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 _____ Холковский К.В

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург

2021

Цель работы

Ознакомиться с методами предобработки данных из библиотеки Scikit Learn

Ход работы

1) Загрузка данных

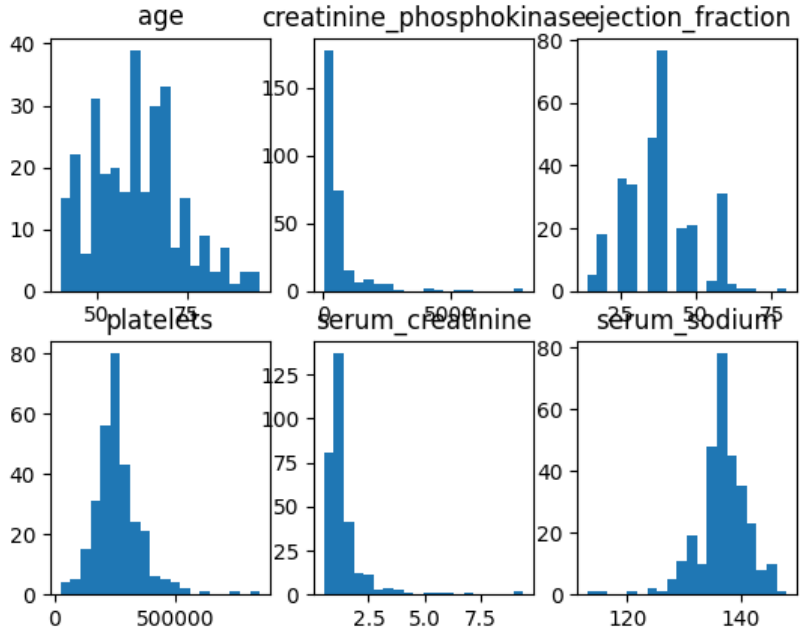

Рис 1 – гистограммы признаков

2) Стандартизация данных

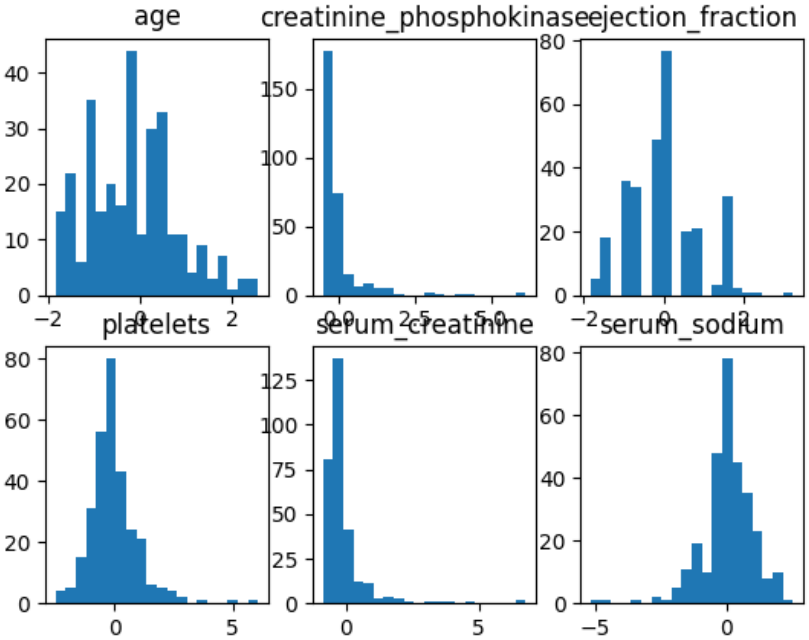

Рис 2 – Гистограммы стандартизированных данных по 150

Изменился диапазон и теперь среднее значение является нулем.

Таблица 1 – Сравнение данных стандартизации

		age	creatinine phosphokinase	ejection fraction	platelets	serum creatinine	serum sodium
До	МатОж	60.83	581.84	38.08	263358	1.39	136.63
	СКО	11.87	968.66	11.82	97640.5	1.03	4.41
[150] После	МатОж	-0.1697	-0.0213	0.0105	-0.0352	-0.1086	0.0379
	СКО	0.9538	0.8142	0.9061	1.0151	0.8854	0.9704
[150] Scaler	МатОж	62.95	607.15	37.95	266746.	1.52	136.45
	СКО	12.45	1189.74	13.04	96191.7	1.166	4.538
[299] После	МатОж	0	0	0	0	0	0
	СКО	1	1	1	1	1	1
[299] Scaler	МатОж	60.83	581.84	38.08	263358	1.39	136.63
	СКО	11.87	968.66	11.82	97640.5	1.03	4.41

Стандартизация проводилась по формуле:

$$\frac{x - mean}{std}$$

3) Приведение к диапазону

MinMaxScaler

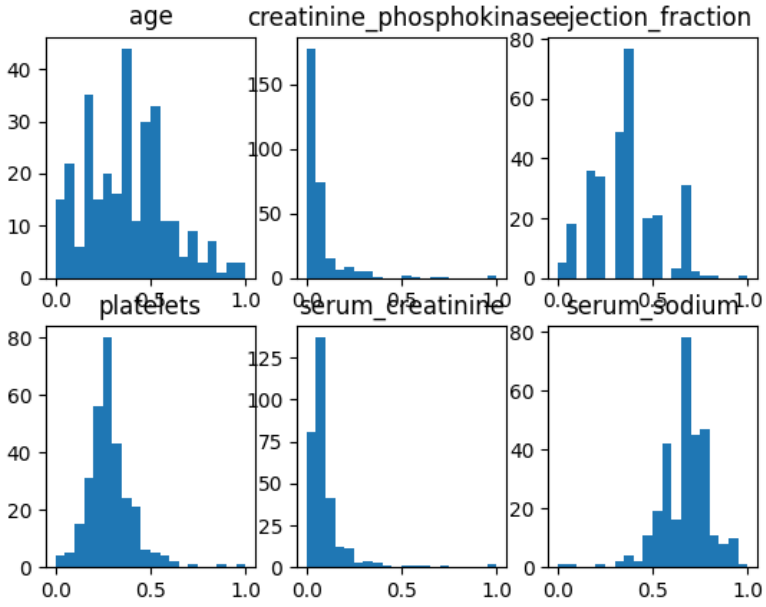

Рис 3 – Гистограммы мин макс

Данные приводятся к диапазону, где min это 0, а max это 1.

Таблица 2 – Данные в полях min и max

	age	creatinine phosphokinase	ejection fraction	platelets	serum creatinine	serum sodium
max	95	7861	80	850000	9.4	148
min	40	23	14	25100	0.5	113

MaxAbsScaler

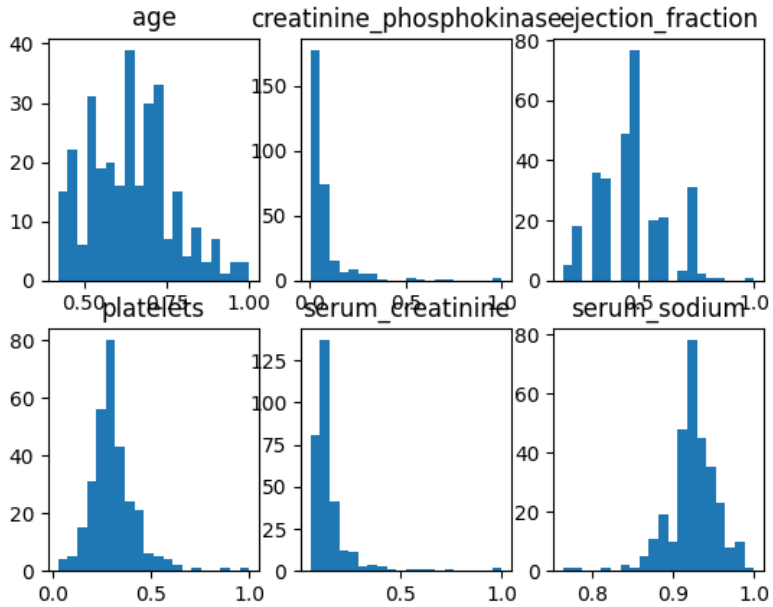

Рис 5 – гистограмма макс абс

Данные приводятся к диапазону, где max_abs_ берется как 1

Таблица 3 – Данные max_abs_

	age	creatinine phosphokinase	ejection fraction	platelets	serum creatinine	serum sodium
МаксАбс	95	7861	80	850000	9.4	148

RobustScaler

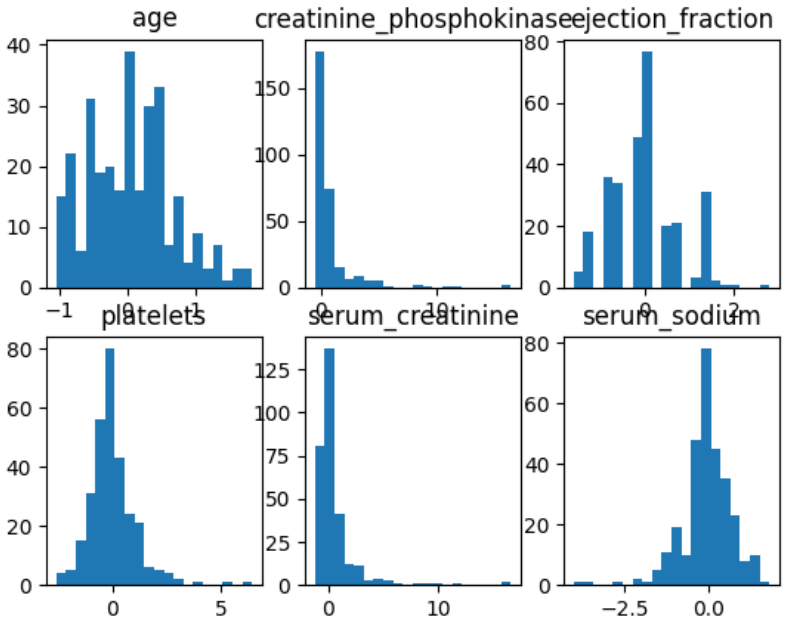

Рис 6 – Гистограмма Робус

Удаляет медианное значение и масштабирует данные в соответствии с квартильным диапазоном.

Таблица 4 – Данные center_

	age	creatinine phosphokinase	ejection fraction	platelets	serum creatinine	serum sodium
центер	60	250	38	262000	1.1	137

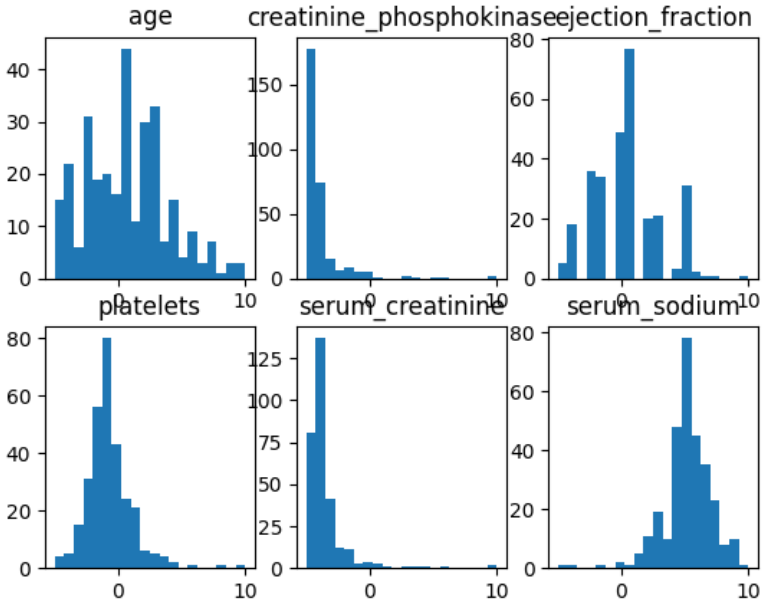

Рис 7 – Гистограммы для диапазона [-5; 10]

4) Нелинейные преобразования

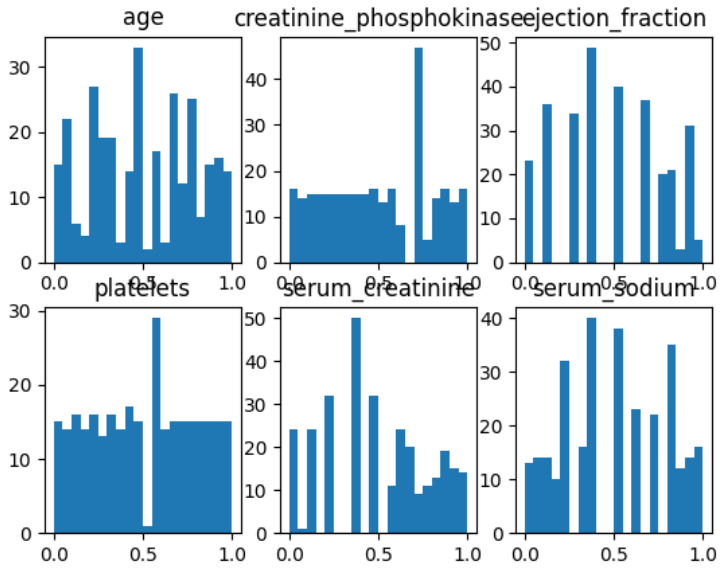

Рис 8 – Гистограмма преобразованных данных

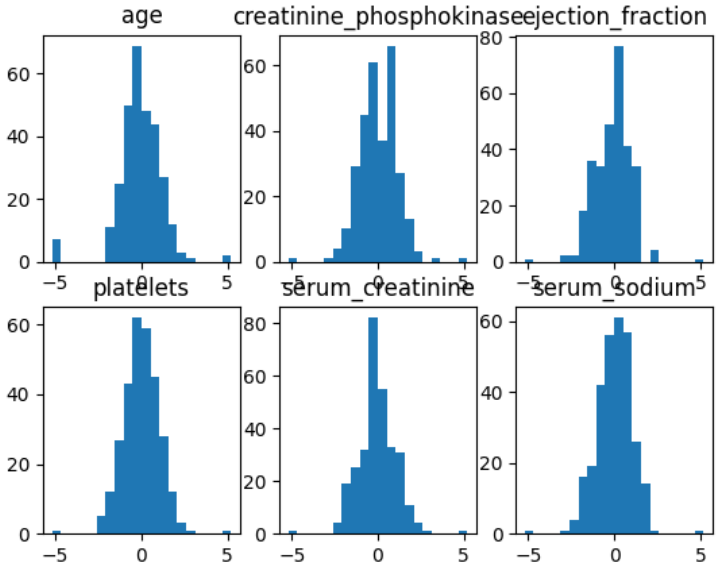

Рис 9 - Гистограмма преобразованных данных в нормальном распределении

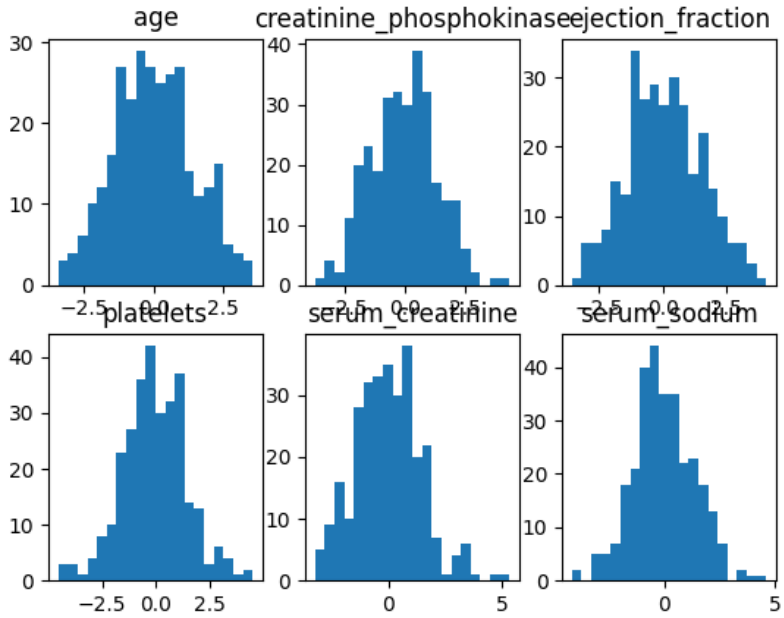

Рис 10 – Нормальное распределение при использовании PowerTransformer

5) Дискретизация признаков

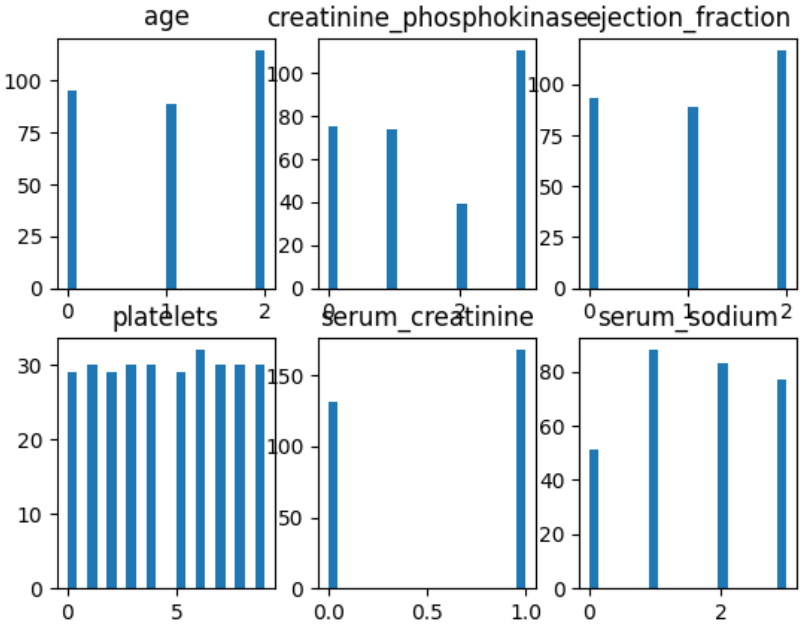

Рис 11 - Гистограммы дискретизации

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

Рис 12 – Диапазоны для каждого признака