**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

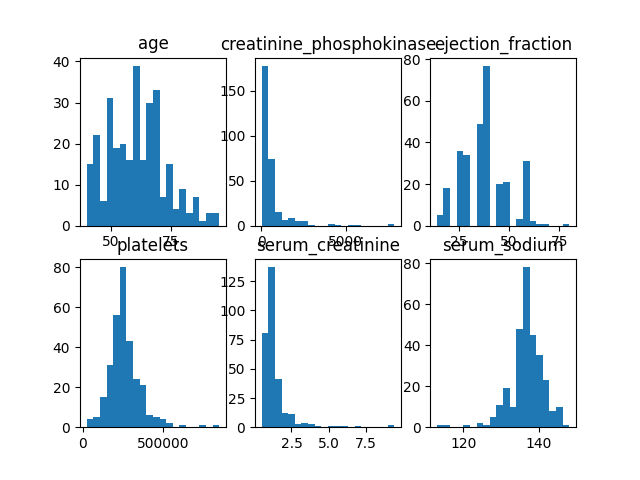


Рис 1 – гистограммы признаков

1. Стандартизация данных

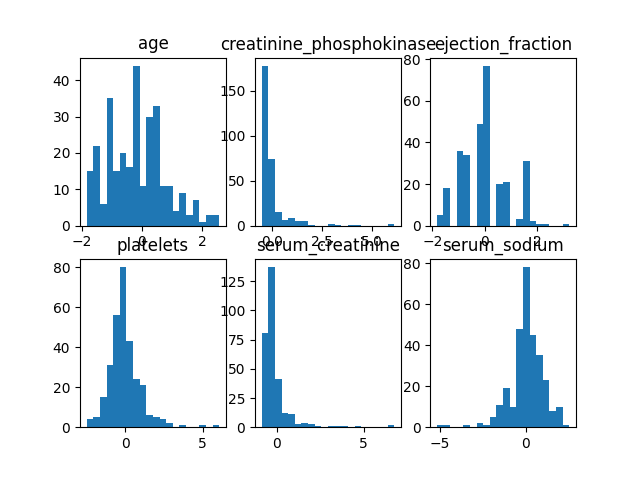


Рис 2 – Гистограммы стандартизированных данных по 150

Изменился диапазон и теперь среднее значение является нулем.

Таблица 1 – Сравнение данных стандартизации

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | age | сreatinine  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:

1. Приведение к диапазону

**MinMaxScaler**

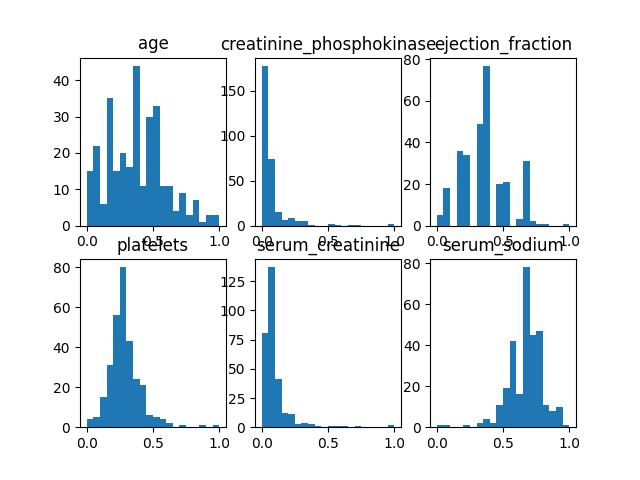


Рис 3 – Гистограммы мин макс

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

Таблица 2 – Данные в полях min и max

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | age | сreatinine  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**

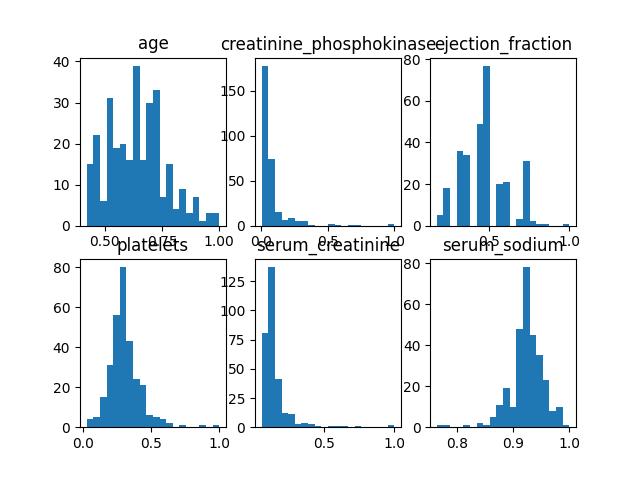


Рис 5 – гистограмма макс абс

Данные приводятся к диапазону, где max\_abs\_ берется как 1

Таблица 3 – Данные max\_abs\_

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

**RobustScaler**

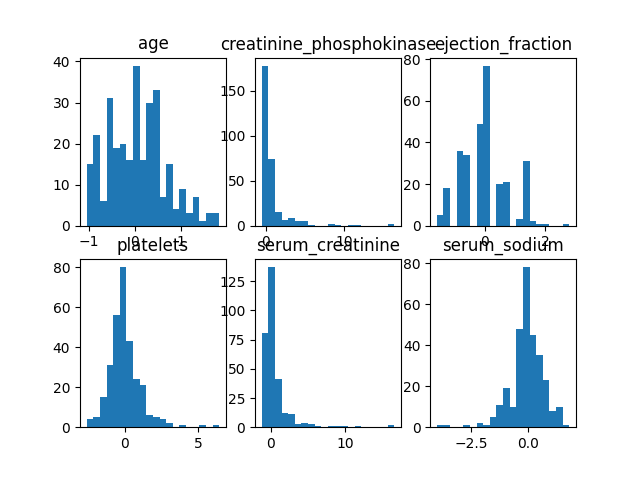


Рис 6 – Гистограмма Робус

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

Таблица 4 – Данные center\_

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

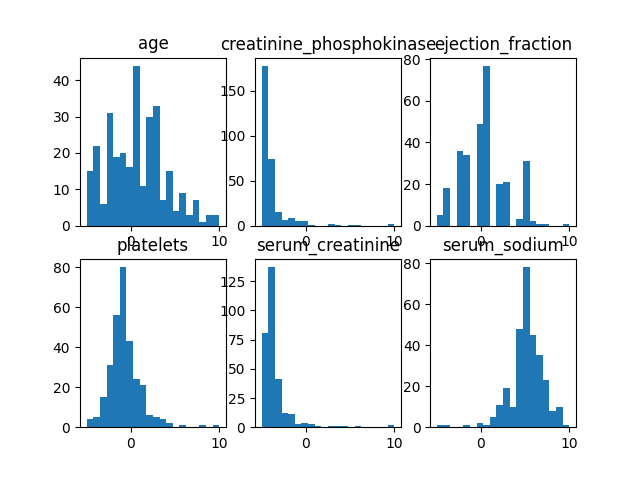


Рис 7 – Гистограммы для диапазона [-5; 10]

1. Нелинейные преобразования

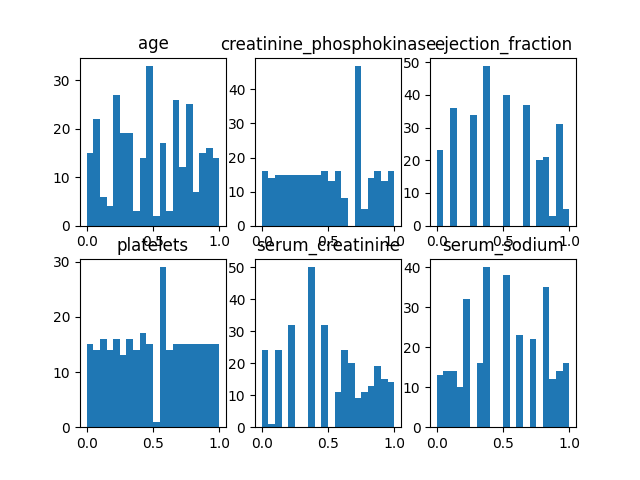


Рис 8 – Гистограмма преобразованных данных

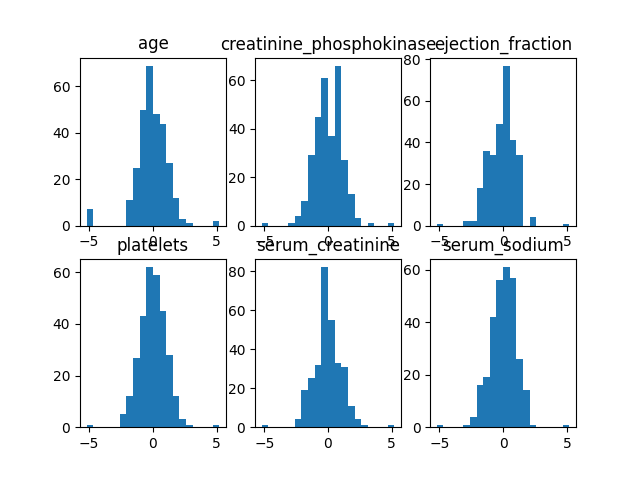


Рис 9 - Гистограмма преобразованных данных в нормальном распределении

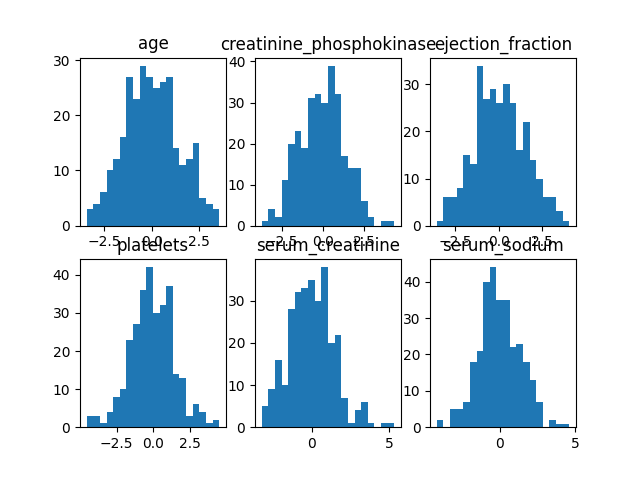


Рис 10 – Нормальное распределение при использовании PowerTransformer

1. Дискретизация признаков

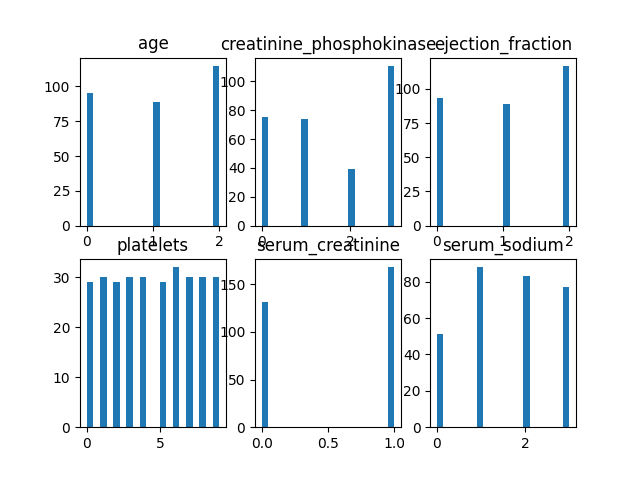


Рис 11 - Гистограммы дискретизации

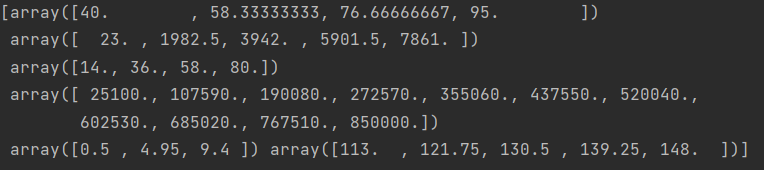


Рис 12 – Диапазоны для каждого признака