**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 |  | Сергеев А.Д. |
| Преподаватель |  |  |

Санкт-Петербург

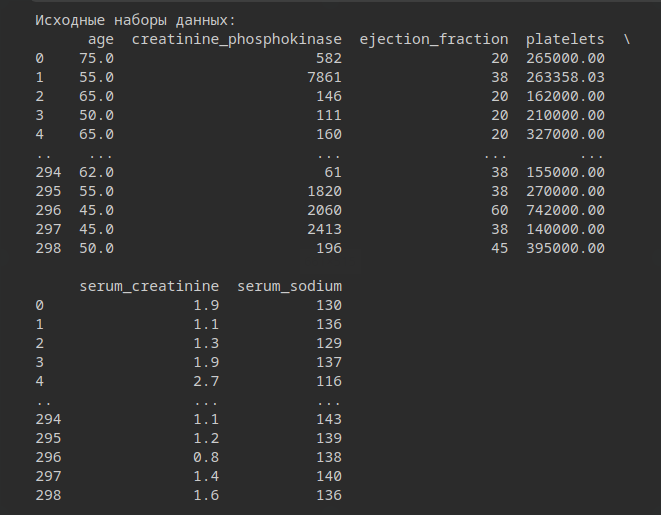
2021

**Цель работы.**

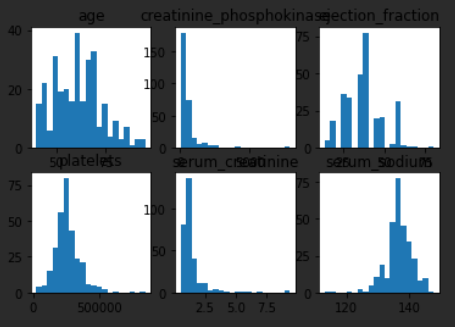
Ознакомиться с методами предобработки данных из библиотеки Scikit Learn.

**Ход работы.**

Данные были загружены в датафрейм, ненужные аттрибуты исключены. Полученные данные:

Рисунок 1 - исходные данные в датафрейме

Были построены гистограммы признаков:

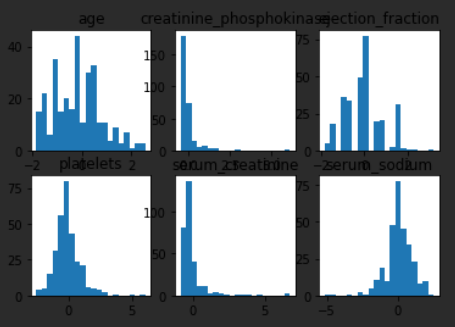
Рисунок 2 — гистограммы исходных данных

На основании гистограмм были определены диапазоны и медианы для каждого из признаков:

|  |  |  |  |
| --- | --- | --- | --- |
| Название признака | Минимум | Максимум | Медиана |
| age | 40 | 95 | 60 |
| creatinine\_phosphokinase | 23 | 7861 | 250 |
| ejection\_fraction | 14 | 80 | 38 |
| platelets | 25100 | 850000 | 262000 |
| serum\_creatinine | 0.5 | 9.4 | 1.1 |
| serum\_sodium | 113 | 148 | 137 |

Датафрейм преобразован к двумерному массиву.

Была настроена стандартизация на основе первых 150 наблюдений, после чего все данные были стандартизованы при помощи *StandartScaler*. Были построены гистограммы стандартизованных данных:

Рисунок 3 — гистограммы стандартизованных данных

По внешнему виду гистограммы стандартизованных данных похожи на гистограммы исходных данных. Масштаб оси ординат остался прежним, а масштаб оси абсцисс сильно изменился. В общем можно сказать, что значения на оси абсцисс на всех графиках принадлежат промежутку [-10 .. 10].

Для каждого параметра были высчитаны значения математического ожидания и среднеквадратического отклонения до и после стандартизации:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Название признака | МО до | МО после | СКО до | СКО после |
| age | 60.834 | -0.170 | 11.875 | 0.954 |
| creatinine\_phosphokinase | 581.839 | -0.021 | 968.664 | 0.814 |
| ejection\_fraction | 30.084 | 0.011 | 11.815 | 0.906 |
| platelets | 263358.029 | -0.035 | 97640.548 | 1.015 |
| serum\_creatinine | 1.394 | -0.109 | 1.033 | 0.885 |
| serum\_sodium | 136.625 | 0.038 | 4.405 | 0.970 |

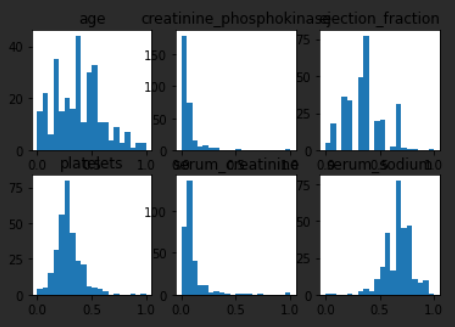
Из полученных данных можно сделать вывод о том, что для стандартизации данных была использована формула: *z ≈ (x - u) / s*, где *x* — исходное значение, *u* — математическое ожидание параметра, *s* — среднеквадратическое отклонение параметра.

Значения полей *mean\_* и *var\_* объекта *scaler* подтверждают прдположение:

|  |  |  |
| --- | --- | --- |
| Название признака | scaler.mean\_ | scaler.var\_ |
| age | 62.947 | 154.997 |
| creatinine\_phosphokinase | 607.153 | 1415488.823 |
| ejection\_fraction | 37.947 | 170.024 |
| platelets | 266746.749 | 9252860499.079 |
| serum\_creatinine | 1.521 | 1.361 |
| serum\_sodium | 136.453 | 26.608 |

Была проведена настройка стандартизации на всех данных. Результаты стали точнее, так как значения математического ожидания приблизились к 0, а среднеквадратичееского отклонения — к 1.

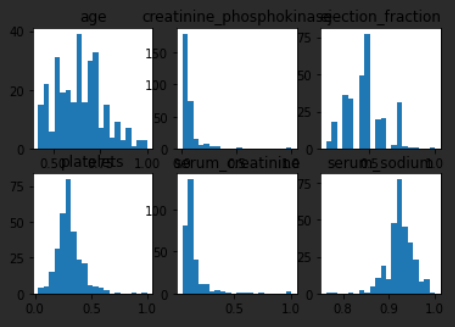
При помощи MinMaxScaler данные были приведены к диапазону [0, 1]:

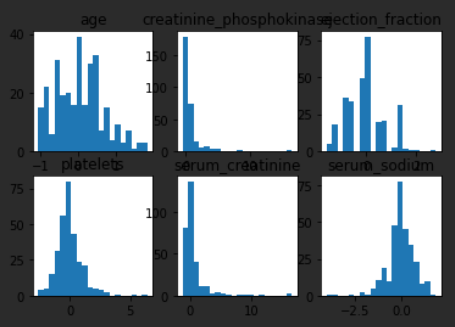
Рисунок 4 — гистограмма приведенных к диапазону данных

По внешнему виду гистограммы приведенных к диапазону данных похожи на гистограммы исходных данных. Масштаб оси ординат остался прежним, а масштаб оси абсцисс изменился, теперь все данные принадлежат отрезку [0, 1].

Таблица, в которой указаны минимальные и максимальные значения для каждого признака, была приведена выше. Значения не изменились.

Данные были трансормированы с помощью *MaxAbsScaler* и *RobustScaler*. Были построены гистограммы:

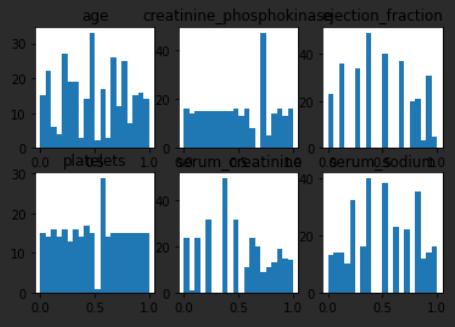
Рисунок 5 — гистограмма данных, обработанных *MaxAbsScaler*

Рисунок 6 - гистограмма данных, обработанных *RobustScaler*

Основываясь на документации *MaxAbsScaler* и *RobustScaler*, можно сказать, что первый также приводит данные к диапазону [0, 1], а второй — к интерквартильному диапазону (между 25 и 75 квантилью).

Для того, чтобы привести данные к диапазону [-5, 10] достаточно передать в конструктор *MinMaxScaler* параметр *feature\_range*, равный кортежу (-5, 10).

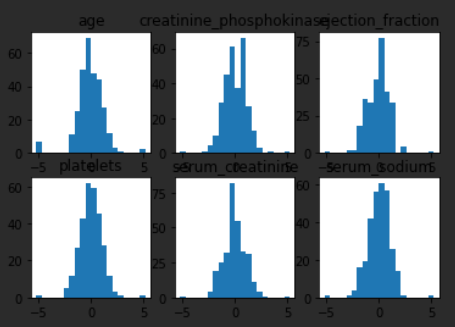
Данные были обработаны *QuantileTransformer* для получения равномерного распределения:

Рисунок 7 — гистограмма равномерно распределенных данных

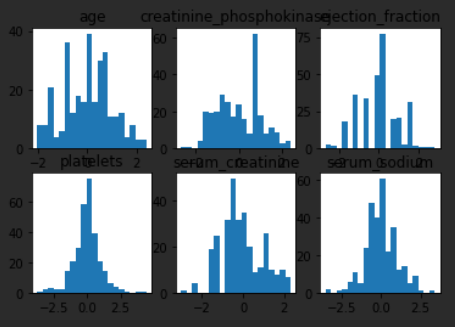
По внешнему виду гистограммы равномерно распределенных данных не похожи на гистограммы исходных данных.

Параметр *n\_quantiles* представляет из себя количество фрагментов, на которое будет разбита квантильная функция при вычислении. Согласно документации, максимально точное значение дает количество квантилей, равное количеству измерений.

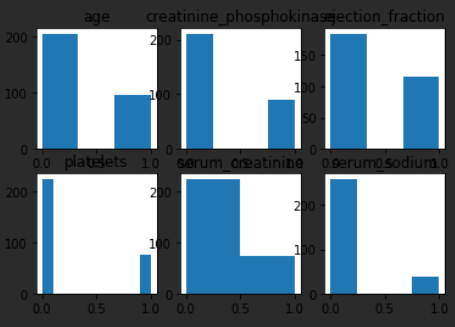
При передаче в конструктор *QuantileTransformer* параметра *output\_distribution*, равного *"normal"*, данные будут распределены нормально:

Рисунок 8 — нормально распределенные данные при помощи *QuantileTransformer*

Такжеданные могут быть обработаны и распределены нормально при использовании *PowerTransformer*.

Рисунок 9 — нормально распределенные данные при помощи *PowerTransformer*

Была проведене дискретизация данных с использованием *KBinsDiscretizer*.

Рисунок 10 — дискретизированные данные при помощи *KBinsDiscretizer*

Количество столбцов в гистограмме определенного признака равно значению параметра *n\_bins* для него. Внешний вид гистограммы похож на внешний вид гистограммы исходных данных с уменьшенным количеством столбцов.

Через параметр *bin\_edges\_* были получены границы диапазонов для каждого признака:

|  |  |
| --- | --- |
| Название признака | scaler.bin\_edges\_ |
| 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.] |

**Выводы.**

В ходе лабораторной работы было успешно произведено ознакомление с методами предобработки данных из библиотеки Scikit Learn.