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. 6304 _____ Пискунов Я.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Цель работы.

Ознакомиться с методами предобработки данных из библиотеки Scikit Learn.

Ход работы.

Датасет скачан и загружен в датафрейм согласно инструкции. Исключены бинарные признаки и признаки времени. После всех манипуляций действительно осталось 299 рядов и 6 столбцов. На рис. 1 представлены гистограммы по каждому из признаков.

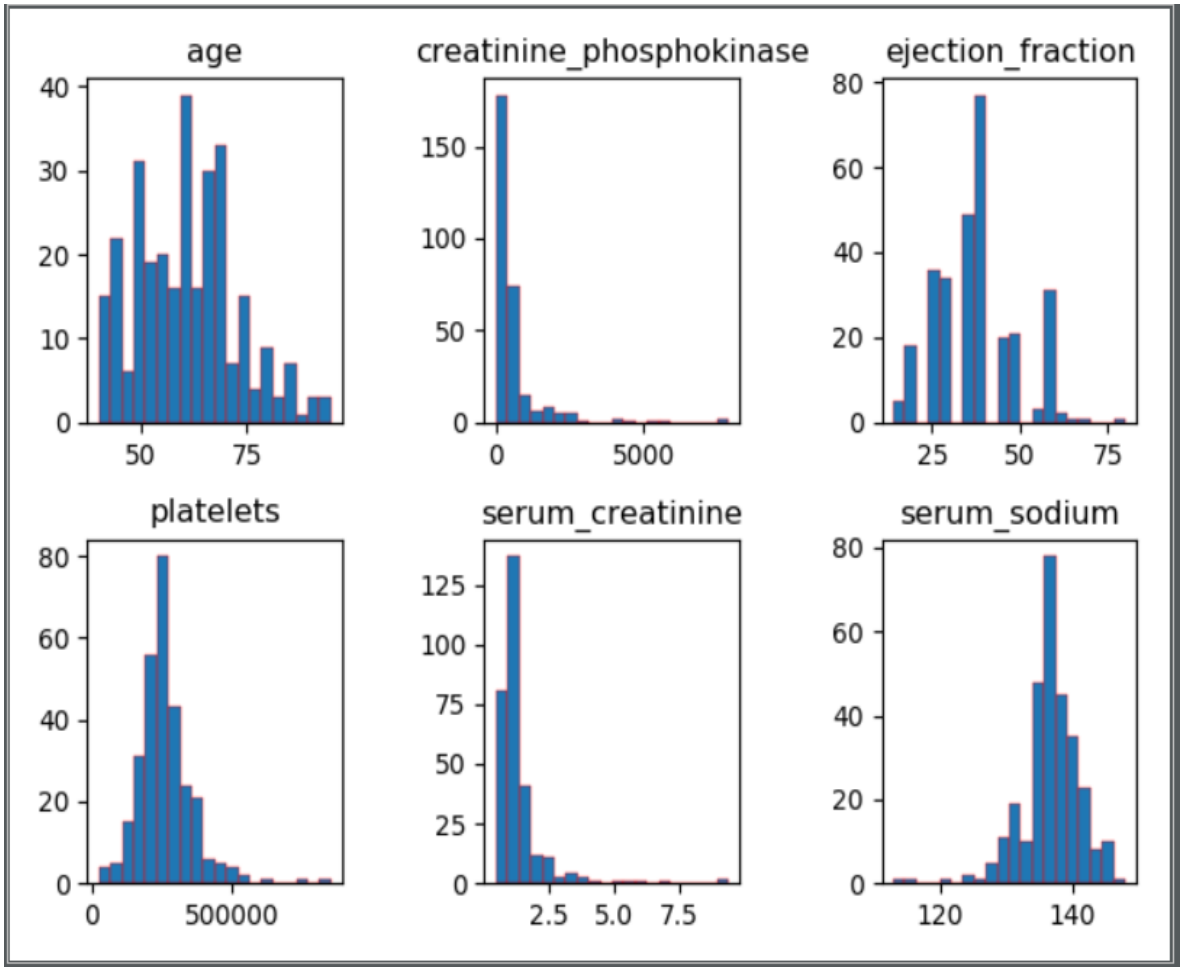

Рисунок 1 – Гистограммы признаков

На основании полученных гистограмм определены границы диапазонов значений признаков, а также значение, возле которого лежит наибольшее количество наблюдений. Эти данные представлены в табл. 1.

Далее датафрейм приведен к формату numpy.

Таблица 1. Диапазоны и значения с наибольшим количеством наблюдений

Признак	**Диапазон значений**	**Значение с наибольшим количеством наблюдений**
age	(40, 100)	60
creatinine_phosphokinase	(0, 8000)	0
ejection_fraction	(0, 80)	40
platelets	(0, 900000)	250000
serum_creatinine	(0, 10)	1.25
serum_sodium	(110, 150)	137

После выполнена стандартизация данных на основе первых 150 результатов. По итогам стандартизации построены гистограммы признаков. Они представлены на рис. 2. Новые значения диапазонов и значения с наибольшим количеством наблюдений представлены в табл. 2.

Таблица 2. Диапазоны и значения после стандартизации

Признак	**Диапазон значений**	**Значение с наибольшим количеством наблюдений**
age	(-2, 2.5)	-0.2
creatinine_phosphokinase	(-0.3, 6.3)	-0.3
ejection_fraction	(-2, 3)	0
platelets	(-3, 6)	0
serum_creatinine	(-1.5, 7)	-1
serum_sodium	(-5.1, 3)	0

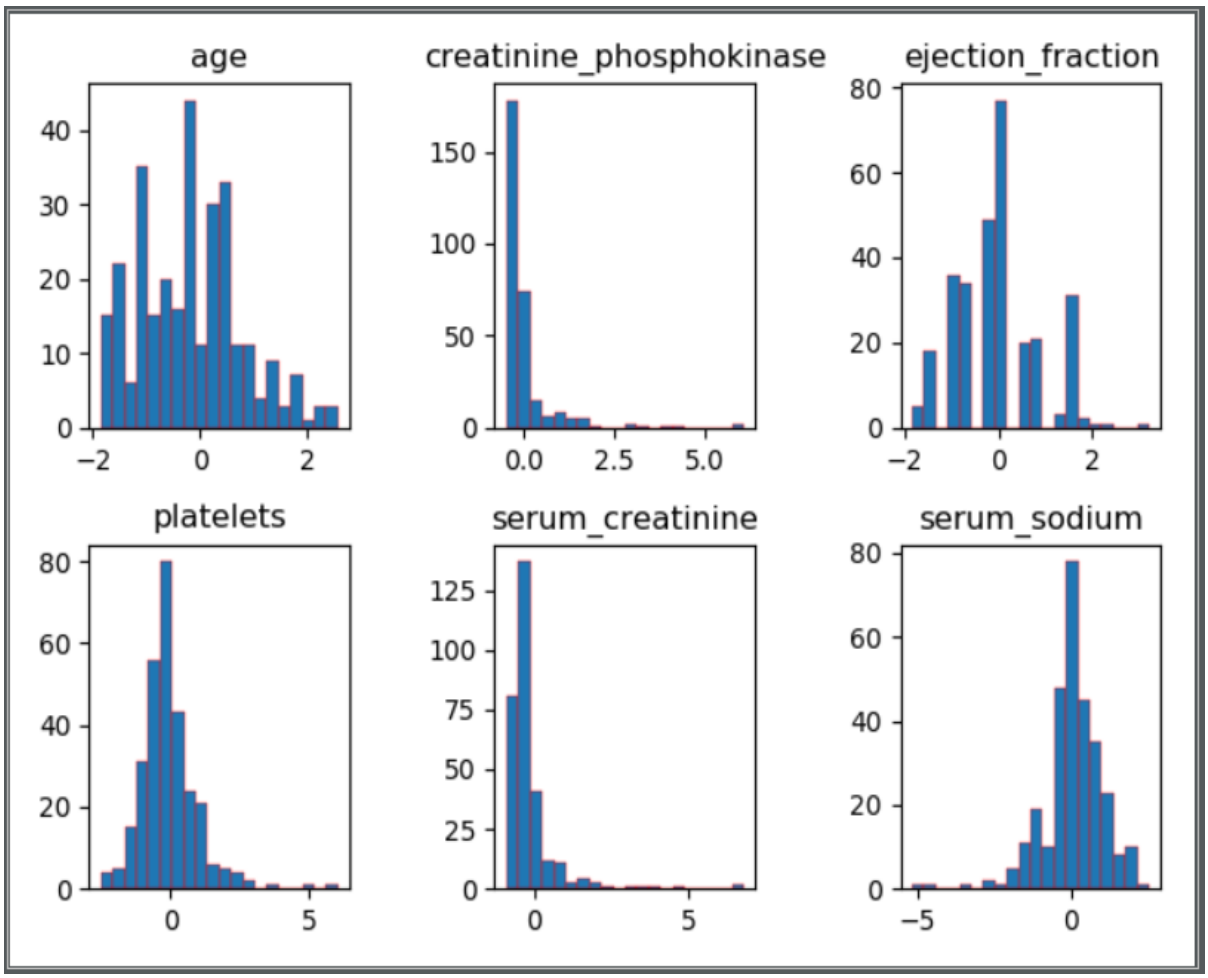

Рисунок 2 – Гистограммы признаков после стандартизации

Диапазоны и значения с наибольшим количеством наблюдений изменились таким образом, чтобы наибольшее число наблюдений оказалось вблизи нулевого значения. Причина изменений – преобразования, произведенные с данными. Далее проведена стандартизация данных по всему диапазону. В табл. 3 представлены значения математического ожидания и дисперсии для каждого признака до и после стандартизации. Так как значение, возле которого больше всего наблюдений по сути МО, то можно предположить, что для его смещения к нулю происходит вычитание из значения МО. Кроме того, происходит сжатие шкалы, так что можно предположить, что происходит деление на СКО. Таким образом, если X – исходные данные, а Y – преобразованные, то

$$Y_i \sim \frac{X_i - M(X)}{D(X)}.$$

Таблица 3. Значения мат. ожидания и дисперсии

Признак	**Изначальные данные**		**Ст. по первым 150**		**Ст. по всей выборке**		**Scaler**	
	МО	СКО	МО	СКО	МО	СКО	МО	СКО
1	60.83	11.87	-0.16	0.95	5.70e-16	1.00	62.95	155.00
2	581.84	968.66	-0.02	0.81	0	1.00	607.15	1415489
3	38.08	11.82	0.01	0.91	-3.27e-17	1.00	37.95	170.02
4	263358	97641	-0.04	1.02	7.72e-17	1.00	266746	9e9
5	1.39	1.03	-0.11	0.89	1.43e-16	1.00	1.52	1.36
6	136.63	4.41	0.04	0.97	-8.67e-16	1.00	136.45	20.61

Далее произведено приведение данных к диапазону [0, 1] с помощью MinMaxScaler. Гистограммы признаков после приведения представлены на рис. 3.

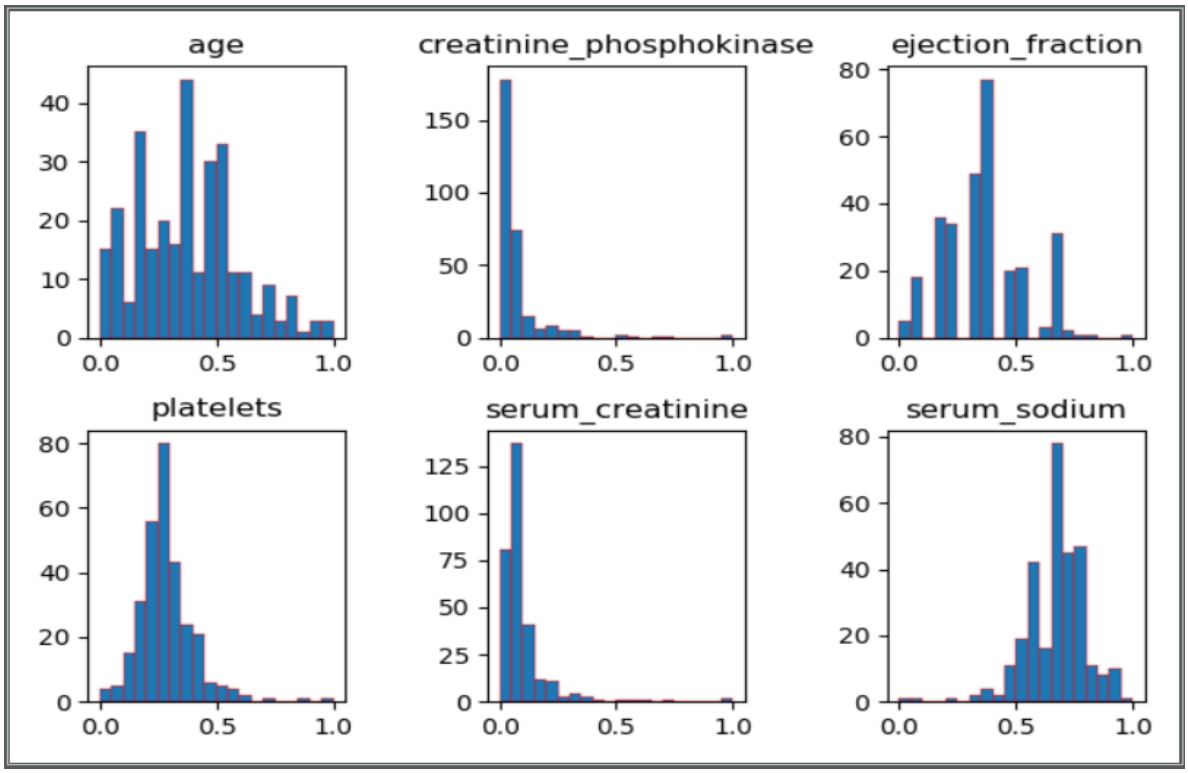

Рисунок 3 – Гистограммы признаков после приведения к [0, 1]

Результаты минимального и максимального значений каждого из признаков по данным MinMaxScaler представлены в табл. 4.

Таблица 4. Максимальные и минимальные значения согласно MinMaxScaler

Признак	**Минимальное значение**	**Максимальное значение**
age	40.0	95.0
creatinine_phosphokinase	23.0	7861.0
ejection_fraction	14.0	80.0
platelets	850000.0	25100.0
serum_creatinine	0.5	9.4
serum_sodium	113.0	148.0

Далее произведены аналогичные преобразования с помощью MaxAbsScaler и RobustScaler. Гистограммы признаков представлены соответственно на рис. 4 и 5.

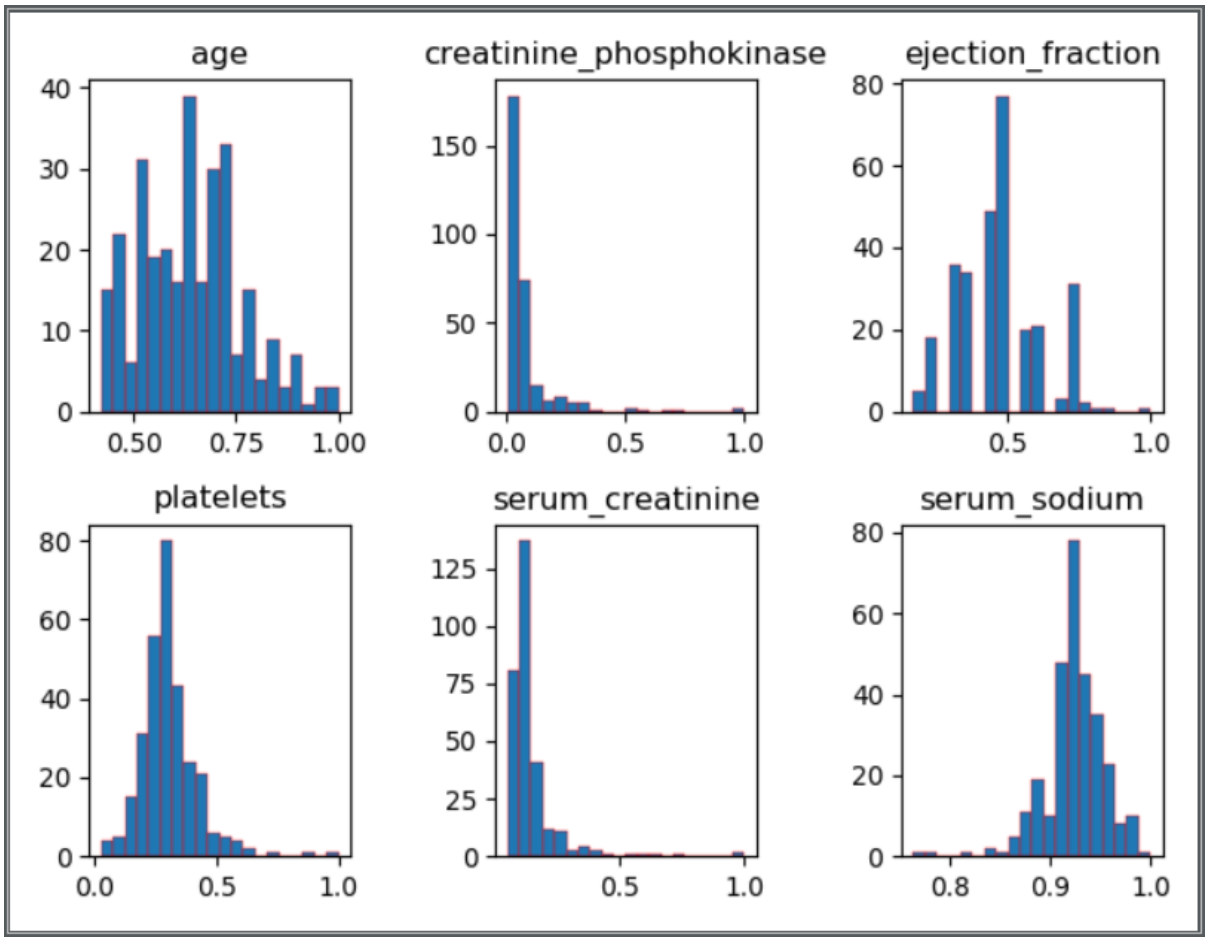

Рисунок 4 – Гистограммы после приведения с MaxAbsScaler

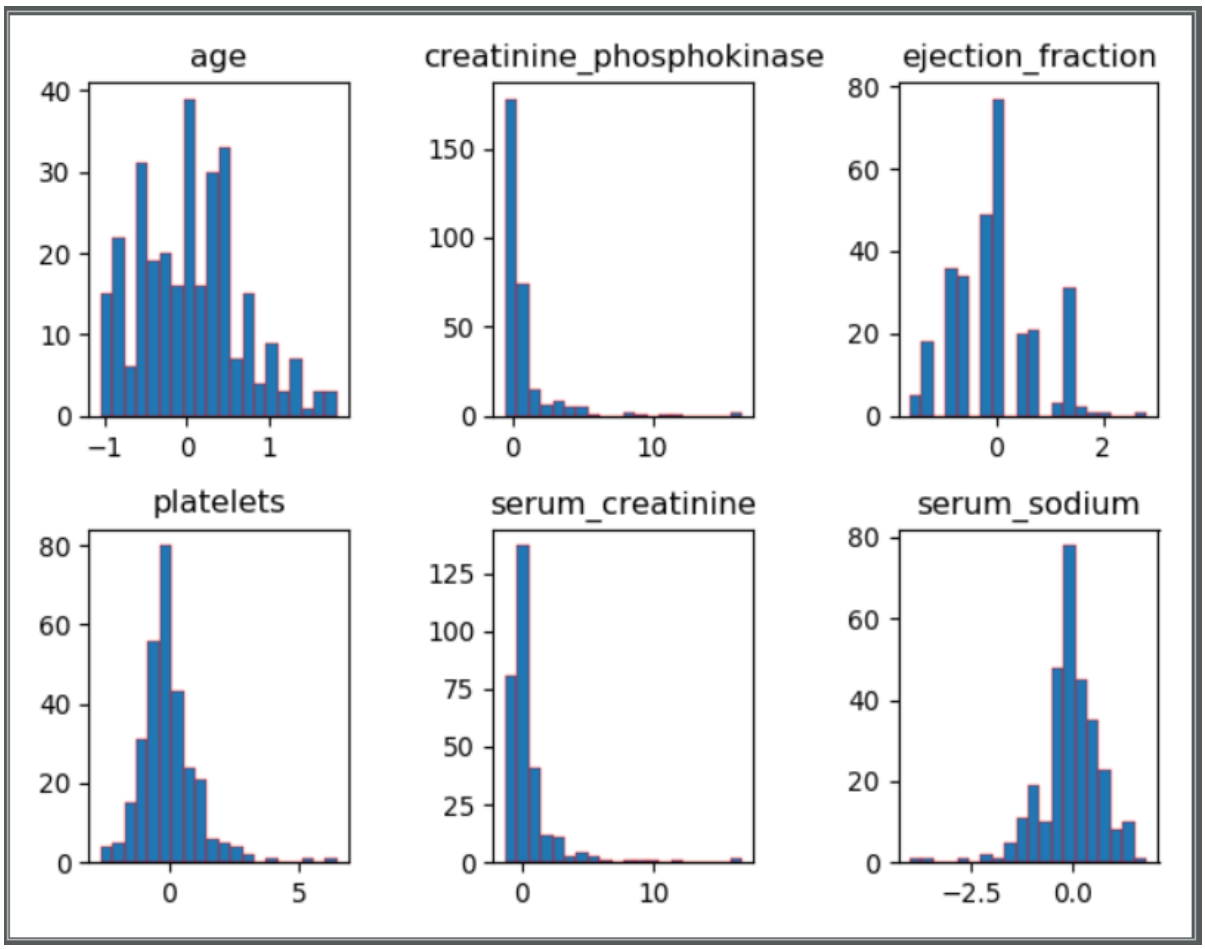

Рисунок 5 – Гистограммы после приведения с RobustScaler

MaxAbsScaler, как можно заметить, производит преобразования таким образом, что максимальное значение становится равно 1, а остальные так, чтобы сохранить исходное соотношение. Что, касается RobustScaler, то видно, что опять производится вычитание МО. Кроме того, производится некое масштабирование.

Далее создана функция, которая приводит данные к диапазону [-5, 10]. Гистограммы после преобразования представлены на рис. 6.

Следующими преобразованиями являются нелинейные преобразования к нормальному и равномерному распределению. Гистограммы соответственно представлены на рис. 7 и рис. 8.

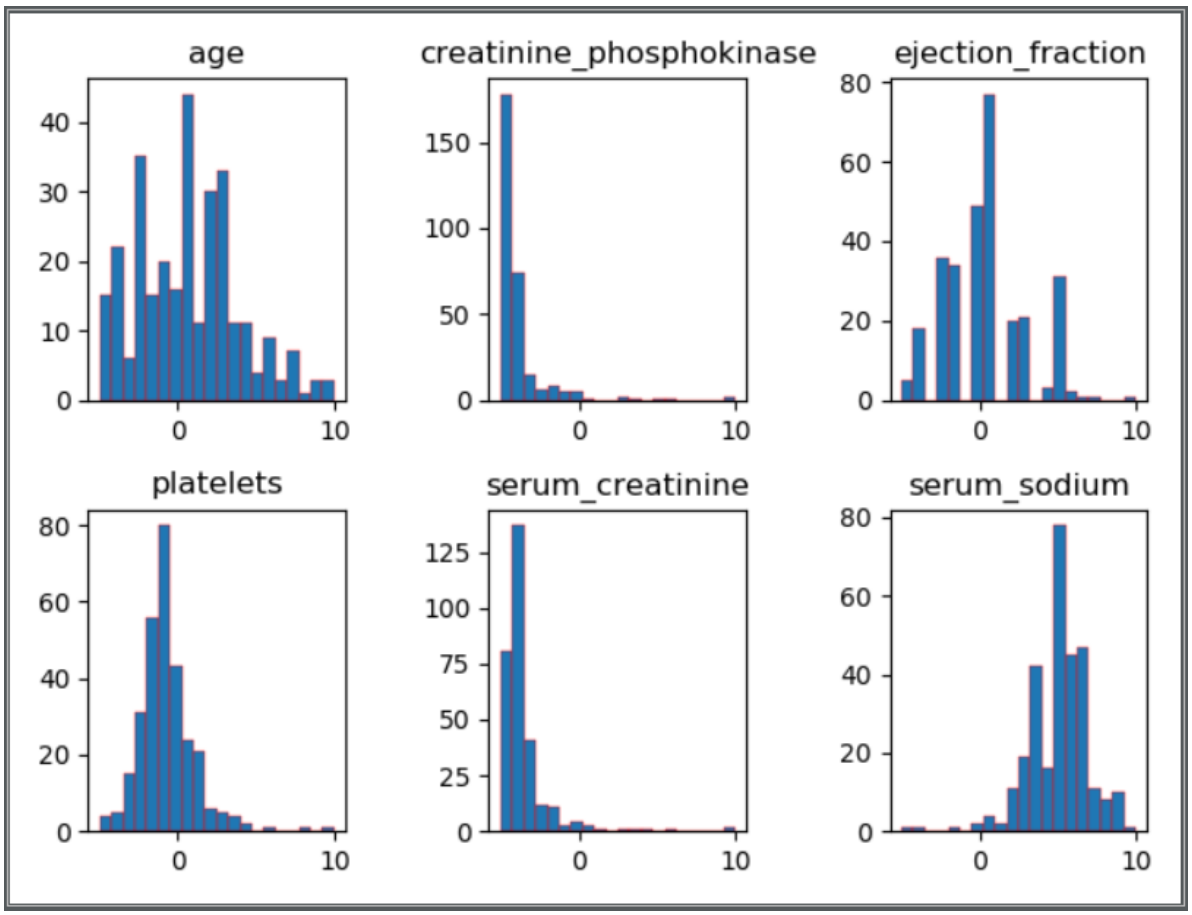

Рисунок 6 – Гистограммы признаков после приведения к [-5, 10]

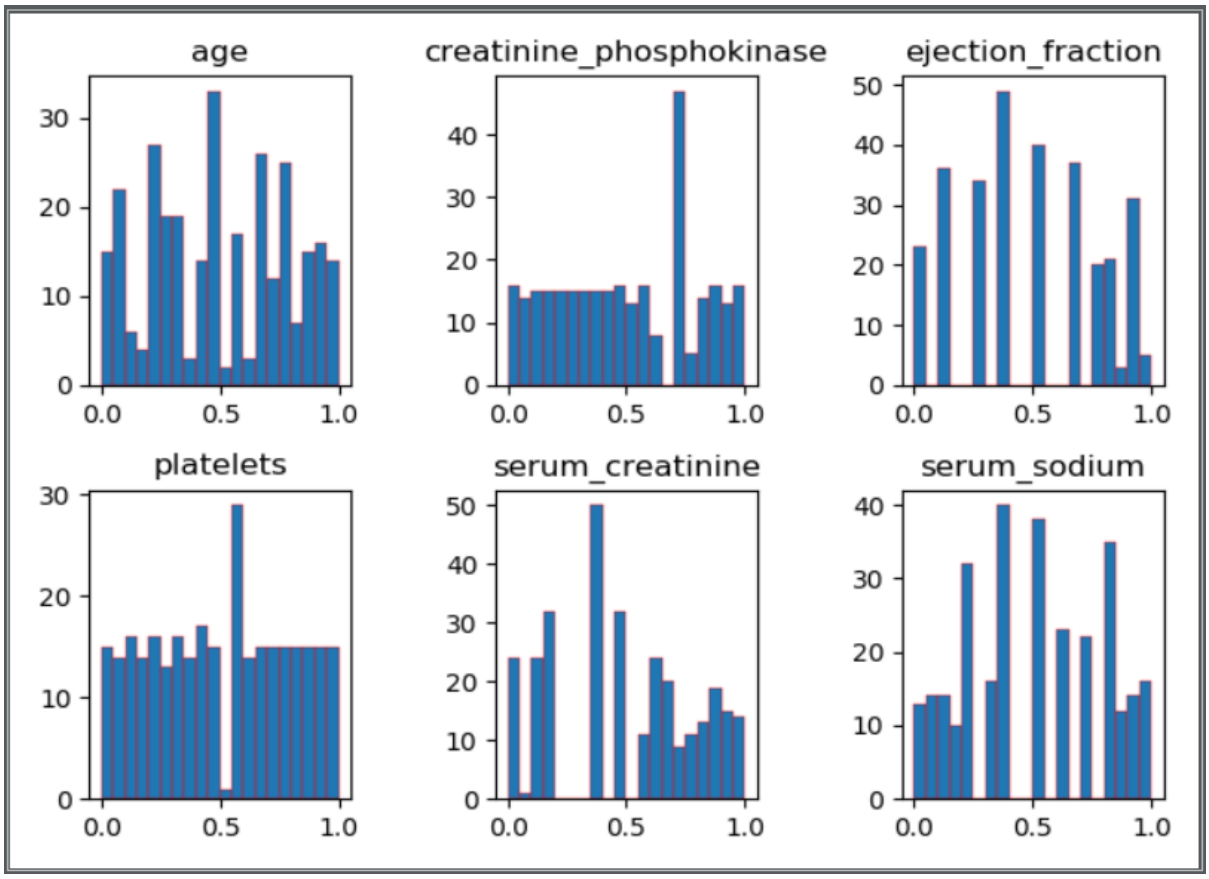

Рисунок 7 – Гистограммы приведенные к равномерному распределению

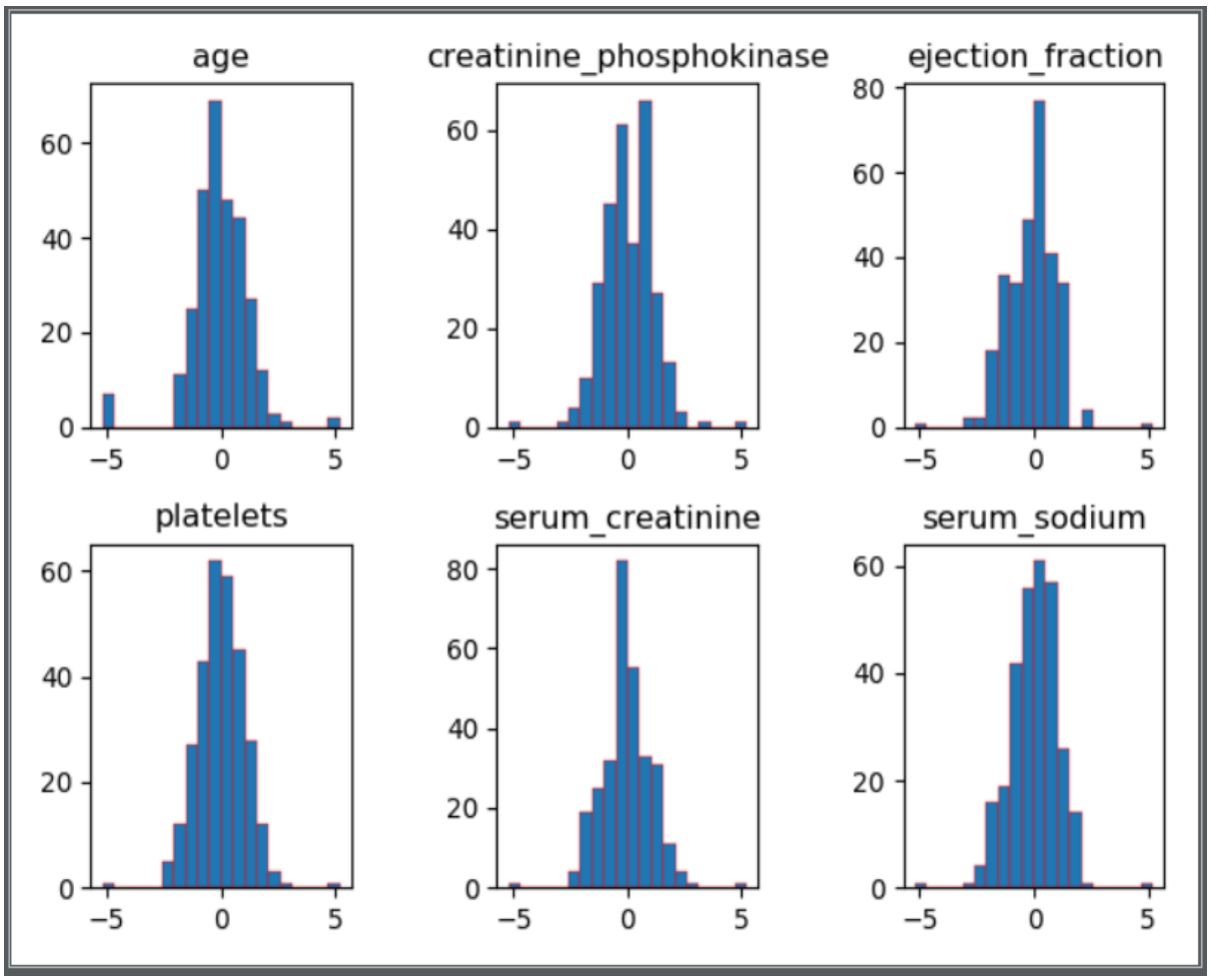

Рисунок 8 – Гистограммы приведенные к нормальному распределению

Параметр n_quantiles определяет количество квантилей. Чем больше данное число, тем выше частота дискретизации функции, что в свою очередь приводит к более точному соответствию получаемой функции ожидаемой. Также произведено приведение к нормальному распределению с помощью PowerTransformer. Гистограммы представлены на рис. 9.

Далее выполнена дискретизация признаков согласно указанному количеству интервалов. Результаты дискретизации представлены на рис. 9. Полученные диапазоны интревалов:

- 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.].

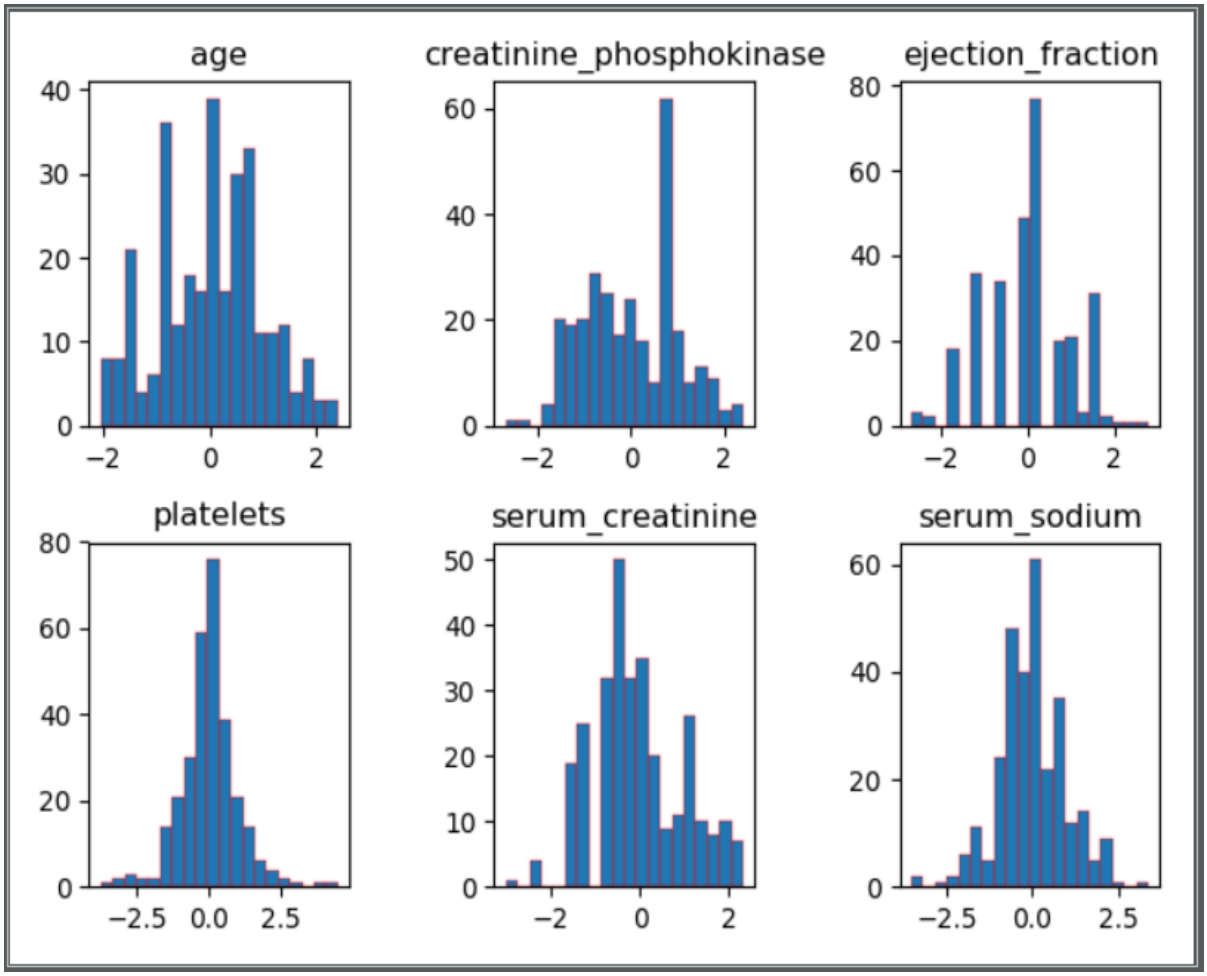

Рисунок 9 – Приведение к нормальному распределению с помощью PowerTransformer

Выводы.

В ходе выполнения данной работы было проведено ознакомление с методами предобработки данных с помощью методов библиотеки Scikit Learn. Были изучены и опробованы различные способы предобработки, такие как дискретизация признаков, нелинейные преобразования, приведение к диапазону, стандартизация данных, загрузка данных. Установлено, что стандартизация без учета всех данных приводит к снижению точности данных; приведение к диапазонам оказывает небольшое влияние на точность