованием *MinMaxScaler*.

3. Определить минимальное и максимальное значение для каждого признака (табл. 7)

	min до	max до	min после	max после
age	40	95	0	1
creatine_phosphokinase	23	7861	0	1
ejection_fraction	14	80	0	1
platelets	25100	850000	0	1
serum_creatine	0.5	9.4	0	1
serum_sodium	113	148.0	0	1

Таблица 7. Диапазоны значений каждого признака до и после *MinMaxScaler*.

4. Трансформировать данные используя *MaxAbsScaler* и *RobustScaler*. Построить гистограммы, определить диапазоны.

Гистограмма трансформации *MaxAbsScaler* приведена на рисунке 5.

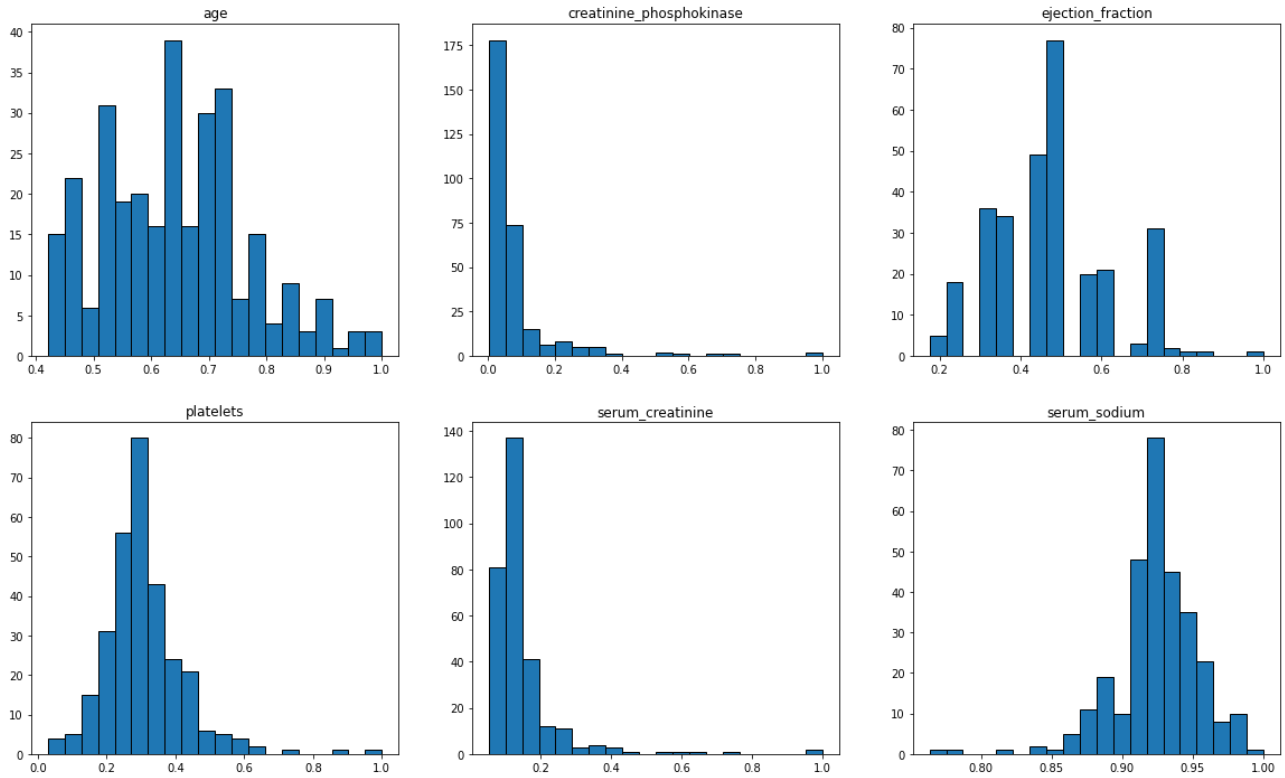

Рисунок 5. Гистограмма значений, отмасштабированных с использованием *MaxAbsScaler*.

Данные отмасштабированы так, что модуль максимального значения равен единице.

Гистограмма трансформации *RobustScaler* приведена на рисунке 6.

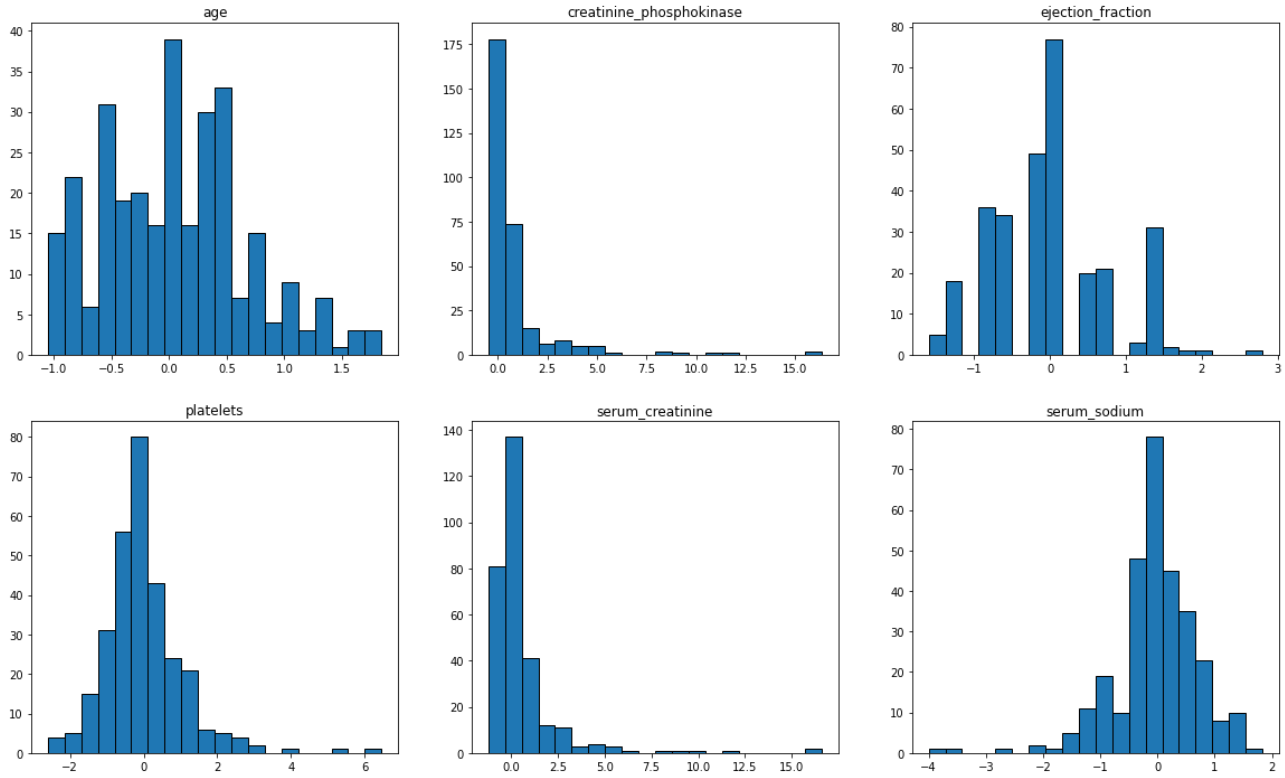

Рисунок 6. Гистограммы значений, отмасштабированных с использованием *RobustScaler*.

Данные отмасштабированы следующим образом. Сначала медиана устанавливается в ноль, затем конец первого и начало третьего квартиля становятся минимумом и максимумом соответственно.

5. Написать функцию, которая приведет данные к диапазону [-5, 10].

Для приведения к диапазону [-5, 10] (рис. 7) будет использоваться следующая формула:

$$y = \frac{x - \min(X)}{\max(X) - \min(X)}(10 + 5) - 5$$

```
1.  def min_max_scaler(data, min, max):
2.      d_min, d_max = min, max
3.
4.      def map_range(value, s_min, s_max, d_min, d_max):
5.          return (value - s_min) / (s_max - s_min) * (d_max - d_min) + d_min
6.
7.      def inner_min_max(values):
8.          s_min, s_max = np.min(values), np.max(values)
9.          l = lambda x: map_range(x, s_min, s_max, d_min, d_max)
10.         return np.array(list(map(l, values)))
11.
12.     return np.array(list(map(inner_min_max, data.T))).T
```

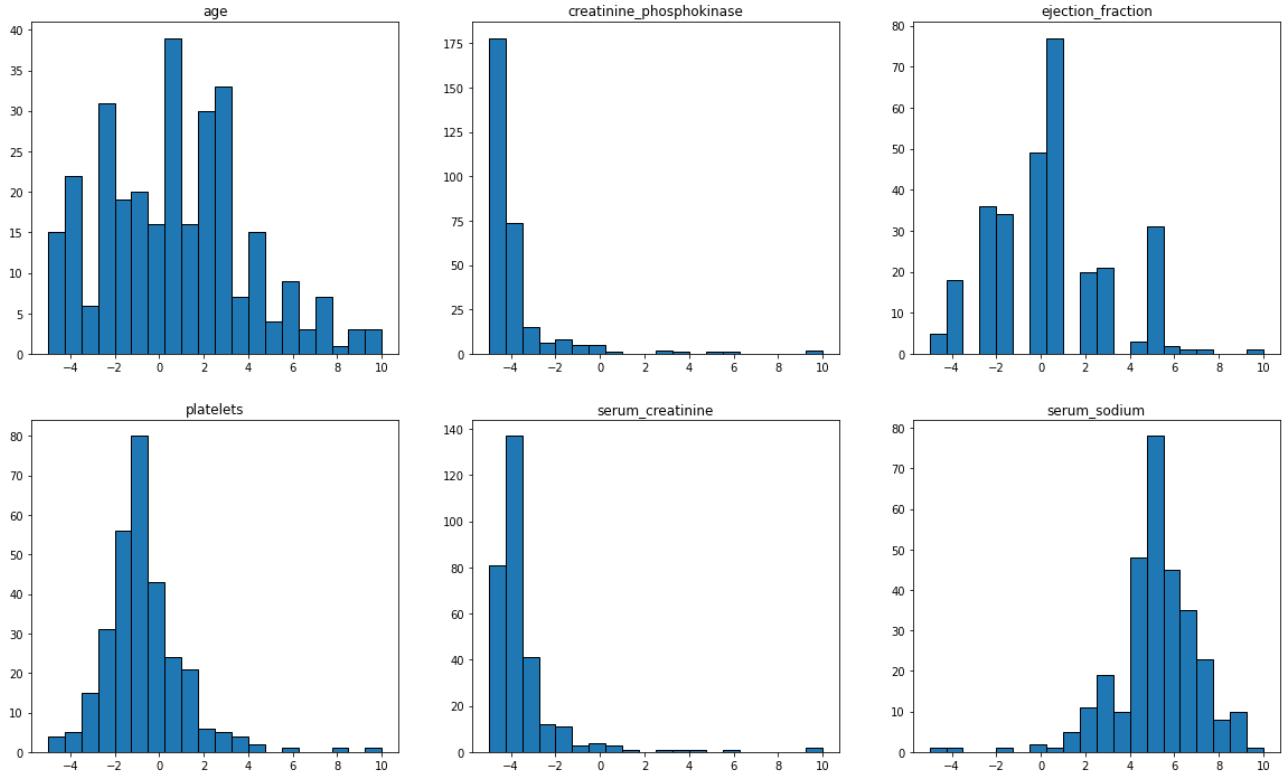

Рисунок 7. Гистограммы значений, отмасштабированных по диапазону с использованием собственной функции.

Нелинейные преобразования

1. Привести данные к равномерному распределению, используя *QuantileTransformer*.
2. Построить гистограммы (рис. 8)

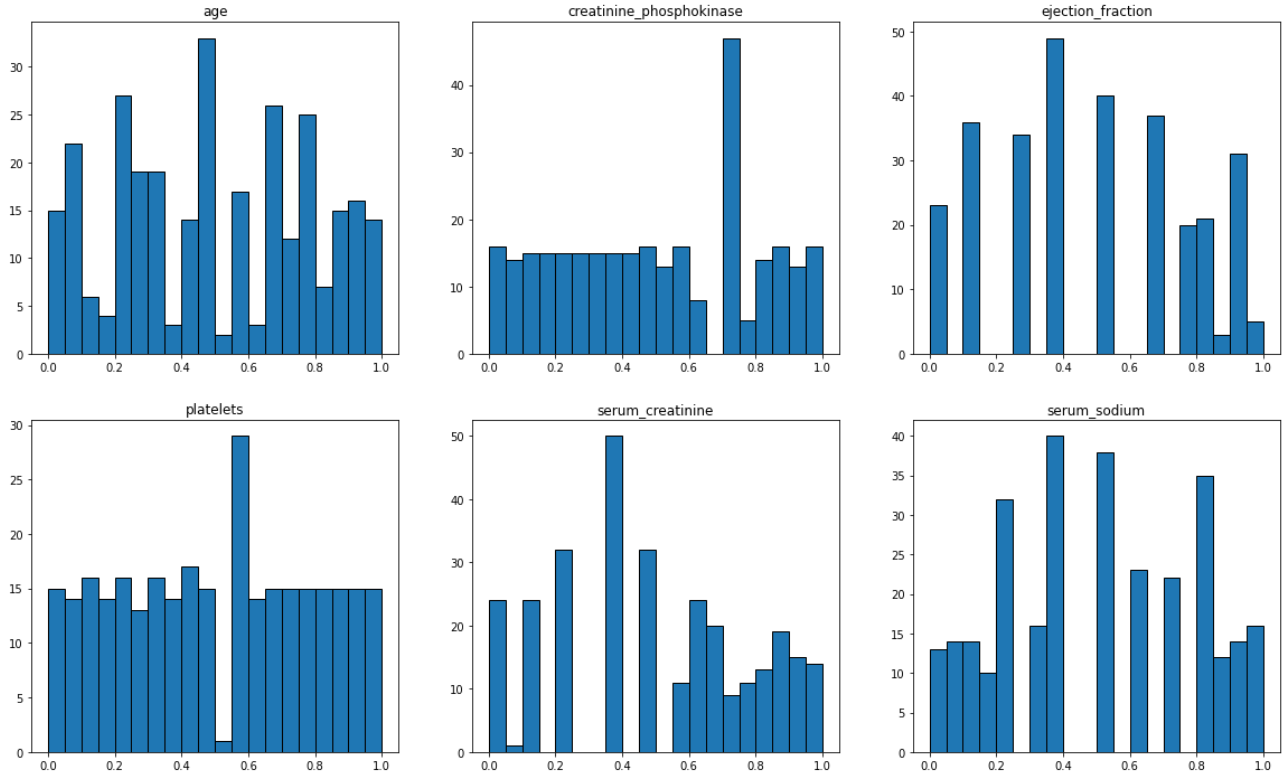

Рисунок 8. Гистограммы значений, приведенных к равномерному распределению с использованием *QuantileTransformer*.

3. Определить влияние параметра n_quantiles

   Параметр определяет количество квантилей, которое будет использовано для дискретизации функции распределения.

4. Привести данные к нормальному распределению, используя *QuantileTransformer*.

5. Построить гистограммы (рис. 9)

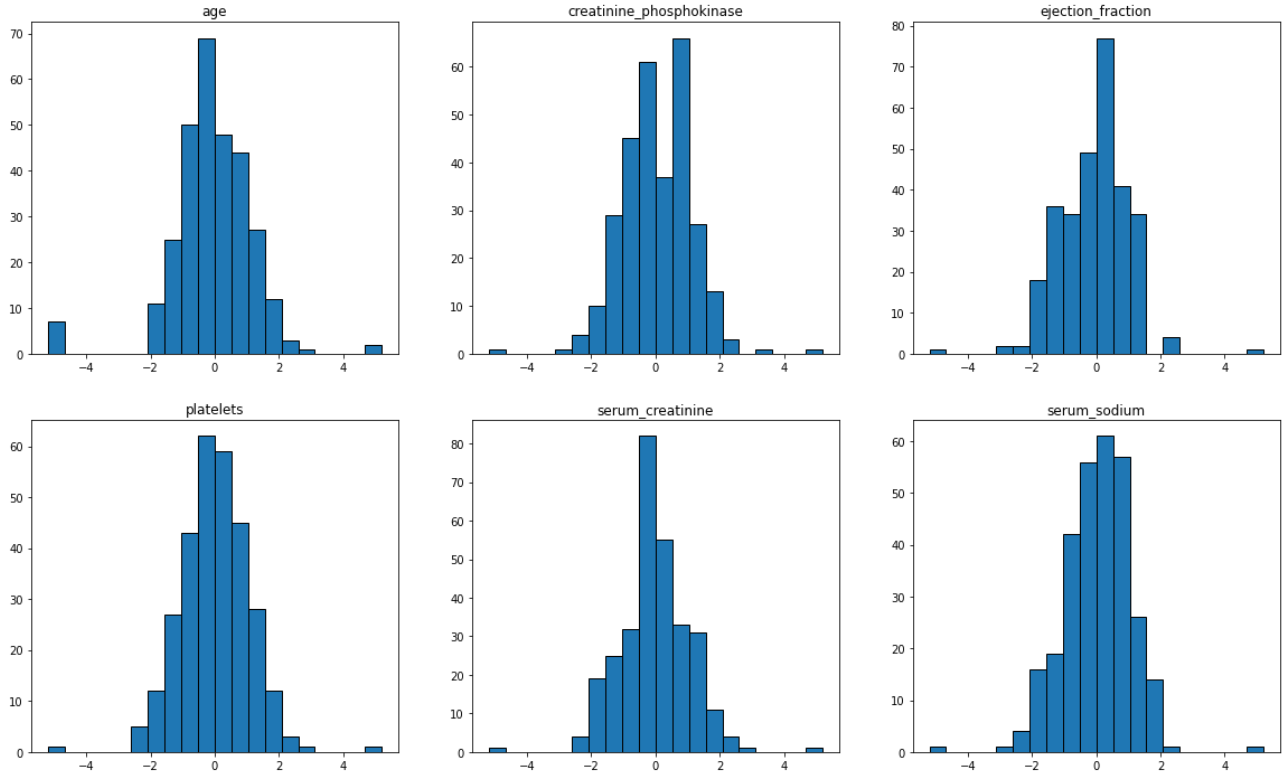

Рисунок 9. Гистограммы данных, приведеных к нормальному распределению с использованием QuantileTransformer.

6. Привести данные к нормальному распределению, используя *PowerTransformer*.

   Гистограмма приведена на рисунке 10.

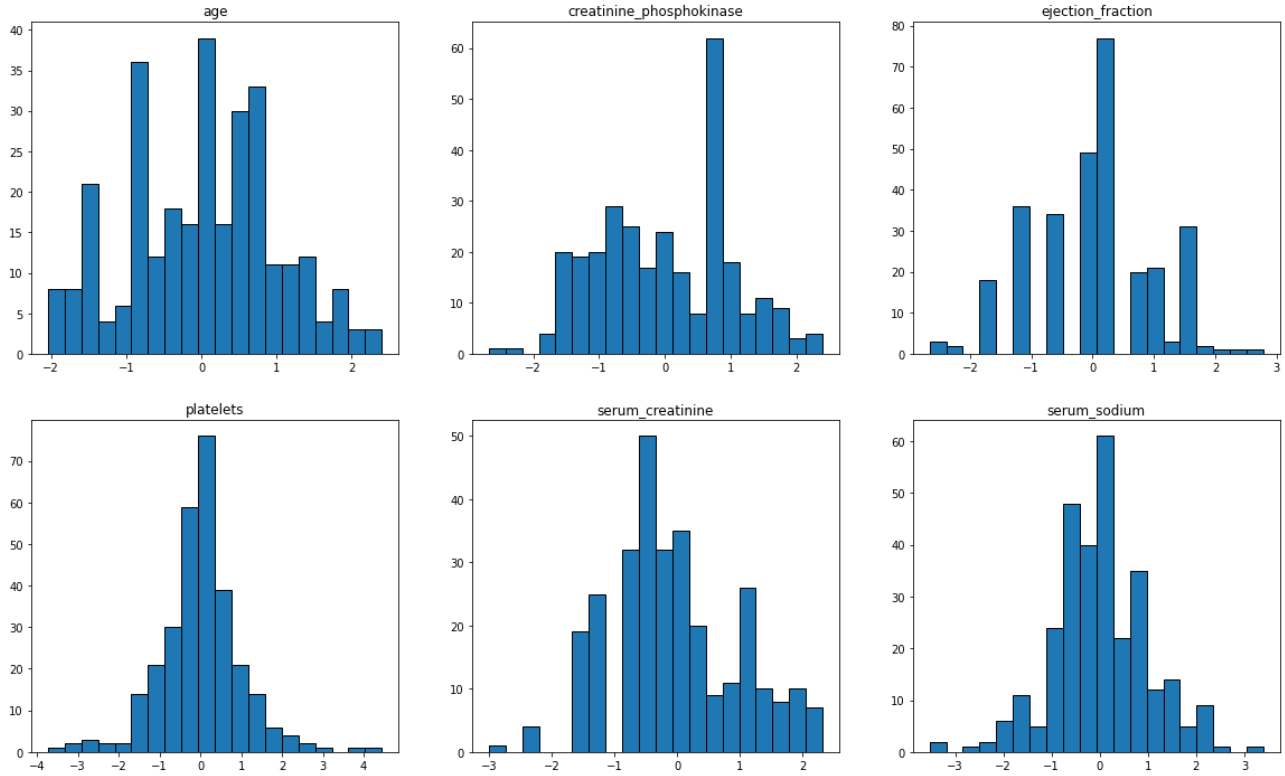

Рисунок 10. Гистограммы данных, приведеных к нормальному распределению с использованием PowerTransformer.

Дискретизация признаков

1. Провести дискретизацию признаков, используя KBinsDiscretizer.

   Данные дискретизированы на следующее количество диапазонов:

   - age – 3
   - creatine_phosphokinase – 4
   - ejection_fraction – 3
   - platelets – 10
   - serum_creatine – 2
   - serum_sodium – 4

2. Построить гистограммы (рис. 11). Объяснить результат.

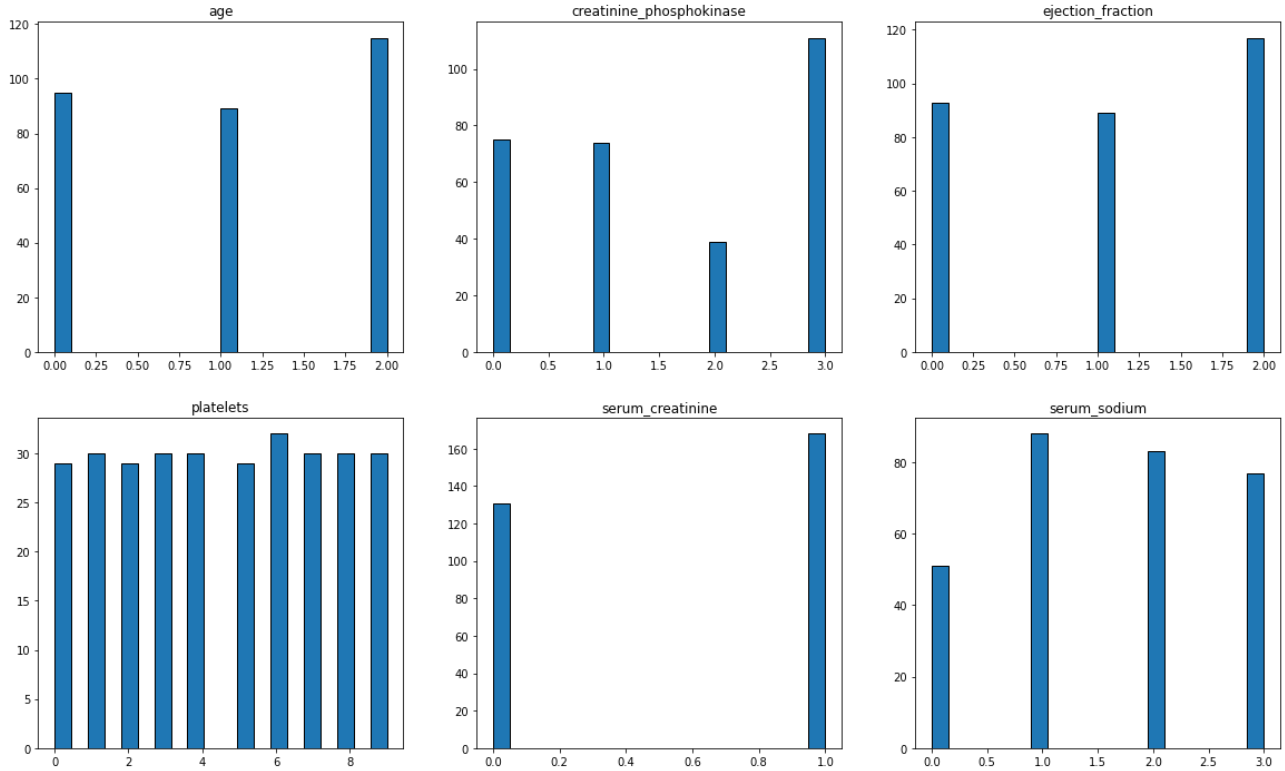

Рисунок 11. Гистограммы признаков, дискретизированных с использованием KBinsDiscretizer.

3. Вывести диапазоны каждого интервала (табл. 8).

age	[40, 55, 65, 95]
creatine_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_creatine	[0.5, 1.1, 9.4]
serum_sodium	[113, 134, 137, 140, 148]

Вывод

В ходе лабораторной работе был изучены методы предобработки данных. Изучен и применен инструмент *Scikit Learn*. Исходя из полученных результатов можно прийти к следующим заключениям:

- Трансформация по неполным значениям приводит к снижению качества выходных данных.
- Линейные преобразования изменяют диапазон наблюдений, но при этом никак не влияют на форму распределения.
- Нелинейные преобразования позволяют привести распределение к любой другой форме.